If we watch the particle moving along its great circle orbit L at unit speed for time Δt then the radius will rotate within the plane of L by $\Delta\theta = (\Delta t/R)$. Thus the radius sweeps out area $\mathcal{A} = \frac{1}{2}R^2\Delta\theta = \frac{1}{2}R\Delta t$, and so area is swept out at a *constant rate*: the same area $\mathcal{A}$ will be swept out in each equal period of time Δt. Historically, this situation is summarized by saying that *equal areas are swept out in equal times*.

Now let us project this orbit vertically downward onto the equatorial plane Π. This projection $\mathcal{P}$ is a linear transformation, and therefore all areas are contracted by the same factor, the determinant of the transformation. In fact it is easy to see geometrically [exercise] that if the plane of L makes an angle γ with Π, then $\det \mathcal{P} = \cos\gamma$.

It is also easy to see that the projection of the circular orbit L is an elliptical orbit $\mathcal{P}(L)$ in Π. It follows that as the particle travels round this ellipse, the radius in Π *also* sweeps out equal areas in equal times. In greater detail,

$$\frac{d\widetilde{\mathcal{A}}}{dt} = \cos\gamma \frac{d\mathcal{A}}{dt} = \frac{1}{2}R\cos\gamma.$$

Since $\mathcal{P}(L)$ is an ellipse, this steady generation of area is only possible if the projected particle moves slower when it is further away and faster when it is closer in. This variation in speed is readily confirmed and quantified by noting [exercise] the following:

If $\mathbf{q}(t)$ *is any trajectory in space, the velocity of the orthogonal projection* $\mathcal{P}(q)$ *onto a plane* Π *is the projection of the velocity:* $\frac{d}{dt}\mathcal{P}[\mathbf{q}] = \mathcal{P}\left[\frac{d\mathbf{q}}{dt}\right]$. (11.7)

The tangent plane T_q (not drawn) is spanned by the tangents to the orthogonal circles of longitude and latitude. As illustrated, the *direction* of the geodesic at q can be described by the angle ψ it makes it makes with the meridian through q. If the particle q moves for a short time ϵ it will travel a small distance ϵ along the great circle orbit. Thus the horizontal component of the motion along the circle of latitude is $\epsilon\sin\psi$, as illustrated. It follows that the unit velocity vector to the geodesic, lying within T_p can be decomposed into a horizontal component $\sin\psi$ along the circle of latitude, and a component $\cos\psi$ along the circle of longitude (meridian).

The component along the meridian projects to radial motion in Π, generating no area. The area generation is entirely due to the horizontal component $\sin\psi$, which projects to an equal component perpendicular to the radius ρ in Π; note that ρ is the distance of the original particle from the axis of symmetry, as illustrated.

Since the rate of generation of area in Π is $\frac{1}{2}\rho\sin\psi$, we deduce that this quantity is constant along the original great circle orbit, and is given by,

$$\tfrac{1}{2}\rho\sin\psi = \tfrac{1}{2}R\cos\gamma = const.$$

While this formula could have been demonstrated directly from the geometry of the sphere, the advantage of the above argument is that we will soon be able to generalize it to obtain,

Clairaut's Theorem. *Let* $\mathcal{S}$ *be a surface of revolution generated by rotating a curve* $\mathcal{C}$ *about an axis* $\mathcal{L}$. *If* ρ *is the distance from the axis* $\mathcal{L}$ *to a point* q *on a geodesic* g, *and* ψ *is the angle between the meridian* $\mathcal{C}_q$ *through* q *and the direction of* g, *then* $\rho\sin\psi$ *remains constant as* q *travels along the geodesic* g. *Conversely, if* $\rho\sin\psi$ *is constant along a curve* g *(no part of which is a parallel of* $\mathcal{S}$*), then* g *is a geodesic.* (11.8)

(Recall that a ***parallel*** is a horizontal circle on $\mathcal{S}$, the intersection of $\mathcal{S}$ with a plane perpendicular to $\mathcal{L}$.)

HISTORICAL NOTE: Alexis Claude Clairaut (1713–1765) was a French mathematician, astronomer, and geophysicist, who helped to extend Newton's results in the *Principia*. In 1752 he published an accurate, usable, approximate solution to the three-body problem of the Sun, Earth, and Moon, which Euler declared to be "*... the most important and profound discovery that has ever been made in mathematics.*" (Hankins, 1970, p. 35) The naming of the general theorem above stems from Clairaut's 1733 investigation of quadratic surfaces of revolution.

11.7.2 Kepler's Second Law

In order to understand the general version of Clairaut's Theorem, let us ask, *what is the magnitude and direction of the force that holds* $\mathcal{P}(q)$ *in its orbit* $\mathcal{P}(L)$*?* To answer this, we note the simple generalization of (11.7) from velocity to acceleration:

If $\mathbf{q}(t)$ *is any trajectory in space, the acceleration of the orthogonal projection* $\mathcal{P}(\mathbf{q})$ *onto a plane* Π *is the projection of the acceleration:* $\frac{d^2}{dt^2}\mathcal{P}[\mathbf{q}] = \mathcal{P}\left[\frac{d^2\mathbf{q}}{dt^2}\right]$. (11.9)

Suppose for simplicity's sake that the particle orbiting on the sphere in [11.5] has unit mass, so that force is equal to acceleration. The force $\mathbf{F}$ that holds q in this orbit L is directed along the normal to the sphere, pointing directly at the centre, O, and it has constant magnitude $(v^2/R) = (1/R)$, since the particle has unit speed. Thus the force $\widetilde{\mathbf{F}}$ that holds the projection in its elliptical orbit in Π is *also* directed at O, and by virtue of the illustrated similar triangles,

$$\frac{\left|\widetilde{\mathbf{F}}\right|}{|\mathbf{F}|} = \frac{\rho}{R} \qquad \Longrightarrow \qquad \left|\widetilde{\mathbf{F}}\right| = (1/R^2)\rho.$$

A force field that is directed towards a single point O is called a ***central force field***. We have just established that the central force field in which the magnitude of the force directed to the centre of force is *proportional to the distance* of the particle from O results in an *elliptical orbit centred at* O *that sweeps out equal areas in equal times*. This was first proved by Newton (using a quite different argument) in the *Principia* (Proposition 10).

In practical terms, this force field and orbit can be created as follows. Imagine Π to be a frictionless sheet of ice, with a small hole at O. Take a small ice puck, and attach to it a length l of elastic string. Pass the length of string through the hole, and rest the puck on top of the hole. Take the other end of the elastic, hanging down at distance l below the hole, and attach it to a fixed point. Sitting at O, the puck experiences no pull from the string, because the string is relaxed at its natural length l. But if we move the puck a distance ρ away from O, thereby stretching the string by ρ, Hooke's Law then tells us that the stretched elastic will pull the puck back towards O with a force proportional to ρ. If we now launch the puck across the ice in an arbitrary direction with arbitrary speed, it will indeed trace an elliptical orbit centred at O, sweeping out equal areas in equal times. It is not hard to create an (ice-free) approximation of this experiment at home, and we encourage you to do so.

The fact that the orbit is an *ellipse centred at* O is specifically linked to the fact that the force varies *linearly* with distance. But, as Newton was the first to recognize and prove, the fact that equal areas are swept out in equal times is a remarkable *universal* property of *all* central force fields! We shall describe Newton's beautiful proof momentarily, but first let us pause to understand the pivotal role of this result in the *Principia*.

By analyzing the painstakingly accurate observations of the planets taken over a period of years by Tycho Brahe (1546–1601), Johannes Kepler (1571–1630) was able to discern mathematical

patterns within the mass of data before him. These empirical mathematical facts are now known as Kepler's Three Laws of Planetary Motion. Kepler announced the first two laws in 1609, and the third (after tremendous struggles) in 1618:

Kepler's Laws

(I) *The orbit of a planet is an ellipse with the Sun at one of the two foci.*
(II) *A line segment joining the Sun to a planet sweeps out equal areas in equal times.*
(III) *The square of the orbital period of a planet is proportional to the cube of the semi-major axis of its orbit.*

For the next 70 years these laws would remain a mystery. Finally, in 1687, Newton succeeded in mathematically *explaining* Kepler's Laws as logical consequences of his universal, Inverse-Square Law of Gravitation—a spectacular vindication of his ideas. But in order for Newton to be able to achieve his analysis of dynamics using *geometry alone* (see Prologue) it was essential for him to be able to be able to *represent time as a geometrical quantity*.

Kepler's Second Law, or rather Newton's generalization of it to arbitrary central force fields, was therefore absolutely critical to the entire enterprise—it stands as Proposition 1 of the *Principia*. Newton's legion geometric diagrams and associated proofs in the *Principia* simply would not have been possible without this one absolutely fundamental fact: *area is the clock*.

11.7.3 Newton's Geometrical Demonstration of Kepler's Second Law

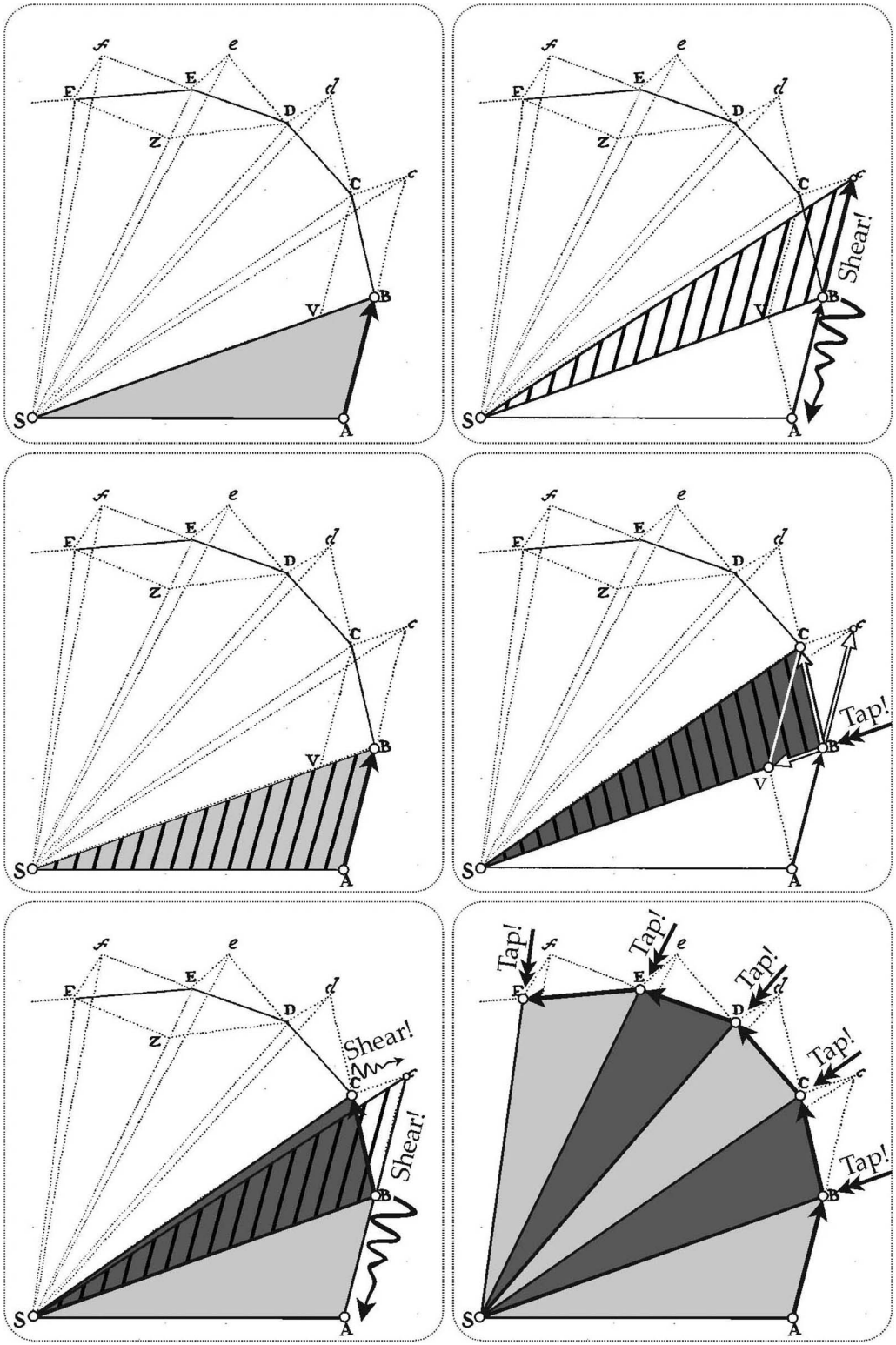
Figure [11.6] contains six copies of Newton's own diagram for Proposition 1 of the *Principia*, embellished so as to tell the story (as in a comic strip) of Newton's argument, establishing that in any central force field directed towards the fixed point *S*, the orbit ABCDEF sweeps out equal areas in equal times.

In the absence of any force, the first panel shows a particle travelling from A to B in a straight line, at uniform speed, as dictated by Newton's First Law of Motion; in this period of time, the radius from a fixed point S sweeps out the shaded area SAB. In the next equal increment of time, the particle will continue on an equal distance from B to c, still in the plane SAB, sweeping out the cross-hatched area SBc. But these two areas are *equal*, for the illustrated shear along BA brings SBc into coincidence with SAB. Thus, in the absence of force, equal areas are swept out in equal times.

Now suppose, instead, that at the moment the particle arrives at B it receives a sharp tap directed towards S. If the particle had been at rest at B initally, then this tap would have caused it to travel from B to v in the same time that it formerly travelled from B to c. The actual motion of the particle with therefore be the *sum* of these two motions, from B to C, still in the plane SAB, sweeping out the darkly shaded, cross-hatched area SBC. But this area is again *equal* to the original area SAB, for the illustrated shear parallel to SB carries SBC to SBc, and then (as before) a second shear brings SBc into coincidence with SAB.

Suppose we give the particle a second tap[2] towards S as it arrives at C, then it will travel from C to D, and, by the same reasoning as before, the area SCD will again equal the area SAB. Continuing in this manner, tapping the particle towards S at equal time intervals (at D, E, F, ...) the polygonal orbit ABCDEF sweeps out equal areas in equal times.

[2]Note that Newton's diagram appears to take the *magnitude* of each tap to be equal, but his argument is equally valid if the taps are unequal.



[11.6] Kepler's Second Law. *Six copies of Newton's own diagram for Proposition 1 of the* Principia *have been embellished so as to tell the story (as in a comic strip) of Newton's argument, establishing that in any central force field directed towards the fixed point* S*, the orbit* ABCDEF *sweeps out equal areas in equal times.*

Newton concludes,

> *Now let the number of triangles be increased and their width decreased indefinitely, and their ultimate perimeter ADF will be a curved line; and thus the centripetal force by which the body is continually drawn back from the tangent of the curve will act uninterruptedly, while any areas described will be proportional to those times Q.E.D.*

The remarkable elegance and economy of Newton's reasoning was not lost on Richard Feynman, who said[3] to his Caltech class in 1964,

> *The demonstration you have just seen is an exact copy of one in the* Principia Mathematica *by Newton, and the ingenuity and delight you may or may not have gotten from it is that already existing in the beginning of time.*

Lastly, observe [exercise] that the *converse* is also true; this is Proposition 2 of the *Principia*:

> *Every body that moves in some curved line described in a plane and, by a radius drawn to a point describes areas around that point proportional to the times, is urged by a centripetal force tending towards that same point.*

11.7.4 *Dynamical Proof of Clairaut's Theorem*

We are now but a step away from a satisfying explanation of Clairaut's Theorem on a general surface of revolution, as illustrated on the vase shown in [11.7].

As the point q moves along a geodesic g on this surface, the acceleration of the orbit is (by definition) always directed along the surface normal $\mathbf{n}$, and its direction therefore intersects the axis of symmetry $\mathcal{L}$. But, by virtue of (11.9), the projection of g onto Π is an orbit $\mathcal{P}(g)$ whose acceleration is therefore directed towards O, and so by Newton's generalized version of Kepler's Second Law, $\mathcal{P}(q)$ sweeps out area at a constant rate in Π. But, in the short time $\delta t = \epsilon$, the particle q travels a distance ϵ along the geodesic on the surface, and its projection sweeps out an area $\delta\mathcal{A}$ on the plane Π that is ultimately equal to the area of the white triangle with base ρ and height $\epsilon \sin\psi$. Thus $\delta\mathcal{A} \asymp \frac{1}{2}\rho\epsilon \sin\psi$, and so $\frac{d\mathcal{A}}{dt} = \frac{1}{2}\rho \sin\psi$ is constant. This completes the explanation of the first part of Clairaut's Theorem.

Note that meridians ($\psi = 0$) are exceptional geodesics, in the sense that their projections in Π move in and out radially, generating no area.

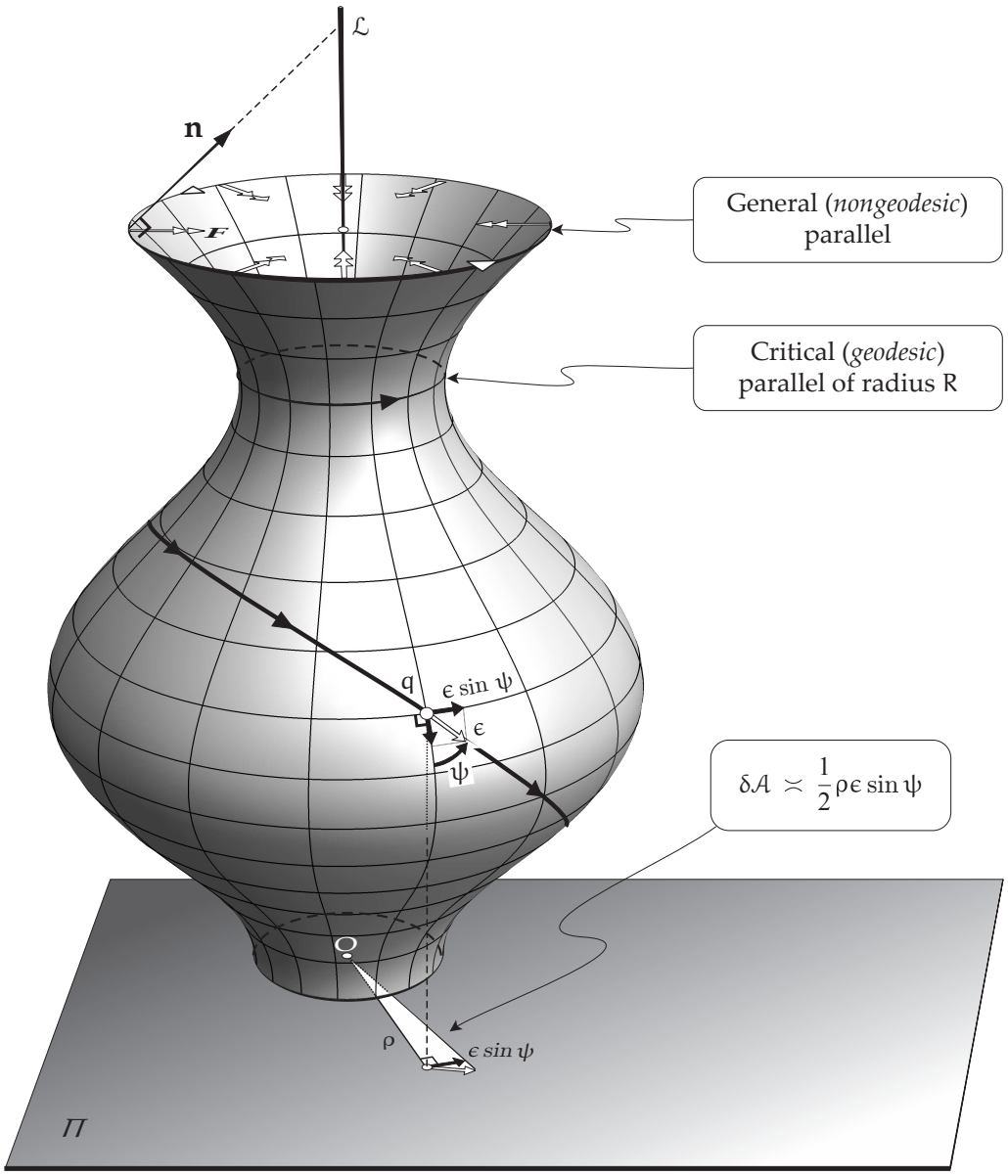
As for the converse, suppose $\rho \sin\psi$ is constant along a curve g on $\mathcal{S}$. Then $\mathcal{P}(q)$ sweeps out area in Π at a constant rate. Therefore, by Proposition 2 of the *Principia* (above), it follows that the acceleration of $\mathcal{P}(q)$ is always directed towards O. And from this it follows that the acceleration of g itself is always directed at the axis $\mathcal{L}$; equivalently, it lies within the vertical plane through q and $\mathcal{L}$, containing $\mathbf{n}$; equivalently, it is perpendicular at q to the parallel through q.

Next, assume the small segment of g containing q is *not* a parallel. Then $\mathbf{v}$ is not tangent to the parallel ($\psi \neq \frac{\pi}{2}$). Thus *the acceleration is perpendicular to two distinct directions in* $\mathcal{S}$: the direction of the parallel, and the direction $\mathbf{v}$ of g. Thus the acceleration is directed along $\mathbf{n}$, and so g is a geodesic, as was to be shown.

But suppose the segment of g containing q *is* part of a horizontal, circular parallel $\wp$, such as the top rim of the vase in [11.7]. Note that $\rho \sin\psi$ is indeed constant on $\wp$, by virtue of the fact that ρ and ψ are both separately constant, with $\psi = (\pi/2)$. But now, as illustrated, the acceleration is *horizontal* and directed towards the centre of the circular parallel, and this horizontal direction is in general *not* perpendicular to the surface, and hence $\wp$ is (in general) *not* geodesic.

---

[3] See his so-called *Lost Lecture* (Goodstein and Goodstein, 1996, p. 156), the audio recording of which is available on the internet.



[11.7] ***Clairaut's Theorem***. *In the short time $\delta t = \epsilon$, the particle q travels a distance ϵ along the geodesic on the surface, and its projection onto Π sweeps out an area $\delta\mathcal{A}$ that is ultimately equal to the area of the white triangle with base ρ and height $\epsilon \sin \psi$. Thus $\frac{d\mathcal{A}}{dt} = \frac{1}{2}\rho \sin \psi$, and therefore this quantity is constant, by virtue of Newton's generalization of Kepler's Second Law.*

However, $\wp$ *is* geodesic in the exceptional case that $\mathbf{n}$ is horizontal along $\wp$; an example is indicated in [11.7]. Such parallels $\wp$ are called *critical*, and can be characterized in various ways. For example, if the vertical cylinder with axis $\mathcal{L}$ containing $\wp$ touches $\mathcal{S}$ along $\wp$, then $\wp$ is geodesic. Most textbooks instead describe this situation by supposing that the generating curve $\mathcal{C}$ of $\mathcal{S}$ can be described by a graph $\rho = \rho(z)$, where z is vertical distance along $\mathcal{L}$. Then the parallel $\wp$ is geodesic if and only if it is "critical" in the sense that $\rho'(z) = 0$; in other words, the distance of the profile curve from $\mathcal{L}$ has a maximum, a minimum, or a point of inflection.

For readers who have studied some physics, we note that the quantity $\rho \sin\psi = \Omega$ is the ***angular momentum*** about the axis $\mathcal{L}$ of the (unit-mass) particle. The fact that Ω remains constant can be understood physically[4]: it is a consequence of the fact that the force that holds q in its geodesic orbit on $\mathcal{S}$ passes through $\mathcal{L}$ and therefore has no ***moment*** about $\mathcal{L}$. In exactly the same way that a spinning ice-skater spins faster as she pulls her arms in, so, in order to conserve its angular momentum, the particle orbiting on the surface must spin around $\mathcal{L}$ faster as it comes closer to $\mathcal{L}$. But since the particle on the surface has constant *linear* speed along the geodesic, it can only increase it *angular* speed around the axis by directing its velocity towards the horizontal.

Note that the angular momentum Ω of a geodesic g actually tells us the closest g can get to $\mathcal{L}$:

$$\rho = \frac{\Omega}{\sin\psi} \geqslant \Omega = \rho_{\min}.$$

For example, suppose g starts below the critical latitude (of radius R) at the throat of the vase in [11.7], and then climbs upward towards it. If the angular momentum is too large ($\Omega = \rho_{\min} > R$) then the g cannot ever reach the throat of the vase; instead, it bounces back down the vase. This is the case for the illustrated geodesic, which will likewise be forced to bounce back *up* again as it approaches the narrow base. To get a better feel for all this, try using a real vase, constructing geodesics using the sticky-tape construction, (1.7), page 14. See Exercise 11, page 26.

11.7.5 *Application: Geodesics in the Hyperbolic Plane (Revisited)*

Let us now apply Clairaut's Theorem to the pseudosphere, so as to gain a fresh view of the geodesics of the hyperbolic plane.

Since the pseudosphere becomes arbitrarily narrow the higher up we go, it follows from the discussion above that if a particle is travelling up the pseudosphere and has *any* angular momentum at all, it must eventually turn back and head back down. Thus the meridian, tractrix generators (which have zero angular momentum) are the *only* geodesics than can continue indefinitely upward.

Recall the construction of the Beltrami–Poincaré upper half-plane map of the pseudosphere, which we derived geometrically in [5.4], page 55. What we had called X in that figure is now called ρ. Thus the height y in the map (given by (5.4), page 55) is

$$y = \frac{1}{\rho},$$

where we have chosen the radius of the pseudosphere to be $R = 1$, in order to obtain the standard model of the hyperbolic plane, for which

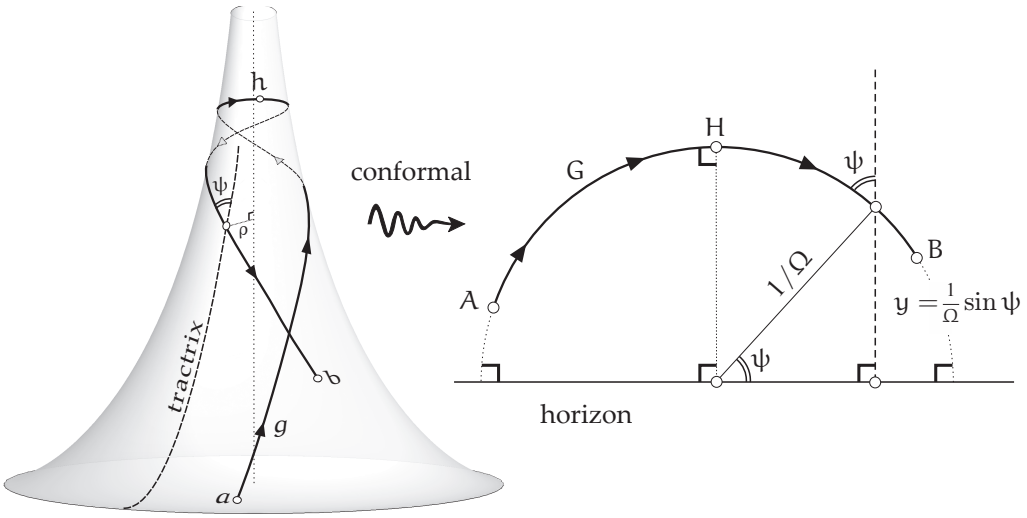
$$\mathcal{K} = -(1/R^2) = -1.$$

Figure [11.8] shows a geodesic g (starting at a and ending at b) on the pseudosphere and its image in the map (starting at A and ending at B). Recall that we previously proved (using optics) that this image is a semicircle meeting the horizon $y = 0$ at right angles. We can now give a fresh proof of this fact using Clairaut's Theorem, and in the process give a new interpretation of the *size* of the semicircle representing g.

Recall that this map is, by construction, *conformal*. That means that the angle ψ between g and the tractrix generator (meridian) of the pseudosphere is preserved: *The image of* g *in the map makes the* same *angle* ψ *with the vertical half-line image of the tractrix.* If the unit-mass particle travelling along g has angular momentum Ω, then Clairaut's Theorem yields,

$$\rho \sin\psi = \Omega \qquad \Longrightarrow \qquad y = \frac{1}{\Omega}\sin\psi.$$

[4] Pressley (2010, p. 230) also provides this physical explanation.



[11.8] *A unit-mass particle travels at unit speed along a geodesic* g *on the pseudosphere of radius* $R = 1$. *By Clairaut's Theorem, its angular momentum* $\Omega = \rho \sin\psi$ *is constant. But since the Beltrami–Poincaré map is* conformal, *the angle* ψ *is preserved, and it follows that the image* G *in the map has equation* $y = \frac{1}{\Omega}\sin\psi$, *which is a semicircle of radius* $(1/\Omega)$ *meeting the horizon* $y = 0$ *at right angles.*

Thus, using exactly the same reasoning as in [5.8], page 60, we conclude that,

> *If a unit-mass particle travels at unit speed along a geodesic* g *on the pseudosphere, and its angular momentum about the axis of symmetry is* Ω, *then its image in the Beltrami–Poincaré upper half-plane travels along a semicircle meeting* $y = 0$ *at right angles. Furthermore, the radius of this semicircle is* $(1/\Omega)$. *In other words, the (Euclidean) curvature of the semicircle is the angular momentum of the particle.*

Lastly, let h be highest point that the particle can reach before its angular momentum Ω forces it to head back down the pseudosphere. Clearly, h is mapped to the highest point H on the image G of g, both h and H corresponding to $\psi = (\pi/2)$. It follows from (5.1) that the arc length σ_{max} along the segment of the tractrix generator (not shown) going from the rim straight up to h, is given by the logarithm of the radius of G:

$$\sigma_{max} = \ln \frac{1}{\Omega}.$$

Chapter 12

The Extrinsic Curvature of a Surface

12.1 Introduction

We have seen how the two principal curvatures (together with their associated principal directions) characterize the extrinsic geometry of a surface in great detail, via Euler's curvature formula, (10.1). But is there is *single number* (without any associated direction) that can characterize the overall extrinsic geometry of a surface at a point, in the same way as the Gaussian curvature $\mathcal{K}$ characterizes the intrinsic geometry?

To characterize the overall extrinsic geometry we must presumably take some kind of *average* of the principal curvatures. The two most obvious ways of doing this are to take their arithmetic mean, $\frac{\kappa_1+\kappa_2}{2}$, or else their geometric mean, $\sqrt{\kappa_1\kappa_2}$. *Both* of these averages are geometrically natural and extremely important.

The arithmetic mean is usually denoted by the letter H (or $\overline{\kappa}$) and it is simply called the ***mean curvature***:

$$\boxed{H = \overline{\kappa} \equiv \tfrac{\kappa_1+\kappa_2}{2}.} \tag{12.1}$$

Look again at the significance in [10.2] of H to the graph of Euler's curvature formula: it is the centre about which the curvature oscillates sinusoidally. This mean curvature H is fundamental to understanding the shape of so-called ***minimal surfaces***, which include the shapes of all possible soap films spanning a complicated curved wire frame. These minimal surfaces must, by definition, satisfy $H = 0$ at every point, so that $\kappa_2 = -\kappa_1$, and the surface is saddle-shaped. As noted in the Prologue, we shall not explore these fascinating surfaces in this work; instead, we refer you to the *Further Reading* section, where several excellent works are recommended.

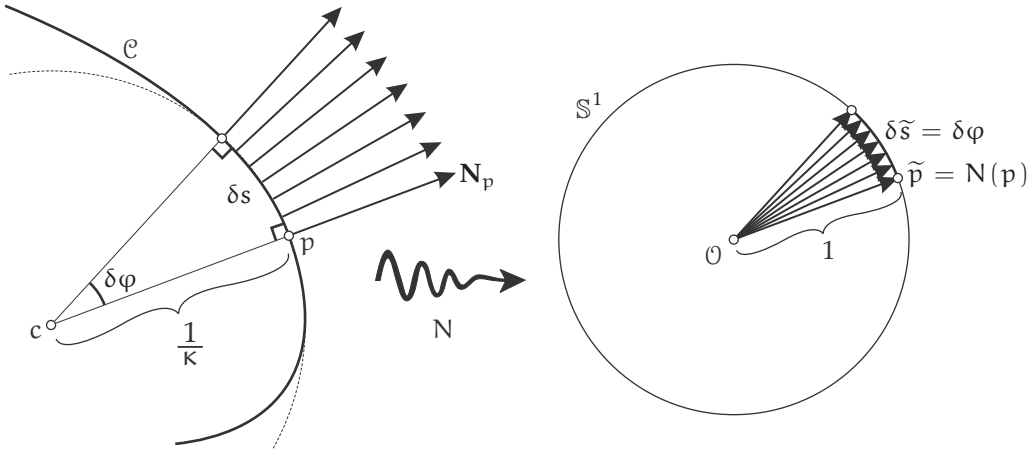
The geometric meaning of $\sqrt{\kappa_1\kappa_2}$, or rather its *square*, $\kappa_1\kappa_2$, turns out to be even more fundamental than that of H, but we shall deliberately keep you in suspense a little longer as to what that significance might be. Forgive us, but even the great Gauss declared it to be one of the greatest punchlines in all of mathematics, so a few pages of drum-roll is in order. (If you cannot bear to wait, jump ahead to the answer: (13.3) on p. 142.)

12.2 The Spherical Map

As we know, the curvature κ of a plane curve $\mathcal{C}$ can be defined as the rate of turning of its tangent. Equivalently, κ may be viewed as the rate of turning of the normal. This latter interpretation will permit us to generalize this extrinsic definition of curvature from curves to surfaces, for the latter also admit a unique ($\pm$) normal vector.

First, however, it will be helpful if we reinterpret the rate of turning of a plane curve's normal vector as the *spread* of the directions of the normal vectors that occurs over a short piece of the curve, say of length δs.

To quantify this, imagine taking each of these unit normal vectors and transplanting them so that they all emerge from a common point $\mathcal{O}$. See [12.1]. In this way, the normal $\mathbf{N}$ may be viewed



[12.1] *The map* N *is defined to send the point* p *on* $\mathcal{C}$ *to the point on the unit circle lying in the same direction as* $\mathbf{N}_p$*. Then* $\kappa \asymp (\delta\varphi/\delta s)$ *can be reinterpreted as the local length magnification factor of* N*, measuring the rate of spreading of the normals.*

as a mapping N from the point p of the plane curve $\mathcal{C}$ to the the point $\widetilde{p} = N(p)$ on the unit circle $\mathbb{S}^1$ (centred at $\mathcal{O}$) that lies at the tip of $\mathbf{N}_p$.

Over the segment δs of $\mathcal{C}$, the directions of the normal vectors are spread over angle $\delta\varphi$, and therefore the tips of these normal vectors fill an arc of length $\delta\widetilde{s} = \delta\varphi$ on the unit circle. The local spreading of the normal vectors can now be quantified by considering the local magnification of arc length under the normal map:

$$\kappa = \text{local length magnification factor of the N map} \asymp \frac{\delta\widetilde{s}}{\delta s}. \tag{12.2}$$

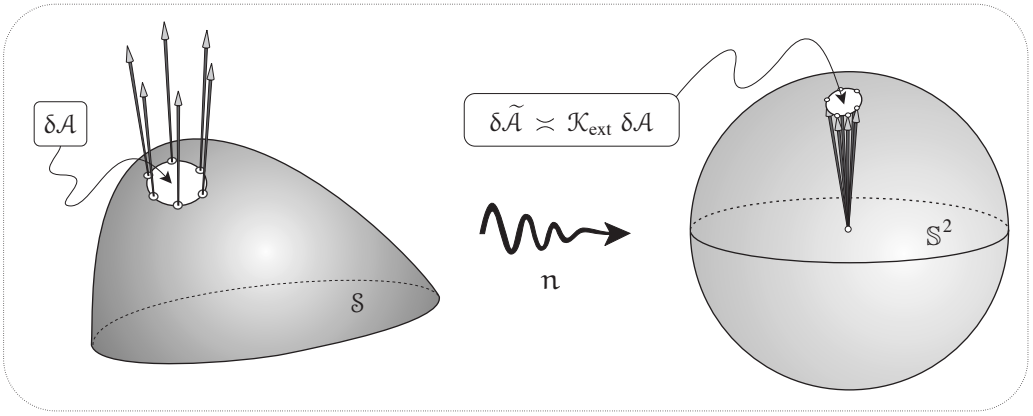
This suggests a way to generalize the construction to a surface $\mathcal{S}$. Consider a small patch of $\mathcal{S}$ of area $\delta\mathcal{A}$ and containing the point p, the normal vector there being $\mathbf{n}_p$. See [12.2]. By analogy with [12.1], we introduce the ***spherical map***—most commonly called the ***Gauss map***, or the ***normal map***—from the surface to the unit sphere ($\mathfrak{n} : \mathcal{S} \to \mathbb{S}^2$) sending the point p to the point $\widetilde{p} = \mathfrak{n}(p)$ on the unit sphere lying in the same direction as $\mathbf{n}_p$.

HISTORICAL NOTE ON TERMINOLOGY: In essentially all other texts, the "spherical map" is instead called the "Gauss map," but this is historically inaccurate. Yes, Gauss did indeed publish it in 1827 (and privately used it years earlier), but it was Olinde Rodrigues (1795–1851) (a French banker and amateur mathematician) who first published this concept in 1815, employing it in a penetrating study of the curvature of surfaces. In recognition of this fact, Marcel Berger (one of the foremost geometric authorities of the late twentieth century) calls it the ***Rodrigues–Gauss map***, which seems to us to strike an appropriate balance. That said, we shall generally prefer the term "spherical map,"[1] by virtue of its clarity and brevity.

12.3 Extrinsic Curvature of Surfaces

We are thus led to a brand new *extrinsic* measure of surface curvature in terms of the spread of the normal vectors, which we shall temporarily denote $\mathcal{K}_{\text{ext}}$.

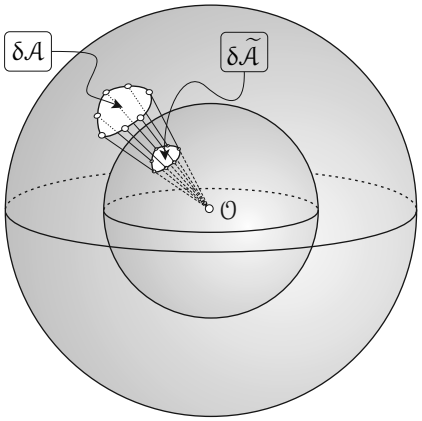
[1] A strong precedence argument can also be made for "spherical map": it was the name preferred by such legendary figures as Hilbert (1952) and Hopf (1956). It has likewise been employed by I. M. Singer, V. A. Toponogov, and other famous differential geometers of the modern era.



[12.2] *The extrinsic curvature $\mathcal{K}_{ext}$ is the local area magnification factor of the spherical map: $\mathcal{K}_{ext} \asymp \frac{\delta\widetilde{\mathcal{A}}}{\delta\mathcal{A}}$.*

In [12.2] we shrink the small shape down towards p and define

$$\boxed{\mathcal{K}_{\text{ext}} \equiv \text{local area magnification factor of the spherical map} \asymp \frac{\delta\widetilde{\mathcal{A}}}{\delta\mathcal{A}}.} \tag{12.3}$$



[12.3] *If $\mathcal{S}$ is a sphere of radius R centred at $\mathcal{O}$, and the image S^2 under n is imagined to be concentric to it, then n is simply radial projection from $\mathcal{O}$. Thus distances within $\mathcal{S}$ are compressed by $(1/R)$ and areas on $\mathcal{S}$ are compressed by $(1/R)^2$.*

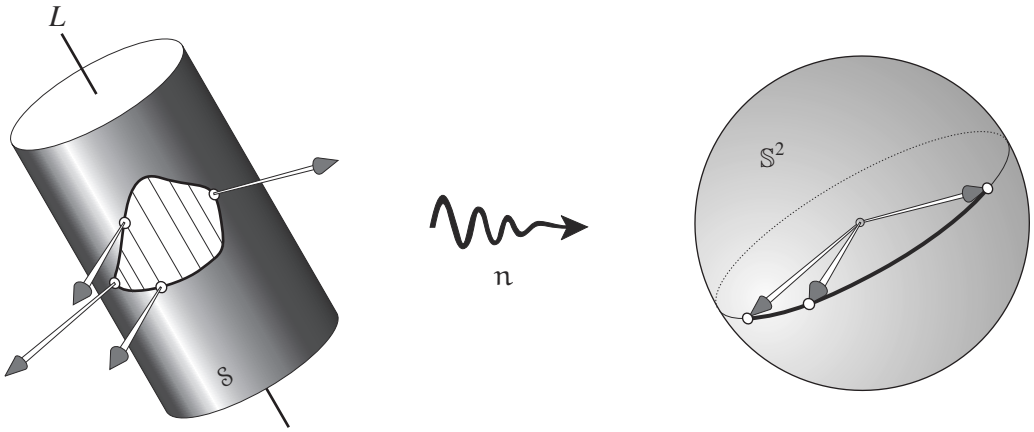
For example, suppose $\mathcal{S}$ is a sphere of radius R centred at $\mathcal{O}$, and picture the image S^2 under the spherical map as having the same centre. Then n is simply radial projection from $\mathcal{O}$, as illustrated in [12.3]. Linear dimensions clearly shrink by $(1/R)$ and areas therefore shrink by $(1/R)^2$, so

$$\boxed{\textit{The sphere of radius } R \textit{ has extrinsic curvature } \mathcal{K}_{\text{ext}} = (1/R^2).} \tag{12.4}$$

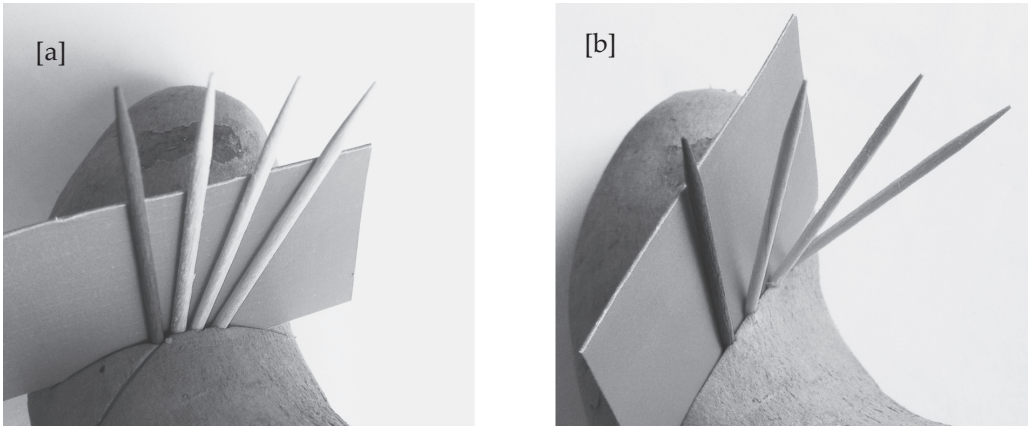
As a second example, let $\mathcal{S}$ be a cylinder of radius R with axis L, as illustrated in [12.4]. All points on the same generator of the cylinder have the same normal vector and hence the same image under the spherical map. Since all the normals of $\mathcal{S}$ are perpendicular to L their images under n lie on the great circle of S^2 that lies in the plane perpendicular to L. Thus any area on $\mathcal{S}$ is crushed down by n to an arc of this great circle, having zero area. Note that the same crushing of generators also occurs on a *cone*, thus

$$\boxed{\textit{The cylinder and the cone both have } \mathcal{K}_{\text{ext}} = 0.} \tag{12.5}$$

In contrast to (12.2), it is not immediately obvious that the magnification factor in (12.3) is uniquely defined in the general case. However, later we shall be able to show that it really is: all



[12.4] *The spherical map* $\mathfrak{n}$ *crushes each generator of the cylinder (parallel to its axis* L*) to a single point on the great circle perpendicular to* L*, so* $\mathcal{K}_{ext} = 0$.



[12.5] [a] *If we move in a principal direction then the normal vector tips in that direction and initially stays within the normal plane;* **[b]** *If we move in a general direction then the normal vector immediately tips out of the normal plane.*

infinitesimal areas at a given point undergo the *same* magnification, regardless of their shape. For now, let us assume this and seek a specific shape that will reveal an explicit formula for $\mathcal{K}_{\text{ext}}$.

To help us find a special shape for which the area magnification factor is geometrically self-evident, two lemmas are needed, the first of which is this:

> *As* p *begins to move in a principal direction,* $\mathbf{n}_{\mathrm{p}}$ *remains within the normal plane* Π *in that direction.* (12.6)

A bit more precisely, we mean by this that if ζ is defined to be the angle between $\mathbf{n}$ and Π, so that $\zeta(\mathrm{p}) = 0$, then $\dot{\zeta}(\mathrm{p}) = 0$. This follows from the local mirror symmetry of $\mathcal{S}$ in Π, for if $\mathbf{n}$ were to immediately rotate out of Π to one side or the other ($\dot{\zeta} > 0$ vs. $\dot{\zeta} < 0$) this would violate the symmetry (10.4).

Fig. [12.5a] illustrates (12.6) on a yam. On the same yam, and at the same point, [12.5b] illustrates the fact that if we instead move off in a *general* direction then $\dot{\zeta}(\mathrm{p}) \neq 0$ and $\mathbf{n}$ immediately

rotates out of Π. We strongly encourage you to conduct your own such experiments, using whatever fruits or vegetables are readily available.

If we think of the left-hand side of [12.2] as a cross-section of $\mathcal{S}$, drawn in the plane Π, and the right-hand side as a cross section of $\mathcal{S}^2$ in the plane parallel to Π through $\mathcal{O}$, the second lemma follows immediately:

If p *moves along a short vector* $\mathbf{v}_i$ *in the* i^{th} *principal direction of* $\mathcal{S}$, *then* $n(p)$ *ultimately moves along* $-\kappa_i \mathbf{v}_i$ *on* $\mathcal{S}^2$. (12.7)

Note that the minus sign in this formula is due to our conventions: in [12.5a] $n(p)$ moves in the *same* direction as $\mathbf{v}_i$, a *positive* multiple of $\mathbf{v}_i$, but (by our convention) here $\kappa_i < 0$, because the normal section bends *away* from our chosen $\mathbf{n}$.

We are thus guided to consider the fate under the spherical map of a small rectangle whose sides are aligned with the principal directions. Let the lengths of these sides be ϵ_1 and ϵ_2, along the first and second principal directions, respectively. By virtue of (12.7), the spherical map ultimately sends this rectangle to another *rectangle* on $\mathcal{S}^2$, parallel to the original, but with the sides ultimately stretched to $\kappa_1\epsilon_1$ and $\kappa_2\epsilon_2$. Thus,

$$\delta\widetilde{\mathcal{A}} \asymp (\kappa_1\epsilon_1)(\kappa_2\epsilon_2) \asymp (\kappa_1\kappa_2)\,\delta\mathcal{A},$$

and therefore (12.3) yields

$$\mathcal{K}_{\text{ext}} = \kappa_1\kappa_2. \tag{12.8}$$

For example, on a sphere of radius R, $\kappa_1 = \kappa_2 = (1/R)$, yielding $\mathcal{K}_{\text{ext}} = (1/R^2)$, in agreement with (12.4). Note that if the principal curvatures of a general surface have the same sign, then we obtain the following interpretation of their geometric mean: the extrinsic curvature of the surface is the same as that of a sphere of radius $1/\sqrt{\kappa_1\kappa_2}$.

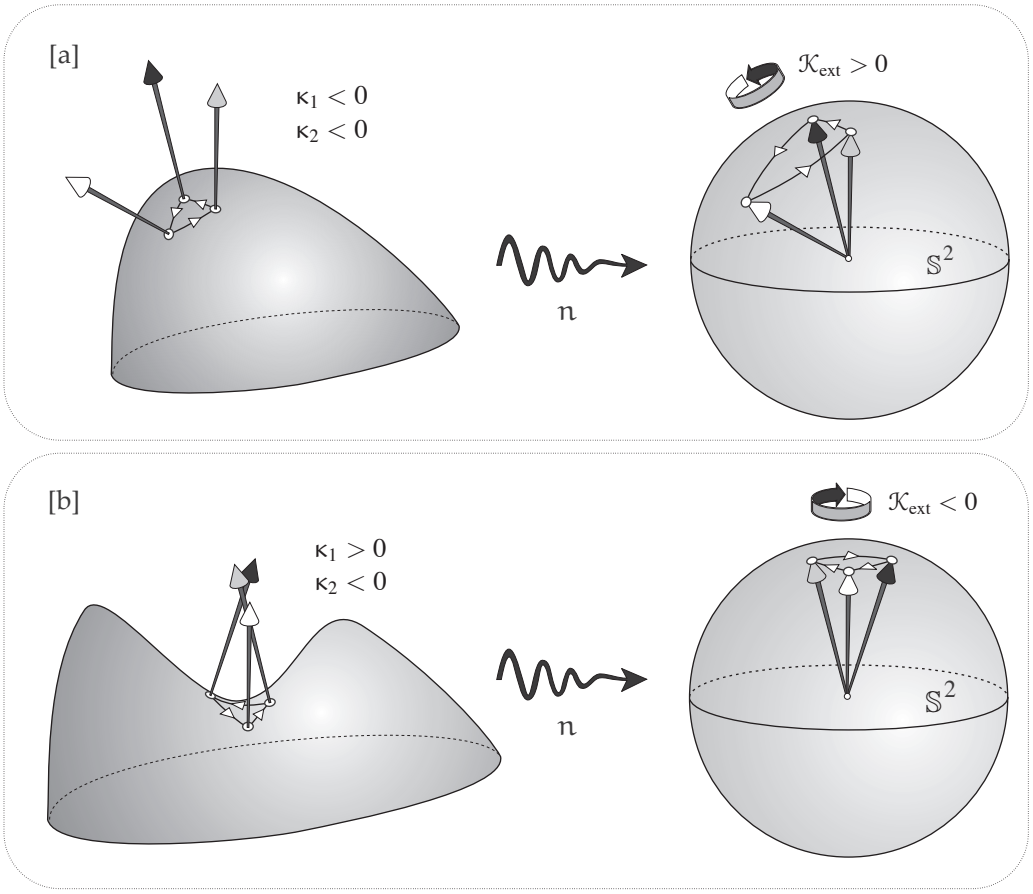
As another example, a cylinder of radius R has $\kappa_1 = (1/R)$ and $\kappa_2 = 0$, yielding $\mathcal{K}_{\text{ext}} = 0$, in agreement with (12.5). A cone also has one vanishing principal curvature, and hence $\mathcal{K}_{\text{ext}} = 0$ in this case, too.

HISTORICAL NOTE: Both the extrinsic definition (12.3) of curvature and the explicit formula (12.8) are almost universally attributed to Gauss (1827). However, as with the spherical map (aka the Rodrigues–Gauss map) itself, both of these insights were in fact first published by Rodrigues in 1815, twelve years prior to Gauss. It appears that Gauss (along with most twentieth-century mathematicians!) was simply unaware of Rodrigues's discoveries. For more on this history, see Kolmogorov and Yushkevich (1996, p. 6) and Knoebel (2007, p. 118).

The formula (12.8) attaches a *sign* to $\mathcal{K}_{\text{ext}}$: if p is elliptic (i.e., κ_1 and κ_2 have the same sign) then $\mathcal{K}_{\text{ext}} > 0$; if it is hyperbolic (i.e., κ_1 and κ_2 have opposite signs) then $\mathcal{K}_{\text{ext}} < 0$; and if it is parabolic (i.e., one of the κ_i vanishes) then $\mathcal{K}_{\text{ext}} = 0$.

In order to make sense of this sign of $\mathcal{K}_{\text{ext}}$ in terms of its original definition (12.3), we take $\delta\mathcal{A}$ to be positive and we use a simple geometric property of the spherical map to attach a sign to $\delta\widetilde{\mathcal{A}}$, as follows.

Imagine that, with $\mathbf{n}$ pointing at your eye, you see the boundary of $\delta\mathcal{A}$ on $\mathcal{S}$ traced counterclockwise, as illustrated in [12.6]. The idea is to attach a sign to the area $\delta\widetilde{\mathcal{A}}$ on $\mathcal{S}^2$ according to whether the spherical map preserves or reverses the orientation of the boundary. That is, we take $\delta\widetilde{\mathcal{A}}$ to be *positive* if its boundary is traced in the *same* counterclockwise sense (as seen from outside the sphere) that the original was, and we take it to be *negative* if it is traced in the *opposite*, clockwise sense.



[12.6] *The sign of $\mathcal{K}_{ext}$ depends on whether the spherical map n preserves orientation (***[a]***: $\mathcal{K}_{ext} > 0$) or reverses it (***[b]***: $\mathcal{K}_{ext} < 0$).*

12.4 What Shapes Are Possible?

Let us analyze systematically what shapes are possible for an arbitrary surface, at least *locally*.

In the immediate vicinity of a generic point, we have seen that the surface is described by the quadratic in (10.6) and repeated here:

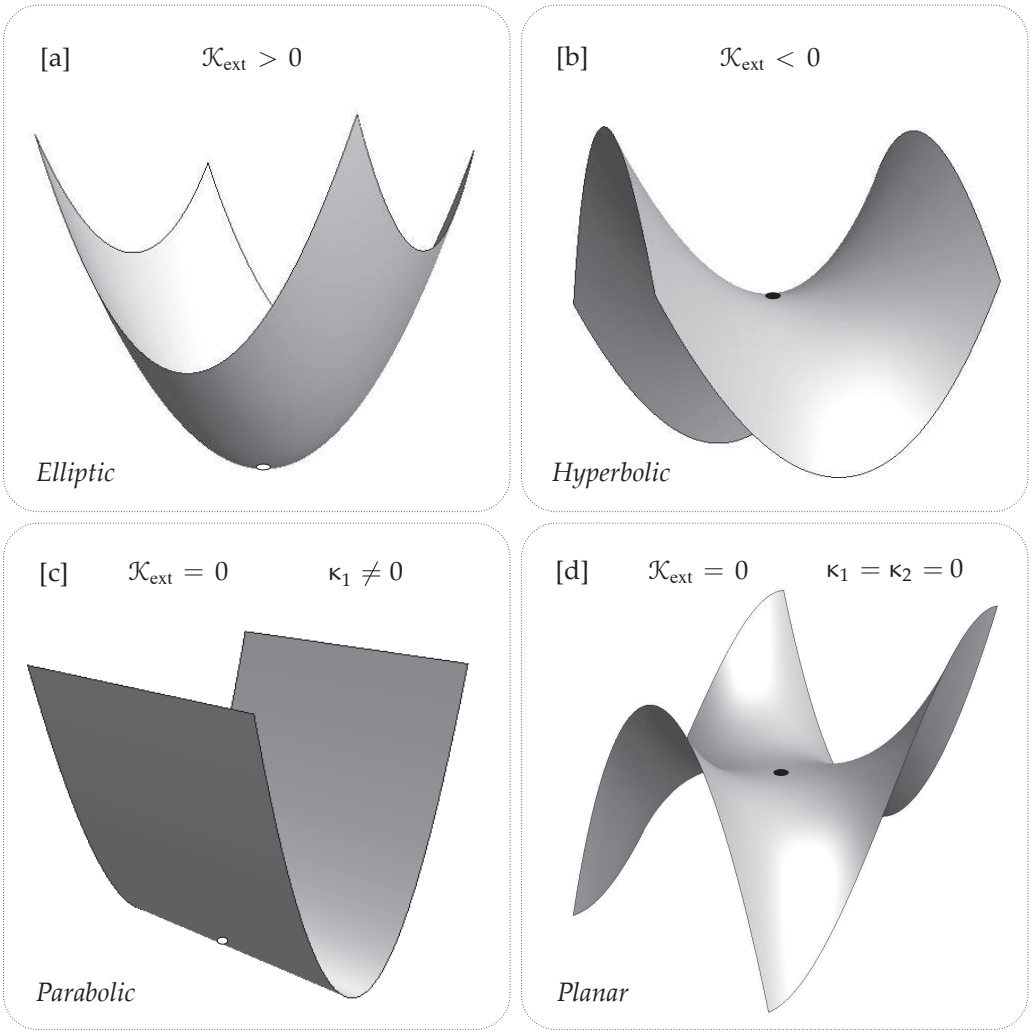
$$z \asymp \tfrac{1}{2}\kappa_1 x^2 + \tfrac{1}{2}\kappa_2 y^2.$$

In general, neither principal curvature vanishes, so $\mathcal{K}_{\text{ext}} = \kappa_1\kappa_2 \neq 0$. Recall that a point is called elliptic if $\mathcal{K}_{\text{ext}} > 0$, and called hyperbolic if $\mathcal{K}_{\text{ext}} < 0$. The shape of the surface in the vicinity of such a generic point is *completely* determined by this distinction regarding the *sign* of $\mathcal{K}_{\text{ext}}$:

> *If $\mathcal{K}_{ext} > 0$ then the surface is locally a bowl, as in [12.7a].*
>
> *If $\mathcal{K}_{ext} < 0$ then the surface is locally a saddle, as in [12.7b].*

(12.9)

The remaining case $\mathcal{K} = 0$ can arise in two different ways: either one principal curvature vanishes, or both do. In the former case, the point is called parabolic, and in the latter case it is called ***planar***, for reasons that will become clear shortly.



[12.7] *The local shape of the surface is* **[a]** *a bowl if* $\mathcal{K}_{ext} > 0$*;* **[b]** *a saddle if* $\mathcal{K}_{ext} < 0$*;* **[c]** *a trough if* $\mathcal{K}_{ext} = 0$ *(and only one of* κ_1 *and* κ_2 *vanishes). If* both *principal curvatures vanish, then the surface can be arbitrarily complex;* **[d]** *illustrates just one possibility, called the* **monkey saddle**.

In the former parabolic case, suppose that $\kappa_1 \neq 0$ and $\kappa_2 = 0$. Then the surface is locally given by $z \asymp \frac{1}{2}\kappa_1 x^2$, which is a trough (with a parabolic cross section) running in the y-direction, as illustrated in [12.7c].

The surface of a doughnut provides examples of all three cases thus far considered. Every point on the outer half of the doughnut has positive curvature, and looks like [12.7a], and every point on the inner half of the doughnut has negative curvature, and looks like [12.7b]. If we imagine the doughnut sitting on a plate, the circle of contact with the plate separates the two halves of opposite curvature, and is made up of parabolic points where $\mathcal{K}_{\text{ext}} = 0$. Sure enough, the strip of surface surrounding this circle is indeed a trough, like [12.7c]. For a diagram that focusses on this circular trough, look ahead to [13.3a], page 141.

In the final "planar" case, $\mathcal{K} = 0$ again, but now *both* principal curvatures vanish. Since $\kappa_1 = \kappa_2 = 0$, we see that z must be ultimately be equal to a *cubic* (or higher-powered) homogeneous polynomial in x and y. It makes sense that such a point is called planar, because the curvature of the normal section vanishes in every direction, like a plane, and the surface departs from its

tangent plane so slowly that it does indeed look very plane-like locally. But of course as we move further away, or simply look with greater precision, we can see the surface bending away from the tangent plane. As we shall now explain, the shape of the surface near such a planar point can be arbitrarily complicated.

To prepare ourselves for the more complicated behaviour that the surface can exhibit near a $\mathcal{K}_{ext} = 0$ planar point, let us first take a fresh look at a regular $\mathcal{K}_{ext} < 0$ saddle. For simplicity's sake, let us assume that $\kappa_1 = -\kappa_2 = 2$. Now let us think of the tangent plane as the *complex* plane, so that each point within it can be described (with Cartesian and polar coordinates) by the complex number $x + iy = r\,e^{i\theta}$. Then,

$$\text{Height of saddle} = \left(x^2 - y^2\right) = \text{Re}\,(x+iy)^2 = \text{Re}\,\left[r^2\,e^{i2\theta}\right] = r^2\cos 2\theta.$$

This new form of the height formula makes it transparent that as we circle once around the origin, the surface does *two* complete oscillations in height. If we picture an actual saddle on a horse, then the "valley" of the saddle on one side of the horse corresponds to one of these two complete oscillations of height, and it accommodates one of the rider's legs, while the second complete oscillation provides the valley on the other side, for the other leg. And directly in front and behind the rider, the saddle rises up in what we shall call "hills."

Now let us construct a saddle suitable for a *monkey*, who has two legs and a *tail*; note that the official, technical name for this surface is indeed the ***monkey saddle***! To create the three equally spaced valleys the monkey requires, we need only *cube* our complex number:

$$\text{Height of monkey saddle} = \text{Re}\,(x+iy)^3 = \text{Re}\,\left[r^3\,e^{i3\theta}\right] = r^3\cos 3\theta.$$

This monkey saddle is illustrated in [12.7d].

Note that the three valleys are now equally spaced, $(2\pi/3)$ apart, and the three hills are likewise equally spaced in between the valleys: each hill is directly opposite a valley. As you can see, this means that the normal sections necessarily have an *inflection point* at the origin, implying (without calculation) that the curvature must vanish there. Since the normal curvature vanishes in every direction, the monkey saddle does indeed have a planar point at its centre.

It is now easy to see that there exists an infinite menagerie [*sic*] of such saddles, of ever increasing complexity, of which the monkey saddle may be described as the 3-saddle. For example, if we want to create a ***cat saddle***—which we must confess is *not* standard terminology!—we need only replace the third power of the monkey saddle with the *fifth* power, to create the 5-saddle, with height given by $r^5\cos 5\theta$.

Note that it is visually clear that all these saddles have strictly *negative* curvature everywhere except at the planar point itself, where $\mathcal{K}_{ext} = 0$. In fact it can be shown (see Ex. 20) that the curvature $\mathcal{K}_{ext}$ of the general n-saddle is *symmetrical* around the origin: it only depends on r and therefore has a constant negative value on each circle $r = \textit{constant}$.

We have not yet exhausted all possible shapes surrounding a planar point. While all the higher-order saddles have negative curvature surrounding the planar point of vanishing curvature, it is *also* possible to have the planar point be surrounded by a sea of *positive* curvature. For example, consider [exercise] the almost flat-bottomed bowl with equation $z = r^4 = (x^2 + y^2)^2$.

Chapter 13

Gauss's *Theorema Egregium*

13.1 Introduction

In 1827 Gauss announced the ***Theorema Egregium***—Latin for "Remarkable Theorem." Although that year witnessed the death of Beethoven, the appearance of the *Theorema Egregium* meant that it also witnessed the birth of modern Differential Geometry.

From this result sprang fundamental advances in both mathematics and physics, some of which we have touched on already, and some of which must await future chapters. It paved the road to Beltrami's crucial 1868 step in the acceptance of Hyperbolic Geometry, interpreting it as the intrinsic geometry of a saddle-shaped surface of constant negative Gaussian curvature. And Riemann's brilliant generalization of Gauss's intrinsic curvature to higher-dimensional manifolds in turn enabled Einstein in 1915 to give precise mathematical form to his supremely beautiful and supremely accurate General Theory of Relativity, in which gravitation is understood as the curvature impressed upon the intrinsic geometry of space and time by matter and energy.

But in order to properly appreciate the *Theorema Egregium* itself, we shall first describe its origins. The fascinating and insightful work of Dombrowski (1979) uses Gauss's private notebooks, letters to friends, and official publications to carefully piece together a chronology of Gauss's evolving insights into Differential Geometry in general, and into this theorem, in particular. As Dombrowski explains, Gauss's epic explorations began in 1816, when he made a *nonlocal* discovery about the curvature of surfaces that was profound and unexpected in equal measure.

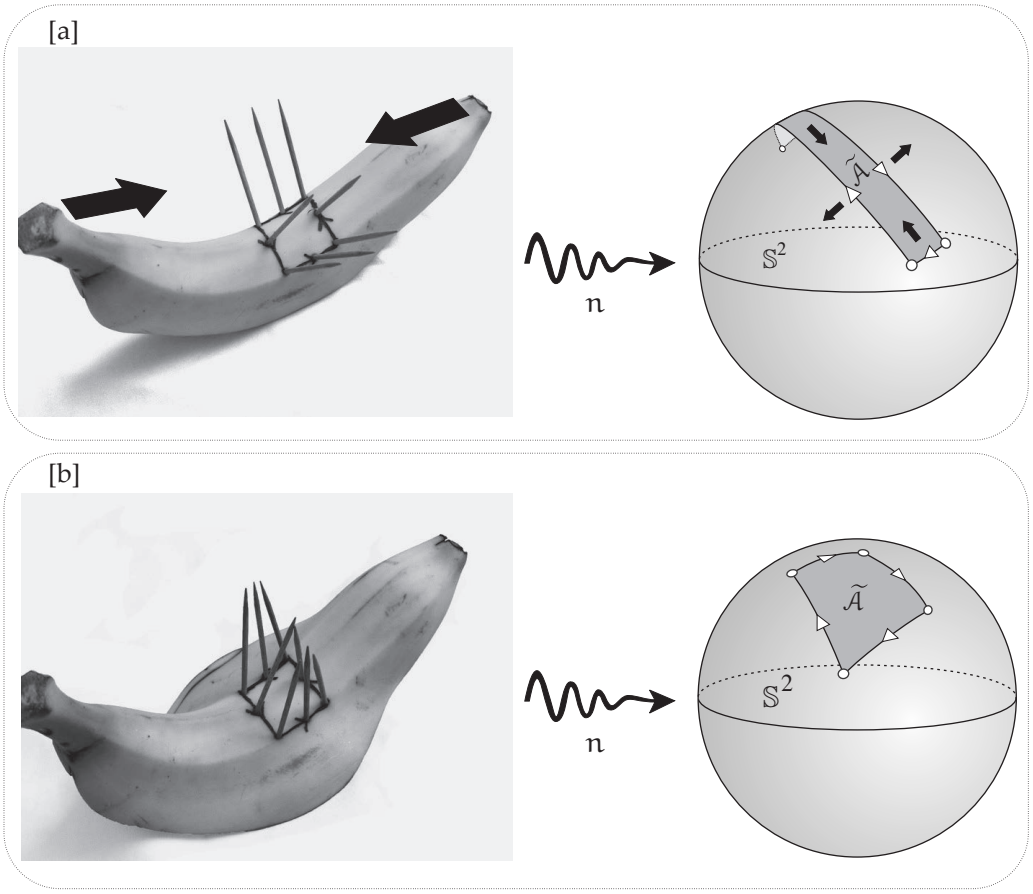
13.2 Gauss's *Beautiful Theorem* (1816)

Gauss was not given to gushing. But if he was stingy in his praise of others, then at least he was not a hypocrite. A succession of his major discoveries lay hidden in his private notebooks till after his death because he deemed them insufficiently perfected to be worthy of publication. Indeed, his Latin motto was *pauca, sed matura*, "few, but ripe." One such unpublished discovery was the nonlocal result of 1816.

Agonizingly, even in his private notebook[1] Gauss left no trace of what led him to suspect the result, nor how he proved it, but with uncharacteristic exuberance he was moved to *name* it:

> "**Beautiful Theorem.** *If a curved surface on which a figure is fixed takes different shapes in space, then the surface area of the spherical image of the figure is always the same.*" (13.1)

[1] (Gauß, 1973, p. 372)



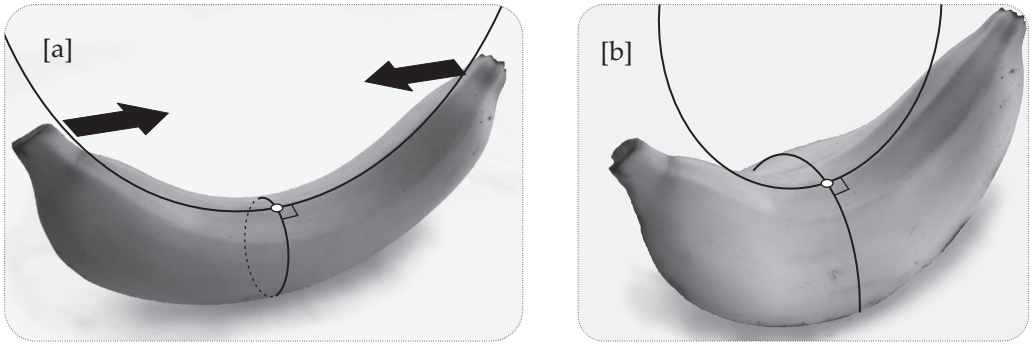
[13.1] *Gauss's* **Beautiful Theorem***:* **[a]** *Under the spherical map,* $\mathfrak{n}$*, the squarish loop on the banana is mapped to a narrow quadrilateral of area* $\widetilde{\mathcal{A}}$ *on* $\mathbb{S}^2$*, traversed in the opposite sense to the original, because of the negative curvature. Bending the banana skin by pulling the ends towards each other distorts the image on* $\mathbb{S}^2$*, yielding* **[b]***: The image on* $\mathbb{S}^2$ *is now a squarer quadrilateral, but* the area $\widetilde{\mathcal{A}}$ is exactly the same as before!

What Gauss means here by the "spherical image" is the image on $\mathbb{S}^2$ under the spherical map. Figure [13.1] illustrates the meaning of the theorem with a banana peel; we strongly encourage you to try this (or similar) experiment yourself. First, we have peeled one strip of peel from the positive-curvature side and have removed the banana itself, leaving the remainder of the peel surface intact. On the negatively curved part of peel we have drawn a squarish, counterclockwise loop, and on this loop we have erected normals (toothpicks).

Under the spherical map this loop is mapped to a narrow quadrilateral on $\mathbb{S}^2$ of area $\widetilde{\mathcal{A}}$, traversed in the *opposite* sense to the original, as illustrated in [13.1a]. Make sure to follow the loop on the banana with your eye, confirming that the tip of the normal does indeed trace such a reversed image on $\mathbb{S}^2$.

If we now pull the ends of the banana towards each other, the skin undergoes an isometry, and the normals along the boundary of our patch deform to produce a squarer image on $\mathbb{S}^2$. According to the Beautiful Theorem, this new spherical image has *exactly the same area as before!*

Such experiments provide exciting and tangible manifestations of the underlying mathematical truth, (13.1), but they *explain* nothing. In Act IV we shall introduce the concept of ***parallel transport***, and with its assistance we shall be able to provide an elegant conceptual *explanation*



[13.2] *Gauss's* **Theorema Egregium***:* **[a]** *The two circles of curvature at the illustrated point on the banana have radii of curvature ρ_1 and ρ_2, and the extrinsic curvature there is $\mathcal{K}_{ext} = 1/\rho_1\rho_2$. Bending the banana skin by pulling the ends towards each other yields* **[b]***: One circle expands while the other contracts, but* the product of their radii remains constant*: $\mathcal{K}_{ext}$ is invariant.*

of the Beautiful Theorem; for now, though, we will simply assume it, and thereby deduce as its consequence the *Theorema Egregium*.

13.3 Gauss's *Theorema Egregium* (1827)

Gauss did not publish a single word on the subject for another *decade*, but in private he returned to Differential Geometry and to his Beautiful Theorem many times, most intensely in 1822 and 1825, even writing and then abandoning a full-length manuscript.[2] At last, in 1827 he was satisfied, publishing the result as the centrepiece of his *Disquisitiones Generales Circa Superficies Curvas* ("General Investigations of Curved Surfaces"[3]), and he allowed pent-up excitement to get the better of him. What he had privately described to himself as "beautiful" he now announced to the world as "remarkable"—the ***Theorema Egregium***.

But by this point Gauss had almost totally covered his tracks from the original nonlocal discovery of 1816, and the form of the result that he presented to the world was purely *local*. To see how he was led to this new local version of the result, take the figure occurring in the Beautiful Theorem and simply shrink it down towards a point p.

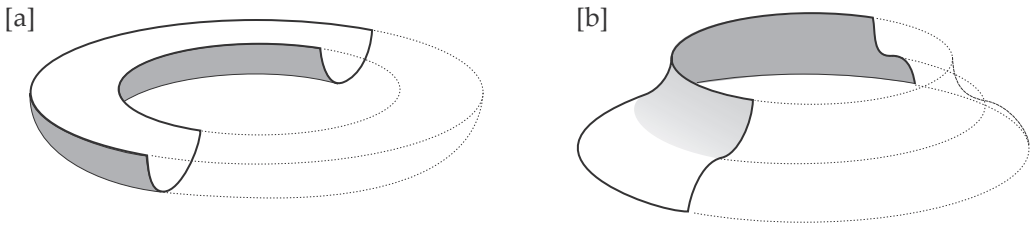
Let $\delta\mathcal{A}$ be the original area on $\mathcal{S}$ surrounding p and let $\delta\widetilde{\mathcal{A}}$ be the image area on $\mathbb{S}^2$ surrounding $n(p)$. Naturally, $\delta\mathcal{A}$ is invariant under isometries of $\mathcal{S}$, and by virtue of the Beautiful Theorem (13.1) $\delta\widetilde{\mathcal{A}}$ is *also* invariant. The "remarkable theorem" follows immediately:

> **Theorema Egregium.** *The extrinsic curvature $\mathcal{K}_{ext} \asymp (\delta\widetilde{\mathcal{A}}/\delta\mathcal{A})$ is invariant under isometries of $\mathcal{S}$ and therefore belongs to the* intrinsic *geometry of $\mathcal{S}$. More explicitly, while the principal curvatures individually depend on the shape of the surface in space, their product does* not*: $\kappa_1\kappa_2$ is invariant under isometries.* (13.2)

Figure [13.2] illustrates this with another banana peel, and again we urge you to try this experiment yourself. Pull the ends of the banana toward each other, and watch what happens at a

[2] An English translation of the aborted 1825 manuscript is appended in (Gauss 1827).

[3] English translations are available in both (Dombrowski 1979) and (Gauss 1827).



[13.3] *When the outer half of the circular trough in* **[a]** *is reflected across the plane upon which it rests, we obtain the isometric surface* **[b]**. *Yet [a]* cannot *be continuously deformed into [b] without stretching the surface. Nevertheless, the* Theorema Egregium *assures us that the two surfaces have equal curvature at corresponding points.*

particular point p: the radius of curvature ρ_1 shrinks (for the principal direction running the length of the banana), while the cross-sectional radius of curvature ρ_2 expands. But according (13.2) the product $\rho_1\rho_2$ remains perfectly constant throughout the bending.

To confirm this experimentally, take two small lengths of fairly stiff wire and bend them to fit the surface of the banana in its natural state (before removing the fruit inside) in the two principal directions at a particular point, thereby obtaining two small pieces of the principal circles of curvature there. Now lay these two small pieces of circle flat on a table and use a ruler to estimate their radii ρ_1 and ρ_2, and hence their product $\rho_1\rho_2$. Now have a friend bend the banana peel and hold it steady while you again fit the pieces of wire to the new surface at the same point as before, in the new principal directions.[4] Finally, confirm that the new value of $\rho_1\rho_2$ is the same as before, within the limits of experimental error.

The word "bending" implies continuous deformation, but this is not actually required by the theorem: there do exist isometries that cannot be carried out via gradual, continuous isometric deformation, but which nevertheless preserve the curvature by virtue of the theorem.[5]

To illustrate this we shall describe an example of Aleksandrov (1969, p. 101). Figure [13.3] is a freehand copy of Aleksandrov's original picture. Figure [13.3a] depicts a circular trough, which we imagine to rest on a plane, touching it along a circle C. To obtain the surface in [13.3b], cut the surface into two along C, then reflect the outer half in the plane. Clearly this new surface is isometric to the original surface, yet it also intuitively clear (and can be proven) that the original surface is rigid and cannot be bent into the new shape without stretching the surface in the process. Note that if we had instead performed an analogous transformation of a straight trough (a half-cylinder) then the new surface *could* have been obtained by a continuous deformation, without stretching.

A word about bending of physical "surfaces" versus mathematical ones. A physical surface, no matter how thin, cannot actually be bent without any stretching. For example, take a piece of paper and roll it into a cylinder, joining the edges together with tape. Both sides of the sheet began with the same length, but the cylinder has an outer circumference that is very slightly greater than the inner one: the outside had to be *stretched*, creating tension within the material. It is for this reason that when you remove the tape, the sheet will spontaneously spring back to it original planar form. Only in the mathematical limit of vanishing thickness can a surface be bent without any stretching.

Gauss's result (13.2) is indeed remarkable, but he took it still further. Since the result shows that $\mathcal{K}_{\text{ext}}$ is actually an *intrinsic* measure of curvature, it is natural to ask how it might be related

[4] We deliberately chose a point and a deformation such that the principal directions did not change. However, in general the principal directions will spin within the surface as the bending occurs.

[5] It is not clear if Gauss himself was aware of this distinction, but it is certainly true that he expressed the *Theorema Egregium* in terms of isometries, rather than bendings. It was only later authors who paraphrased Gauss as saying that $\mathcal{K}_{\text{ext}}$ was "invariant under bending."

to our original intrinsic definition (2.1) (see p. 18) of Gaussian curvature $\mathcal{K}$, as the angular excess per unit area: $\mathcal{K} \asymp \mathcal{E}(\Delta)/\mathcal{A}(\Delta)$. Gauss's answer is *very* remarkable:

> *The extrinsically defined curvature* $\mathcal{K}_{ext} = \kappa_1\kappa_2$ *and the intrinsically defined Gaussian curvature* $\mathcal{K}$ *are numerically* equal*:*
>
> $$\mathcal{K}_{ext} = \mathcal{K}.$$

(13.3)

In light of this result, we can and will drop the distinction between these two measures of curvature, enabling us to speak simply of *the* curvature $\mathcal{K}$ of the surface.

As with the Beautiful Theorem itself, the simplest and most general proof of this wonderful result must await the introduction of the concept of ***parallel transport*** in Act IV. However, in the next chapter we shall be able to lend some plausibility to the result via a more limited argument concerning polyhedra.

In the meantime, let us confirm the validity of (13.3) for some specific surfaces for which we have already calculated both the extrinsic principal curvatures *and* the intrinsic Gaussian curvature.

- *The Cylinder and the Cone.* The intrinsically flat plane may be rolled into a cylinder or a cone, so the Gaussian curvature of these two surfaces vanishes. Combining this fact with (12.5),

$$\mathcal{K}_{\text{ext}} = 0 = \mathcal{K}.$$

- *The Sphere.* Combining the results (12.4) and (1.3) (p. 8),

$$\mathcal{K}_{\text{ext}} = (1/R^2) = \mathcal{K}.$$

- *The Pseudosphere.* Combining the results (10.9) and (5.3) (p. 53),

$$\mathcal{K}_{\text{ext}} = -(1/R^2) = \mathcal{K}.$$

- *The Torus.* Combining the results (10.10) and Exercise 23 on page 89,

$$\mathcal{K}_{\text{ext}} = \frac{1}{r(r + R\sec\alpha)} = \mathcal{K}.$$

These five surfaces are so important, both historically and mathematically, that we have afforded them the dignity of individual treatments, but in fact the truth of (13.3) for all five surfaces follows, in one fell swoop, from this:

- *The General Surface of Revolution.* Combining the results (10.11), (10.12), and Exercise 22 on page 89, in which a particle travelled along the generating curve at *unit speed*,

$$\mathcal{K}_{\text{ext}} = -\ddot{y}/y = \mathcal{K}. \tag{13.4}$$

Chapter 14

The Curvature of a Spike

14.1 Introduction

Thus far we always imagined our surface to be perfectly smooth, with a well-defined tangent plane and associated normal vector at each point. But consider [14.1], which shows the spiked surface of the durian,[1] here shown with one of its interior yellow fruits exposed.

Regardless of whether we approach this surface from the extrinsic or the intrinsic point of view, how on Earth shall we define the curvature at the tip of one of the durian's many spikes?!

Once we have answered this question we shall be able to shed some new light on the *Theorema Egregium*.

[14.1] *How can we measure the curvature of the spiked surface of the durian?*

14.2 Curvature of a Conical Spike

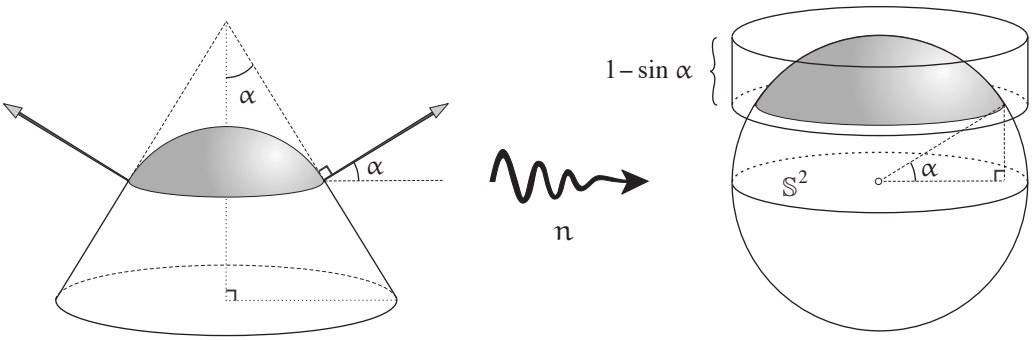
Suppose our spike takes the form of the tip of a cone. As we discussed in the last section, the curvature of the cylinder and the cone both vanish, at least at every point other than the tip of the cone. While the vanishing intrinsic curvature of the cylinder is easy to accept, we balk at the cone's "flatness," and with good reason.

Imagine yourself as a two-dimensional inhabitant of the cone's surface, standing at the vertex. You construct a circle centred there of radius r and measure its circumference $C(r)$. You quickly discover that $C(r)$ is *less than* the Euclidean prediction of $2\pi r$. This discrepancy clearly signals the presence of curvature, and this curvature must reside *at* the tip, and nowhere else, for the curvature at typical points has been shown to vanish.

To discuss this in more detail, suppose that the internal angle from the axis of symmetry to the cone's surface is α. See [14.2]. Intuitively, any reasonable measure of curvature should increase as α becomes smaller and the spike becomes sharper; conversely, it should diminish to zero as α approaches $\pi/2$, for in that limit the cone flattens out into a Euclidean plane.

Unlike a mathematical cone, the tip of a physical cone will not be perfectly sharp, but will instead be slightly blunt and rounded. We may gain some mathematical insight if we now imagine this rounded tip to be in the form of a polar cap of a sphere that fits smoothly onto the cone, in the

[1] This southeast Asian delicacy is justifiably known as "The King of Fruits": it tastes (and smells!) as wonderfully strange as it looks.



[14.2] *The spherical map sends the blunted tip of the cone to the similar polar cap on* $\mathbb{S}^2$*, the area of which equals that of the illustrated cylinder of height* $(1-\sin\alpha)$.

sense that there is a well-defined tangent plane along the join between the cone and the polar cap. See [14.2].

As the figure illustrates, the spherical map merely expands this blunted tip into a similar polar cap on the unit sphere. It now seems reasonable to (provisionally) *define* the curvature of the spiked tip of the original mathematical cone to be the total curvature residing within the blunted tip of the physical cone. For it is clear that this definition does not depend on the *size* of the blunted tip: as we shrink its radius, making it sharper and sharper,[2] the size of the spherical image on the unit sphere does not change. Furthermore, even without calculation, it seems clear that the dependence on α of this measure of curvature conforms to our previously described intuition.

Thus we make this definition:

$$\mathcal{K}(\text{spike}) \equiv \text{total } \mathcal{K} \text{ of blunted tip} = \text{area of the (polar cap) spherical image.}$$

It is also possible to pass directly from the *original* (unblunted) cone tip to this same spherical image. Imagine a horizontal plane resting on the tip of the cone, and imagine a unit normal vector attached to the plane, pointing straight up. Taking this unit normal to be the $\mathbf{n}$ of our spherical map, this horizontal position of the plane is mapped by n to the north pole on the right of [14.2]. The plane is free to rock in any direction until it hits the side of the cone and becomes tangent to the surface, at which point n lies on the boundary of the polar cap. Thus, as the plane resting on the tip assumes all possible positions, n fills the same polar cap as before. We call this the ***generalized spherical map***.

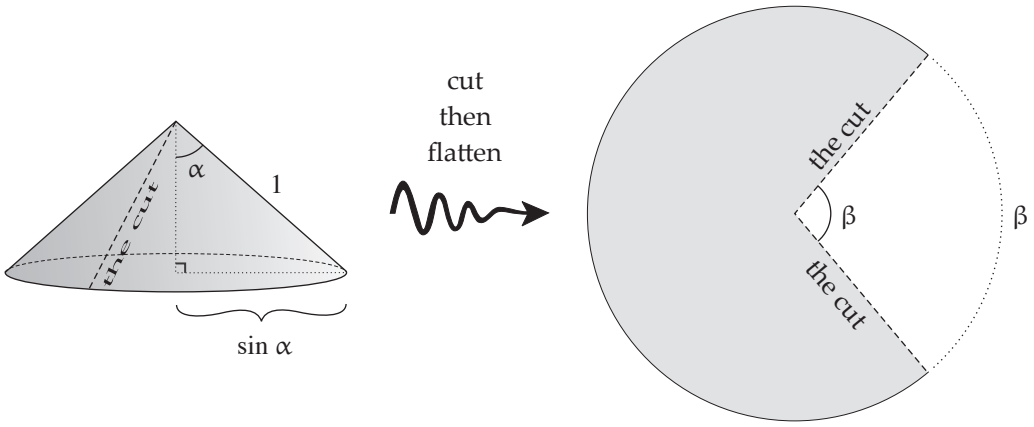
Regardless of how we arrive at this spherical image, let us now find a formula for its area. By virtue of a result by Archimedes (see Ex. 10, page 85), the area of the polar cap equals the area of its illustrated projection onto the cylinder touching the unit sphere along its equator. Thus, since the cylinder has circumference 2π, and this segment of it has height $(1-\sin\alpha)$, then

$$\mathcal{K}(\text{spike}) = \text{area of spherical image} = 2\pi(1-\sin\alpha). \tag{14.1}$$

Note that this formula for the curvature does indeed achieve its maximum when $\alpha = 0$, and it also vanishes when $\alpha = (\pi/2)$, as anticipated.

There is another way way of looking at this result, a way that will shortly offer a natural generalization of curvature to *polyhedral* spikes. See [14.3]. Suppose we cut one unit along a generator

[2] You would not hesitate to press your palm down onto the blunted tip when it's 10-cm wide, but would not wish to when it's 0.01-cm wide!



[14.3] *When the cone is cut by one unit along a generator and flattened out, the base circumference* $2\pi \sin\alpha$ *becomes the arc* $(2\pi - \beta)$. *It follows that* $\mathcal{K}(\textit{spike}) = \beta$.

of the cone, starting at the tip (the vertex), then cut horizontally around the cone. If we flatten this out we obtain a sector of a unit circle, bounded by the two sides of the cut, now split apart by angle β. Clearly β is also a measure of curvature of the original spike. In fact we will now show that it is exactly the *same* measure of curvature.

On the left of [14.3] we see that the circumference of the base is $2\pi \sin\alpha$. But this length does not change when the cone is cut and flattened to produce the sector of the unit circle on the right. Thus,

$$2\pi \sin\alpha = 2\pi - \beta.$$

Therefore (14.1) can be re-expressed as

$$\mathcal{K}(\text{spike}) = \beta = \text{Split angle of flattened spike.} \tag{14.2}$$

Both of these formulas (14.1) and (14.2) for $\mathcal{K}_{\text{spike}}$ employ extrinsic ideas that are unknowable to the inhabitants of the cone: the first involves the 3-dimensional concept of the normal to the surface, while the second involves the flattening of the surface within three-dimensional space. Nevertheless, it is in fact a simple matter to reinterpret the second formula *intrinsically*.

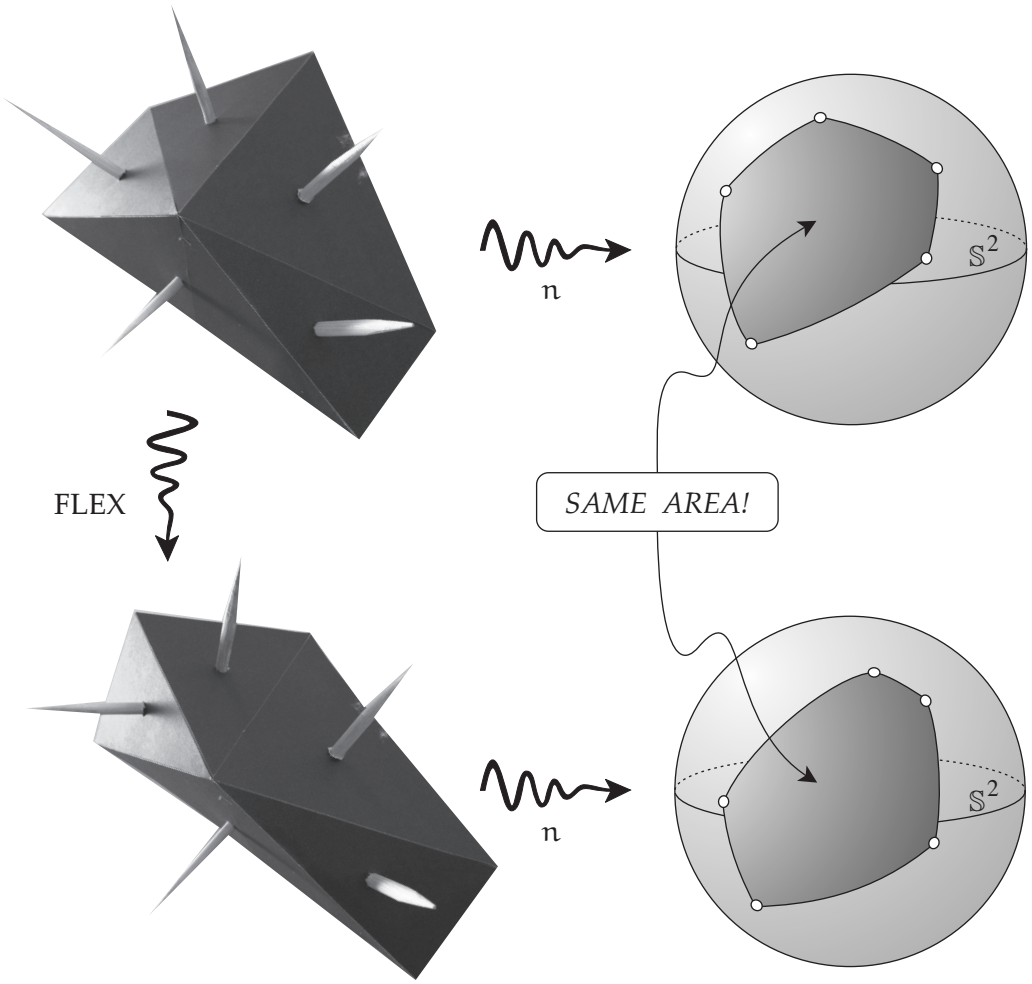
Returning to the discussion at the start of the section, the circumference $C(r)$ of a circle of radius r centred at the vertex is related to the curvature by [exercise] the following *intrinsic* formula for β:

$$\mathcal{K}(\text{spike}) = \frac{2\pi r - C(r)}{r} = 2\pi - C(1). \tag{14.3}$$

14.3 The Intrinsic and Extrinsic Curvature of a Polyhedral Spike

In the case of the cone we have just seen that there is a natural way of measuring the curvature of the spike both intrinsically and extrinsically, and these two measures turn out to *coincide*. As we now explain, the same is true of the vertex v of a polyhedral spike, at which we shall suppose that m flat, polygonal faces $f_1, f_2, \ldots, f_m$ come together, the outward normal of f_j being $\mathbf{n}_j$.

The generalization of the last two interpretations (14.2) and (14.3) of $\mathcal{K}(\text{spike})$ is straightforward, so we begin there. As with the cone, suppose we cut down along one of the edges, starting



[14.4] Polyhedral Theorema Egregium. *The generalized spherical map sends the vertex of the polyhedral spike to a geodesic polygon P_m on S^2, and as the polyhedral spike is flexed, this spherical polygon changes shape, but* its area does not change! *Compare this with the picture of the flexing banana skin in [13.1] on page 139.*

at v and ending at an arbitrary point p, then cut a sideways circuit around v across all the f_j, finally returning to p, thereby cutting off the spike from the polyhedron.

Imagine the polyhedron to made up of stiff cardboard polygons, attached along their edges with sticky tape, so that each edge acts like a hinge. (Such a model is shown in [14.4].) Then, since it has been cut along vp, the detached spike may now be flattened out onto the plane, the two sides of the cut vp ending up split apart by angle β. If θ_j is the angle between the edges of f_j that meet at v, and $\Theta \equiv \sum \theta_j$ is the sum of the angles that meet at v, then clearly $\beta = 2\pi - \Theta$. Lastly, this can be reexpressed *intrinsically* as the circumference $C(1)$ of a unit circle centred at v drawn within the original polyhedral surface:

$$\mathcal{K}_{\text{int}}(v) = \text{split angle of flattened spike} = \beta = 2\pi - \Theta = 2\pi - C(1). \tag{14.4}$$

The extrinsic definition of curvature can also be generalized, via the generalized spherical map. As we did with the cone, imagine a plane Π (with normal $\mathbf{n}$) resting in contact with the spike. As Π rocks and assumes all possible positions, all the while remaining in contact with the spike, what image region does n fill on S^2?

The limits of motion of Π are determined by the faces: we can rock Π until it hits and coincides with a face f_j, at which point $\mathbf{n}$ coincides with $\mathbf{n}_j$, the normal to f_j. The limits of Π's motion are thus determined by the m points n_j on S^2. Next, observe that

> *As the plane Π rocks over the vertex v, the spherical image n on S^2 fills the interior of the m-gon P_m on S^2 with vertices n_j.*

To see this, we need only imagine Π initially coincident with f_j, then rolling over the edge e in which f_j and f_{j+1} meet. As it does so, $\mathbf{n}$ must swing within the plane perpendicular to e, and therefore the image n on S^2 travels along the arc of the great circle connecting n_j to n_{j+1}. The geodesic arcs connecting the successive n_j therefore form the boundary of the region that n covers, as Π rocks over all possible positions.

We have thus arrived at a natural conception of the *extrinsic* curvature of v:

$$\mathcal{K}_{\text{ext}}(v) = \text{area on } S^2 \text{ of } m\text{-gon connecting the polyhedron's normals.} \tag{14.5}$$

14.4 *The Polyhedral Theorema Egregium*

Just as happened with the cone, the intrinsic and the extrinsic measures of curvature of the polyhedral spike turn out to *coincide*, a result which we may reasonably call the ***Polyhedral Theorema Egregium***:

$$\mathcal{K}_{\text{ext}}(v) = \mathcal{K}_{\text{int}}(v). \tag{14.6}$$

Figure [14.4] illustrates the theorem in action: as the spike is flexed, and the normals all rotate in different directions, the spherical image P_m changes, but, according to the theorem, its area is actually intrinsic to the spike, and therefore cannot change! Compare this with the picture of the flexing banana skin in [13.1] on page 139.

To understand this, consider [14.5]. The explanation, in a nutshell, is that the area of the spherical polygon only depends on its *angles*, and the figure demonstrates that these are in turn determined by the intrinsic angles of the polyhedron's faces.

On the left, two edges of the polyhedron meet at angle θ, and the planes perpendicular to these edges meet at angle $\widetilde{\theta}$. It follows from the marked right angles that $\theta + \widetilde{\theta} = \pi$. Thus, under the generalized spherical map, the corresponding edges of the spherical polygon meet at angle $\widetilde{\theta} = \pi - \theta$.

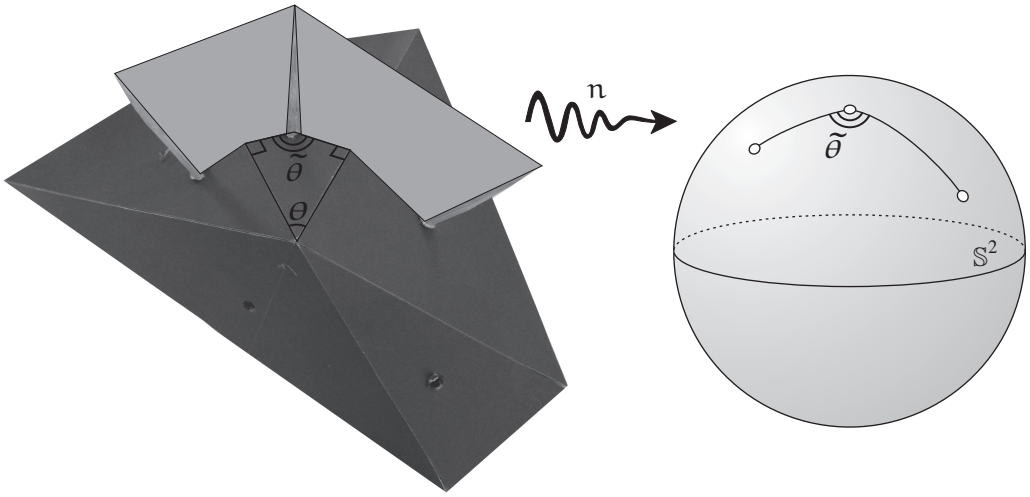
We can now verify the result in detail. Recall that in Exercise 5 on page 83 we generalized Harriot's Theorem from triangles to polygons, showing that the area $\mathcal{A}(P_m)$ of a spherical m-gon P_m is once again equal to its angular excess:

$$\mathcal{A}(P_m) = \mathcal{E}(P_m) \equiv [\text{angle sum}] - (m-2)\pi.$$

Therefore,

$$\mathcal{K}_{\text{ext}}(v) = \mathcal{A}(P_m) = 2\pi - \sum_{i=1}^{m} [\pi - \widetilde{\theta}_i] = 2\pi - \sum_{i=1}^{m} \theta_i = 2\pi - \Theta = \mathcal{K}_{\text{int}}(v),$$

and the Polyhedral *Theorema Egregium* is proved.



[14.5] *On the left, two edges of the polyhedron meet at angle* θ*, and the planes perpendicular to these edges meet at angle* $\tilde{\theta}$*. It follows from the marked right angles that* $\theta + \tilde{\theta} = \pi$*. Thus, under the generalized spherical map, the corresponding edges of the spherical polygon on the right meet at angle* $\tilde{\theta} = \pi - \theta$*.*

HISTORICAL NOTES: The elegant insight above is usually attributed to Hilbert, appearing[3] in the timeless masterpiece, *Geometry and the Imagination* (Hilbert 1952, p. 195), the German original having appeared in 1932. But in fact[4] the same observation was made in 1854 by the great British physicist James Clerk Maxwell, 78 years prior to Hilbert. Maxwell's original 1854 letter to William Thomson can be found in Maxwell (2002, Vol. 1, p. 243), and his eventual paper of 1856 appears in Maxwell (2003, Vol. 1, §4).

Finally, we note that these concepts can be extended to *non*convex polyhedra and to vertices having *negative* curvature. For a beautiful and completely general investigation, see Banchoff (1970).

[3]However, Hilbert does not invoke our rocking plane. Instead, his justification for the transition from the n_j to P_m consists only in this: "In order to relate this to the spherical representation of surfaces, we connect the points $[n_j]$ ... by arcs of great circles ... [to create P_m]"

[4]I stumbled upon this little-known fact entirely by accident while leafing through my copy of Maxwell's collected works (Maxwell 2003), in search of something quite unrelated.

Chapter 15

The Shape Operator

15.1 Directional Derivatives

The Gaussian curvature $\mathcal{K}$, viewed extrinsically, measures the variation of the surface normal $\mathbf{n}$ in the vicinity of a point p, but only in a blurred, average way. Instead of looking at the overall spread of directions of $\mathbf{n}$ over a small patch of the surface $\mathcal{S}$ containing p, we now turn to a more precise method of quantifying the variation of $\mathbf{n}$, by looking at how fast it changes as we move away from p within $\mathcal{S}$ in a *specific direction*.

Let $\hat{\mathbf{v}}$ be a unit vector emanating from p and lying within the tangent plane T_p to $\mathcal{S}$ at p. See [15.1]. We would like to define the rate of change of $\mathbf{n}$ in the direction $\hat{\mathbf{v}}$. If we move a small distance ϵ away from p in this direction then we arrive at a point q whose position vector is therefore $\mathbf{q} = \mathbf{p} + \epsilon\hat{\mathbf{v}}$. It is then tempting to try to define the rate of change of $\mathbf{n}$ as

$$\lim_{\epsilon \to 0} \frac{\mathbf{n}(q) - \mathbf{n}(p)}{\epsilon}.$$

But this will not do. For q actually resides within T_p, not $\mathcal{S}$, and therefore $\mathbf{n}(q)$ is not even defined.

Nevertheless, it seems clear that the concept we are groping for really is well-defined: we need only move a distance ϵ *within the surface*, instead of the tangent plane, but we must do so in the desired *direction* $\hat{\mathbf{v}}$. One way of achieving this end is to drop a perpendicular from q to $\mathcal{S}$, meeting it at r, say. It seems clear (and can be proved) that the length pr of the geodesic segment from p to r is ultimately equal to ϵ:

$$pr \asymp \epsilon.$$

Returning to our abortive definition of rate of change, we can now define

$$\boxed{\delta\mathbf{n} \equiv \mathbf{n}(r) - \mathbf{n}(p)}$$

to be the change in $\mathbf{n}$ that results from moving distance ϵ (ultimately) within the surface in the direction $\hat{\mathbf{v}}$. Since $\mathbf{n}$ has unit length, it merely *rotates* slightly as we move from p to r. Therefore, if we picture both $\mathbf{n}(p)$ and $\mathbf{n}(r)$ as having a common origin, as illustrated in [15.2], then the movement of its tip is ultimately *orthogonal* to $\mathbf{n}$, and therefore $\delta\mathbf{n}$ is a vector that ultimately lies within the tangent plane, T_p.

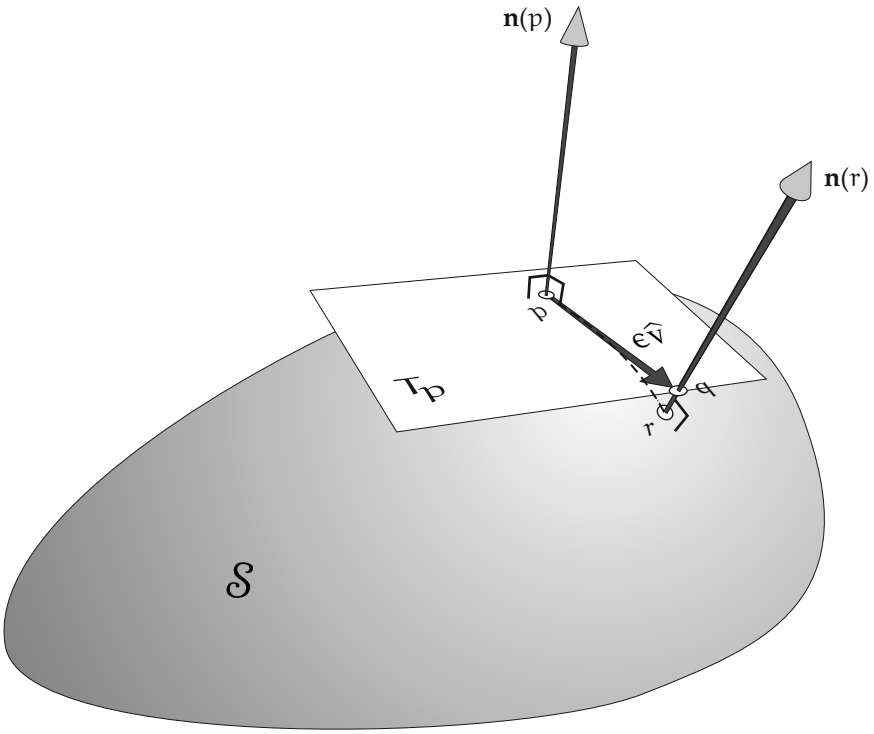
Finally, the ***directional derivative*** $\nabla_{\hat{\mathbf{v}}}$ of $\mathbf{n}$ along $\hat{\mathbf{v}}$ can then be defined as

$$\boxed{\nabla_{\hat{\mathbf{v}}}\mathbf{n} \equiv \lim_{\epsilon \to 0} \frac{\delta\mathbf{n}}{\epsilon}.} \tag{15.1}$$

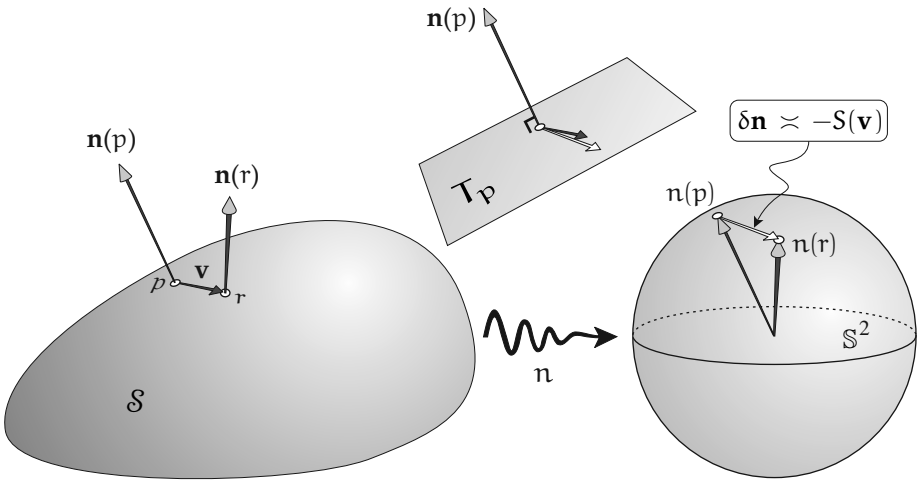
As we have explained, this derivative lives in T_p.

This is the directional derivative per *unit* length. If instead we want the derivative along a general tangent vector $\mathbf{v} = v\hat{\mathbf{v}}$ of length v, then we must simply scale up by this length:

$$\nabla_{\mathbf{v}}\mathbf{n} = \nabla_{v\hat{\mathbf{v}}}\mathbf{n} = v\nabla_{\hat{\mathbf{v}}}\mathbf{n}. \tag{15.2}$$



[15.1] *To find the derivative $\nabla_{\hat{\mathbf{v}}}\mathbf{n}$ of $\mathbf{n}$ in the direction $\hat{\mathbf{v}}$, we consider the change $\delta\mathbf{n} \equiv \mathbf{n}(r) - \mathbf{n}(p)$ that results from moving in that direction a distance ϵ within the surface. Then $\nabla_{\hat{\mathbf{v}}}\mathbf{n} \asymp \frac{\delta\mathbf{n}}{\epsilon}$.*



[15.2] *The* **Shape Operator** *applied to a short tangent vector $\mathbf{v}$ emanating from p in T_p tells us how much the normal changes from the tail to the tip of $\mathbf{v}$. This change $\delta\mathbf{n} \asymp -S(\mathbf{v})$ ultimately lies in the parallel tangent plane to $\mathbb{S}^2$ at $n(p)$. Superimposing the two tangent planes, S may be viewed as a linear transformation of T_p to itself, here depicted detached from $\mathcal{S}$, floating between the two surfaces.*

This makes greater intuitive sense if we think of $\mathbf{v}$ as the *velocity* of a particle moving over the surface, at the moment it passes through p. Then a unit vector $\hat{\mathbf{v}}$ corresponds to a particle moving at unit speed. Thus when $\nabla_{\hat{\mathbf{v}}}$ is applied to *any* quantity defined along the particle's trajectory (not necessarily $\mathbf{n}$) it yields the rate of change of that quantity with respect to *time*. Thus if the particle were to travel three times as fast along the same trajectory, changing $\hat{\mathbf{v}}$ to $3\hat{\mathbf{v}}$, then the quantity would change three times as fast along the trajectory. This is one way of looking at (15.2). Note that this concept of the directional derivative can be applied to *any* quantity (vector or scalar) that is defined on the surface, or merely along the trajectory.

There is another way of looking at the definition of the directional derivative that requires, first, that we extend the concept of "ultimate equality" to vectors. Thus, suppose that two vectors $\mathbf{a}$ and $\mathbf{b}$ depend on a small quantity ϵ that tends to zero. An obvious definition of $\mathbf{a} \asymp \mathbf{b}$ would be that each *component* of $\mathbf{a}$ is ultimately equal to the corresponding component of $\mathbf{b}$. But this definition is problematic [why?] if one or more components is identically zero. Therefore, instead, we adopt the following more satisfactory and geometrical definition, namely, that the magnitudes and directions of the two vectors are ultimately equal:

> *If two vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *depend on a small quantity* ϵ *that tends to zero, we define them to be* **ultimately equal**, *written* $\mathbf{a} \asymp \mathbf{b}$, *iff* $|\mathbf{a}| \asymp |\mathbf{b}|$ *and the angle between their directions has vanishing limit.*

Granted this, we may now rewrite (15.1) as

$$\nabla_{\hat{\mathbf{v}}}\mathbf{n} \asymp \frac{\delta\mathbf{n}}{\epsilon} \quad \Longleftrightarrow \quad \delta\mathbf{n} \asymp \epsilon\nabla_{\hat{\mathbf{v}}}\mathbf{n} = \nabla_{\epsilon\hat{\mathbf{v}}}\mathbf{n}.$$

Here, then, is the resulting point of view, which we shall employ *repeatedly* throughout the remainder of the book:

> *In the limit of vanishing* ϵ, *and hence vanishing* $\mathbf{v} = \epsilon\,\hat{\mathbf{v}}$, $\nabla_{\mathbf{v}}\mathbf{n}$ *is ultimately equal to the change* $\delta\mathbf{n}$ *in* $\mathbf{n}$ *from the tail to the tip of* $\mathbf{v}$*:* $\nabla_{\mathbf{v}}\mathbf{n} \asymp \delta\mathbf{n}$. (15.3)

Of course, as noted earlier, this statement employs a touch of poetic license, for the tip of $\mathbf{v}$ does not actually reside within $\mathcal{S}$. But as ϵ tends to zero, so does the amount of poetic license required! For, in this limit, the distinction between q and r in [15.1] rapidly becomes insignificant.

15.2 The Shape Operator S

The ***Shape Operator***[1] S associated with the surface at the point p tells us how the normal vector changes (in which direction it tips, and how fast) as we move away from p along an arbitrary tangent vector $\mathbf{v}$, which is no longer necessarily short, but rather has arbitrary length v. More precisely, it is simply defined to be the negative of the directional derivative of $\mathbf{n}$ along $\mathbf{v}$:

$$S(\mathbf{v}) \equiv -\nabla_{\mathbf{v}}\mathbf{n}. \tag{15.4}$$

[1]There are two other mathematical objects that encode exactly the same information as the Shape Operator: the ***Second Fundamental Form*** and the ***Weingarten map***. We shall not employ either in this book.

The insertion of the minus sign into the definition is standard; as we shall see shortly, it is motivated by our earlier definition of the sign of the curvature of normal sections.

The surface $\mathcal{S}$ and the sphere S^2 have the *same* normal vector $\mathbf{n}(p)$ at p and $n(p)$, respectively. Thus their tangent planes T_p and $T_{n(p)}$ at these two points are *parallel*, and we may imagine transporting them so as to make them coincide. Thus any tangent vector to S^2 emanating from $n(p)$ may instead be pictured as residing in T_p, emanating from p. In [15.2] we picture these two coincident planes floating between the two surfaces.

The meaning of the Shape Operator now becomes especially clear if we think of a point r very close to p being carried by the spherical map to a point $n(r)$ on S^2 close to $n(p)$. See [15.2]. If $\mathbf{v}$ is the very short vector connecting p to r and $\delta\mathbf{n}$ is the corresponding vector connecting $n(p)$ to $n(r)$ then, by virtue of (15.3),

$$S(\mathbf{v}) \asymp -\delta\mathbf{n}.$$

If we think of p as the origin of these two coincident planes, then S is in fact a *linear* transformation of T_p to itself. That is, if $\mathbf{v}$ and $\mathbf{w}$ are arbitrary tangent vectors at p, and a and b are arbitrary constants, then

$$\begin{aligned} S(a\mathbf{v}+b\mathbf{w}) &= -\left(\nabla_{a\mathbf{v}+b\mathbf{w}}\right)\mathbf{n} \\ &= -\nabla_{a\mathbf{v}}\mathbf{n} - \nabla_{b\mathbf{w}}\mathbf{n} \\ &= -a\nabla_{\mathbf{v}}\mathbf{n} - b\nabla_{\mathbf{w}}\mathbf{n} \\ &= a\,S(\mathbf{v}) + b\,S(\mathbf{w}). \end{aligned}$$

Recall from Linear Algebra that this means that S can be represented as a rectangular array of numbers $[S]$, called the *matrix* of S. In general, the j-th column of the matrix is the image of the j-th basis vector, or rather the numerical components of that vector. In our 2-dimensional case, this means that $[S]$ is a 2×2 square.

Later in this chapter we shall investigate the matrix $[S]$, but for now we illustrate this general idea of "matrixification" with an example we will need momentarily, namely, the matrix $[R_\theta]$ representing a rotation R_θ of the plane through angle θ about the origin.

Figure [15.3] illustrates the effect of R_θ on the basis vectors, from which we immediately deduce that

$$[R_\theta] = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}, \tag{15.5}$$

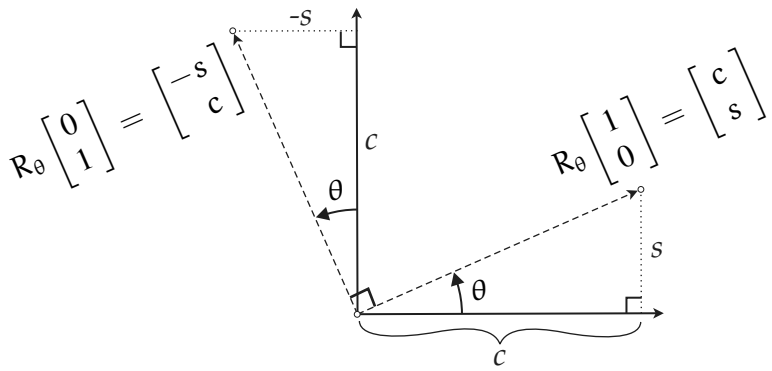
where $c \equiv \cos\theta$ and $s \equiv \sin\theta$.

15.3 The Geometric Effect of S

Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be unit vectors along the (orthogonal) first and second principal directions, respectively. What is the geometric effect of S on vectors in T_p when referred to this special orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2\}$? Our earlier result (12.7) provides the neat answer:

The principal directions are the **eigenvectors** *of the Shape Operator S, and the principal curvatures are the corresponding* **eigenvalues***:*

$$S(\mathbf{e}_i) = \kappa_i \mathbf{e}_i. \tag{15.6}$$



[15.3] *The first/second column of the matrix* $[R_\theta]$ *is the image of the first/second basis vector when it is rotated by* R_θ.

NOTE: It was precisely in order to cancel out the minus sign in (12.7) that a minus sign was introduced into the definition (15.4) of S.

Thus, knowing that the effect of S on the principal directions is to stretch them by their respective principal curvatures, linearity now reveals that the effect of S on a *general* tangent vector is to stretch it by these two factors in these two perpendicular directions.

If we choose our basis vectors to be $\{\mathbf{e}_1, \mathbf{e}_2\}$, then it follows that the matrix of S takes the especially simple, diagonal form,

$$[S] = \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix}. \tag{15.7}$$

Recall that a linear transformation expands the area of every shape by the *same* factor. This universal *area expansion factor* is the ***determinant*** of its matrix. This is particularly clear in the present case, where a unit square aligned with the principal axes is stretched into a rectangle of sides κ_1 and κ_2, and therefore the original square of unit area undergoes an expansion to area $|[S]| = \kappa_1\kappa_2$, so this is the expansion factor for *all* areas. Once again we recover our familiar expression for the extrinsic version of the Gaussian curvature:

$$\mathcal{K}_{\text{ext}} = \text{area expansion factor of the Shape Operator} = |[S]| = \kappa_1\kappa_2. \tag{15.8}$$

Recall that when we first derived this result (see (12.8), p. 134) we did so by considering the area expansion of one *particular* shape, and we merely *claimed* that the result was independent of this choice; the recognition of the Shape Operator's linearity has now proved that claim. Furthermore, given its geometric meaning, this area expansion factor (i.e., the determinant) must have the *same* value in all coordinate systems.

Notice that this matrix $[S]$ is ***symmetric*** (also called ***self-adjoint***), meaning that it has mirror symmetry across its ***main diagonal***, which runs top-left to bottom-right. Since reflection across this diagonal is achieved by swapping rows and columns, which creates (by definition) the ***transpose*** $[S]^T$, this symmetry can be written,

$$[S]^T = [S]. \tag{15.9}$$

The symmetry (15.9) of the matrix $[S]$ is no accident, but rather reflects (pardon the pun!) the underlying symmetry of the linear transformation S itself, and it will persist even if the basis is not aligned with the principal directions, in which case $[S]$ will no longer be diagonal.

15.4 DETOUR: The *Geometry* of the Singular Value Decomposition and of the Transpose

In this optional detour we seek to clarify the symmetry (15.9) of the Shape Operator by first asking the following question: *If a general linear transformation* M *is represented by the matrix* $[M]$*, what transformation* M^T *is represented by the transpose matrix* $[M]^T$*?* In order to answer this question, we shall begin by providing a geometric interpretation and derivation of one of the most crucial results in all of Linear Algebra, the so-called ***Singular Value Decomposition***[2]—or ***SVD*** for short.

The *geometric* interpretations and proofs we shall now present would appear to be appropriate for a first course in Linear Algebra. Yet, surprisingly, we have not been able to find these ideas described in *any* standard introductory text; indeed, we suspect that many working mathematicians may also be unfamiliar with them. Nevertheless, if you are in a hurry to proceed with Differential Geometry, please feel free to skip to the next section.

We begin by recalling the familiar fact that *if* a linear transformation of the plane has two real eigenvalues, its geometric effect is easily visualized, as follows: stretch by these factors along the two (typically nonorthogonal) eigenvector directions. But what if the transformation does *not* have real eigenvalues? A simple example of such a transformation (i.e., one that does not preserve any direction) is a *rotation*, but at least this too enjoys the virtue of being easily visualized. But how shall we make geometric sense of a general linear transformation that does not preserve any direction?

The SVD provides a wonderfully simple and vivid answer to this question, and it applies to *all* linear transformations; therefore even the transformations we thought we already understood (the ones that stretch in two nonorthogonal eigenvector directions) thereby receive new meaning.

> **Singular Value Decomposition (SVD).** *Every linear transformation of the plane is equivalent to stretching in two* orthogonal *directions (by generally different factors,* σ_1 *and* σ_2*, called the* **singular values***), followed by a rotation through angle* τ*, which we call the* **twist**. (15.10)

To understand this, consider the top half of [15.4], which shows the effect (from left to right) of a general linear transformation on an origin-centred circle C. Since the Cartesian equation of C is quadratic, the linear change of coordinates induced by the transformation will lead to another quadratic equation for the image curve. The image curve $\widetilde{C}$ is thus a conic section, and since the finite points of C are not sent to infinity, this conic must be an *ellipse*, shown top right of [15.4].

We have just used an algebraic statement of linearity; next we use the fundamental *geometric* fact is that it makes no difference if we add two vectors and then map the result, or if we map the vectors first and *then* add them. Convince yourself of these two simple consequences:

- *Parallel lines map to parallel lines.*
- *The midpoint of a line segment maps to the midpoint of the image line segment*.

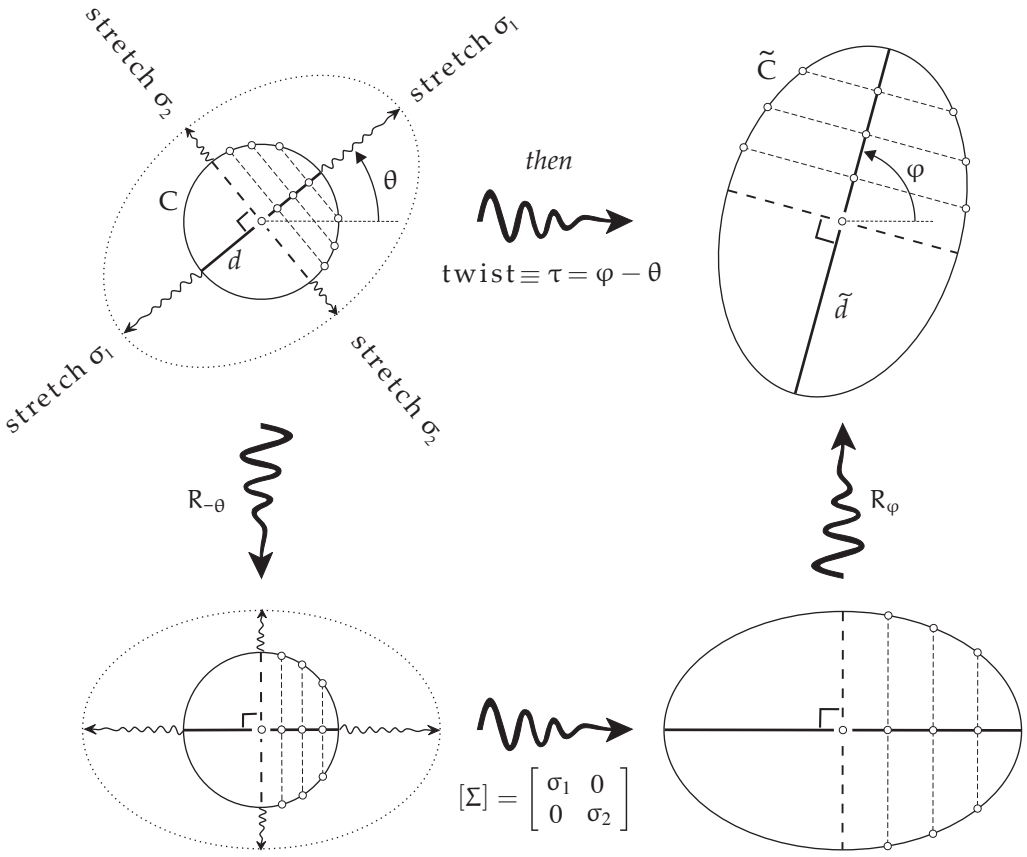
We now apply[3] these facts to $\widetilde{C}$.

Since all the diameters of C are bisected by the centre of C, it follows that the image chords of $\widetilde{C}$ must all pass through a common point of bisection. Thus the centre of the circle C is mapped to the centre of the ellipse $\widetilde{C}$.

---

[2] As we have noted and have sought to correct, Beltrami's name is not currently attached to his models of the hyperbolic plane. To add insult to injury, the average "mathematician in the street" is also unaware of the fact that it was the very same Eugenio Beltrami who first discovered the SVD! See Stewart (1993).

[3] The following argument previously appeared in VCA, page 208.



[15.4] Singular Value Decomposition (SVD): *The top half of the figure illustrates the effect, from left to right, of a general linear transformation* M *on an origin-centred circle. Preservation of midpoints enables us to prove that* M *is equivalent to two perpendicular stretches (by the "singular values"* σ_1 *and* σ_2*), followed by a rotation, which we call the* **twist** $\equiv \tau$*; here* $\tau = \varphi - \theta$*. The bottom half of the figure demonstrates that this SVD is equivalent to being able to write* $M = R_\varphi \circ \Sigma \circ R_{-\theta}$*, where* Σ *stretches horizontally by* σ_1 *and vertically by* σ_2*.*

Drawn in the same heavy line as its image is the particular diameter d of the circle that is mapped to the major axis $\tilde{d}$ of the ellipse. Now consider the chords (shown dashed) of C that are perpendicular to d. Since these are all bisected by d, their images must be a family of parallel chords of $\tilde{C}$ such that $\tilde{d}$ is their common bisector. They must therefore be the family *perpendicular* to $\tilde{d}$. (Convince yourself of this by drawing a family of parallel chords of $\tilde{C}$ in a general direction.)

It is now clear that the linear transformation is a stretch in the direction of d, another stretch perpendicular to it, and finally a twist, thereby confirming the existence of the SVD. Note that we could equally well do the orthogonal stretchings *after* the twist, instead of before. But in that case the stretchings would be in the new, twisted directions, along $\tilde{d}$, and orthogonal to it.

The bottom half of the figure demonstrates that the SVD is equivalent to being able to write

$$\boxed{M = R_\varphi \circ \Sigma \circ R_{-\theta},} \tag{15.11}$$

where Σ stretches horizontally by σ_1 and vertically by σ_2.

Note that this result also makes sense at the level of counting degrees of freedom. Just as the matrix has four independent entries, so the specification of our transformation also requires

four bits of geometric information: the direction of d, the stretch factor in this direction, the perpendicular stretch factor, and the twist.

A vitally important special case arises if we require that the two stretch factors be set to be equal: $\sigma_1 = \sigma_2 = a$, say, in which case the image of a circle is another *circle* that is a times as big. This apparently reduces the number of degrees of freedom from four to three. However, since we are now producing an equal expansion in all directions, the direction chosen for d becomes irrelevant, and we are thus left with only *two* genuine degrees of freedom: the expansion factor a, and the twist τ.

In VCA we sought to exploit geometrically the fact that that the local effect of a differentiable function $f(z)$ of a complex variable is precisely a mapping of the type just described. In this context, $a(z)$ and $\tau(z)$ both typically *vary* with position: they describe the expansion and rotation undergone by an infinitesimal neighbourhood of z.

More explicitly, recapping (4.18), page 42, every tiny complex arrow δz emanating from z is "amplitwisted" to an image arrow $\delta\widetilde{z}$ emanating from $\widetilde{z} = f(z)$, where $\delta\widetilde{z} \asymp [a e^{i\tau}]\,\delta z$. In VCA we called $a(z)$ the ***amplification*** of $f(z)$, and we called the combined effect of the local amplification and twist the ***amplitwist*** of $f(z)$, encoded as a complex number:

$$f'(z) = \text{amplitwist of } f(z) = (\text{amplification})\ e^{i(\text{twist})} = a e^{i\tau}.$$

Let us now resume our quest for the meaning of M^T. To facilitate this quest, first consider the *inverse* of the original linear transformation M (see the top of [15.4]). This inverse M^{-1} is drawn going from right to left in [15.5a]: first it undoes the twist τ of M by twisting $-\tau$, and then it undoes the stretchings of M by squashing by $1/\sigma_1$ and $1/\sigma_2$ in the same two orthogonal directions that M originally did stretching. Reversing the arrows in the bottom half of [15.4], we can also express this as

$$M^{-1} = R_\theta \circ \Sigma^{-1} \circ R_{-\varphi}.$$

At last we can reveal the geometric meaning of M^T. We shall assume that the reader is familiar with the algebraic properties of the transpose operation on matrices, and we shall use them to expose the underlying geometry. First note that taking the transpose of a rotation matrix (15.5) yields the opposite rotation: $[R_\theta]^T = [R_{-\theta}]$. Also, recall that $([P][Q])^T = [Q]^T[P]^T$. Thus (15.11) yields,

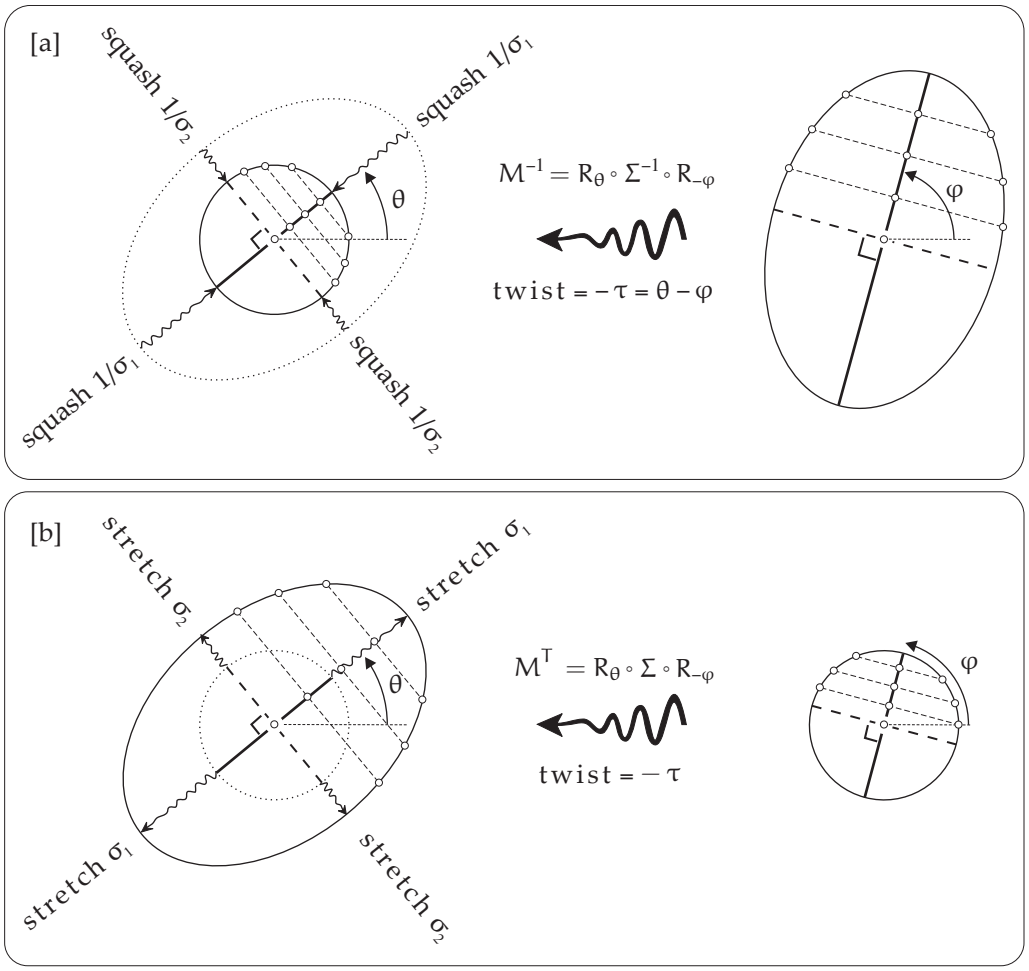
$$M^T = (R_\varphi \circ \Sigma \circ R_{-\theta})^T = R_{-\theta}^T \circ \Sigma^T \circ R_\varphi^T = R_\theta \circ \Sigma \circ R_{-\varphi},$$

and [15.5b] illustrates the meaning of this:

The linear transformation M^T *does the the* opposite *twist to* M*, followed by the* same *two orthogonal expansions as* M. (15.12)

Having grasped the geometric meaning of M^T, the geometric meaning of symmetry suddenly becomes clear, too:

The linear transformation M *is symmetric (i.e.* $M^T = M$*)* iff *the twist vanishes. But if* $\tau = 0$*, then the orthogonal expansion directions become eigenvectors, so we also conclude* M *is symmetric* iff *it has orthogonal eigenvectors.* (15.13)



[15.5] [a] *The inverse* M^{-1} *(going from right to left) of the original linear transformation at the top of [15.4]: it undoes* M *by first doing the opposite twist* $-\tau$*, and then squashing by* $1/\sigma_1$ *and* $1/\sigma_2$ *in the same two orthogonal directions that* M *originally did stretching.* **[b]** *The transpose* M^T *also starts by twisting* $-\tau$*, but then it does the* same *two orthogonal stretches that* M *did.*

If $\tau = \varphi - \theta = 0$, the two rotations in the bottom half of [15.4] are equal and opposite, and so (15.11) specializes:

$$\boxed{\textit{If } M^T = M \textit{ then } M = R_\theta \circ \Sigma \circ R_{-\theta},} \tag{15.14}$$

which is sometimes called the ***Spectral Theorem***.

We could, instead, have taken (15.12) as our *definition* of M^T, and then all the algebraic properties of the transpose would have become readily visualizable *geometric theorems*. For example, check that $R_\theta^T = R_{-\theta}$, which then implies $[R_\theta]^T = [R_{-\theta}]$. Likewise, check that $(P \circ Q)^T = Q^T \circ P^T$, which then implies $([P][Q])^T = [Q]^T[P]^T$.

Here is a fresh example of a result conventionally proved by calculation and now rendered transparent by geometry: M^T *expands area by the same factor as* M:

$$\boxed{|M^T| = \sigma_1\sigma_2 = |M|.}$$

And here is yet another result that is *proved* algebraically in every standard text, but which geometry now allows us to *understand*. If we first do M (top of [15.4]) and then do M^T ([15.5b]) then the opposite twists *cancel*, and we literally *see* that

> *The transformation* $(M^T \circ M)$ *is symmetric, having orthogonal eigenvectors (along* d *and perpendicular to it) with eigenvalues* σ_1^2 *and* σ_2^2. (15.15)

Lastly, we mention another commonly used manifestation of symmetry, leaving it to you to verify this both algebraically and geometrically:

> *A linear transformation* M *is symmetric (i.e.,* $M^T = M$*)* iff
>
> $$\mathbf{a} \bullet M(\mathbf{b}) = \mathbf{b} \bullet M(\mathbf{a}), \tag{15.16}$$
>
> *for all* **a** *and* **b**.

Armed with the geometric meaning of the transpose, Exercise 12 provides *eight more examples* of what we might call "Visual Linear Algebra." These examples underscore our overarching philosophy: direct, geometric reasoning frequently allows us to completely bypass symbolic manipulation to obtain an intuitive, *visual* grasp of mathematical reality.

15.5 The General Matrix of S

The Shape Operator S is a geometric entity, independent of any particular choice of basis vectors. But the matrix [S] that *represents* S certainly *does* depend on this choice.

Suppose that we choose an arbitrary orthonormal basis $\{\mathbf{E}_1, \mathbf{E}_2\}$, without first finding the principal directions and curvatures; how will the matrix look? While we may not yet know the principal directions, they certainly exist, so let us suppose that $\{\mathbf{e}_1, \mathbf{e}_2\}$ are obtained by rotating $\{\mathbf{E}_1, \mathbf{E}_2\}$ through some unknown angle θ. Figure [15.6] illustrates this in a case where $\kappa_1 > \kappa_2 > 0$. On the left, we see the effect of S on the unit circle, which is carried into an ellipse with semimajor axis κ_1 aligned with $\mathbf{e}_1$, and semiminor axis κ_2 aligned with $\mathbf{e}_2$.

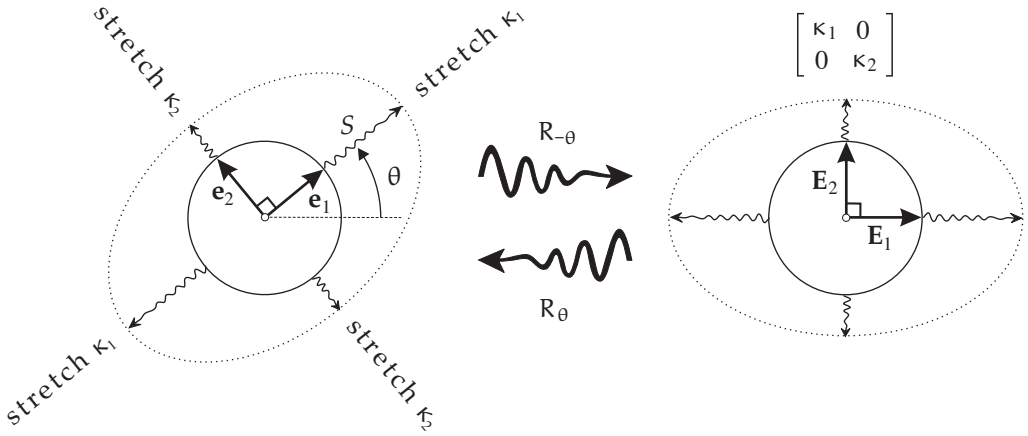
As we see on the right, in terms of $\{\mathbf{E}_1, \mathbf{E}_2\}$ the effect of S is equivalent to the following three successive transformations:

$$S = (\text{rotate by } -\theta) \text{ then } (\text{expand by } \kappa_{1,2} \text{ along } \mathbf{E}_{1,2}) \text{ then } (\text{rotate by } \theta).$$

If you read the previous section, this will already be familiar to you as the result (15.14).

Recall from Linear Algebra that the *product* [B][A] of two matrices is (or at least *should* be!) defined to be the matrix of the composite linear transformation: $[B][A] \equiv [B \circ A]$. It follows from (15.5) that

$$\begin{aligned} [S] &= [R_\theta] \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix} [R_{-\theta}] \\ &= \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \\ &= \begin{bmatrix} \kappa_1 c^2 + \kappa_2 s^2 & (\kappa_1 - \kappa_2)sc \\ (\kappa_1 - \kappa_2)sc & \kappa_1 s^2 + \kappa_2 c^2 \end{bmatrix}. \end{aligned} \tag{15.17}$$



[15.6] *On the left, the Shape Operator S stretches the orthogonal principal directions by factors equal to the principal curvatures, deforming the unit circle into an ellipse aligned with the principal directions. This effect of S is equivalent to first rotating by $-\theta$, producing the figure on the right, then stretching horizontally by κ_1 and vertically by κ_2, then rotating back by θ.*

At least one of these four entries should look familiar. Let $\kappa(\mathbf{E}_1)$ denote the curvature of the normal section of $\mathcal{S}$ taken in the $\mathbf{E}_1$ direction, which makes angle $-\theta$ with the first principal direction. Then Euler's formula (10.1) informs us that the top-left entry is $\kappa_1 c^2 + \kappa_2 s^2 = \kappa(\mathbf{E}_1)$.

But *why* has this simplification occurred?

15.6 Geometric Interpretation of *S* and Simplification of [*S*]

The answer sheds new light on the meaning of the Shape Operator:

> *If $\widehat{\mathbf{v}}$ is an arbitrary unit tangent vector, then the curvature $\kappa(\widehat{\mathbf{v}})$ of the normal section in this direction is given by $\kappa(\widehat{\mathbf{v}}) = \widehat{\mathbf{v}} \bullet S(\widehat{\mathbf{v}})$. Thus,*
>
> $\kappa(\widehat{\mathbf{v}}) =$ *Projection of* $S(\widehat{\mathbf{v}})$ *onto the direction of* $\widehat{\mathbf{v}}$.

(15.18)

(NOTE: By Meusnier's Theorem (11.4), $\kappa(\widehat{\mathbf{v}})$ is not merely the curvature of the normal section, it is also equal to the previously introduced normal curvature $\kappa_n(\widehat{\mathbf{v}})$ of *any* curve on the surface that passes through this point travelling in the direction $\widehat{\mathbf{v}}$.)

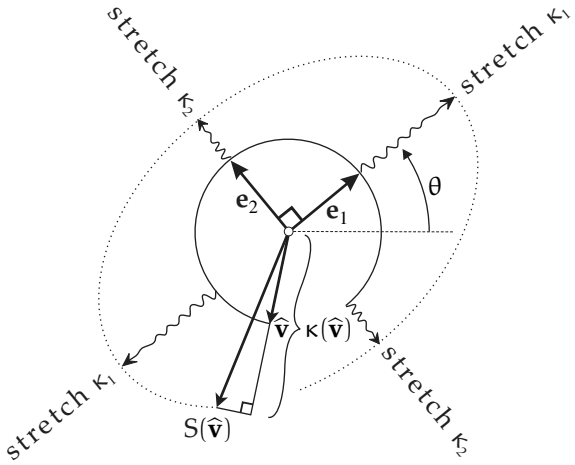
This interpretation is illustrated in [15.7]. Note in particular that this construction correctly yields $\kappa(\mathbf{e}_1) = \kappa_1$ and $\kappa(\mathbf{e}_2) = \kappa_2$.

The essential geometric explanation for this is the same as that shown in [8.5], on page 102. In fact we previously proved the result in the form (11.5), but we repeat the proof here to illustrate our new notation. Thinking of $\widehat{\mathbf{v}}$ as the velocity of a particle travelling along the normal section at unit speed, so that $\nabla_{\widehat{\mathbf{v}}}\widehat{\mathbf{v}} = \kappa(\widehat{\mathbf{v}})\,\mathbf{n}$ is its acceleration towards the centre of curvature, we find

$$0 = \nabla_{\widehat{\mathbf{v}}}(\widehat{\mathbf{v}} \bullet \mathbf{n}) = \widehat{\mathbf{v}} \bullet \nabla_{\widehat{\mathbf{v}}}\mathbf{n} + \mathbf{n} \bullet \nabla_{\widehat{\mathbf{v}}}\widehat{\mathbf{v}} = -\widehat{\mathbf{v}} \bullet S(\widehat{\mathbf{v}}) + \kappa(\widehat{\mathbf{v}}),$$

from which the result (15.18) follows immediately.

We can give an alternative derivation if we assume Euler's formula, (10.1). Currently we are using θ to denote the angle from $\mathbf{E}_1$ to $\mathbf{e}_1$, so, to avoid confusion with our earlier discussion of



[15.7] *The curvature of a normal section of a surface taken in the direction of a general unit tangent vector* $\hat{\mathbf{v}}$ *is given by* $\kappa(\hat{\mathbf{v}}) =$ *projection of* $S(\hat{\mathbf{v}})$ *onto the direction of* $\hat{\mathbf{v}}$.

Euler's formula, let us instead use α to denote the angle that a general $\hat{\mathbf{v}}$ makes with $\mathbf{e}_1$, so that Euler's formula now reads

$$\kappa(\hat{\mathbf{v}}) = \kappa_1 \cos^2\alpha + \kappa_2 \sin^2\alpha.$$

Since (15.18) is a purely geometric statement, as witnessed in [15.7], it is sufficient to demonstrate its truth any particular basis. And in the principal basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ a simple coordinate calculation confirms the result:

$$\hat{\mathbf{v}} \bullet S(\hat{\mathbf{v}}) = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} \bullet S \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} \bullet \begin{bmatrix} \kappa_1 \cos\alpha \\ \kappa_2 \sin\alpha \end{bmatrix} = \kappa(\hat{\mathbf{v}}),$$

by Euler's formula.

Armed with the result (15.18), we may now return to the interpretation and simplification of (15.17). Recall that even if S were an *arbitrary* linear transformation, the j-th column of its matrix would be the image of the j-th basis vector, so the first column of $[S]$ is the vector $S(\mathbf{E}_1)$, and therefore its projections along $\{\mathbf{E}_1, \mathbf{E}_2\}$ (i.e., its components) are the dot products of this vector with these basis vectors. Likewise, this is true for the second column of $[S]$, so

$$[S] = \begin{bmatrix} \mathbf{E}_1 \bullet S(\mathbf{E}_1) & \mathbf{E}_1 \bullet S(\mathbf{E}_2) \\ \mathbf{E}_2 \bullet S(\mathbf{E}_1) & \mathbf{E}_2 \bullet S(\mathbf{E}_2) \end{bmatrix}. \tag{15.19}$$

Thus (15.18) explains[4] why the top-left entry of (15.17) turned out to be $\kappa(\mathbf{E}_1)$. Now look at the other diagonal entry. Whether we use Euler's formula [exercise] or refer to (15.18) and (15.19), we see that the bottom-right entry of (15.17) is $\kappa_1 s^2 + \kappa_2 c^2 = \kappa(\mathbf{E}_2)$. If you read the previous (optional) section, also note that (15.16) shows that the symmetry of $[S]$ now follows from the symmetry of S, for $\mathbf{E}_1 \bullet S(\mathbf{E}_2) = \mathbf{E}_2 \bullet S(\mathbf{E}_1)$.

Before we do any more computation, let us pause to discuss more explicitly the geometric meaning of the entries in (15.19). As we have said, the first column is $S(\mathbf{E}_1) = -\nabla_{\mathbf{E}_1}\mathbf{n}$, which measures how the normal turns as we move off in the $\mathbf{E}_1$ direction. Let Π_1 be the normal plane in this direction, i.e., the plane spanned by $\mathbf{E}_1$ and $\mathbf{n}$. See [12.5b] on page 133 for a concrete example. Then

[4] Turning this around, we have obtained another *derivation* of Euler's formula.

the first component of $S(\mathbf{E}_1)$, namely $\mathbf{E}_1 \bullet S(\mathbf{E}_1)$, measures how fast $\mathbf{n}$ tips towards $\mathbf{E}_1$ *within* Π_1 as we begin to move along $\mathbf{E}_1$; this is determined by the curvature of the normal section, indeed it *is* the curvature $\kappa(\mathbf{E}_1)$. The second component of $S(\mathbf{E}_1)$, namely $\mathbf{E}_2 \bullet S(\mathbf{E}_1)$, measures how fast $\mathbf{n}$ rotates perpendicularly to the direction of motion, i.e., how fast it rotates *out of* Π_1.

Of course the meaning of the second column $S(\mathbf{E}_2)$ is completely analogous. The first component $\mathbf{E}_1 \bullet S(\mathbf{E}_2)$ measures how fast $\mathbf{n}$ rotates *out of* Π_2 as we begin to move along $\mathbf{E}_2$. The second component $\mathbf{E}_2 \bullet S(\mathbf{E}_2)$, measures how fast $\mathbf{n}$ rotates *within* Π_2, and this is $\kappa(\mathbf{E}_2)$.

The κ-related tipping within Π, and the rotation out of Π, are both clearly visible for the typical direction shown in [12.5b]. If, however, we move off in a *principal* direction then *all* of the tipping is in the direction of motion, within Π, and there is *no* initial rotation out of Π, and this is what we see in [12.5a].

Let us look more closely at the off-diagonal term, which measures how fast $\mathbf{n}$ swings out of $\Pi_{1,2}$. If we once again let $\Delta\kappa = (\kappa_1 - \kappa_2)$, as in (10.2), then the matrix representing the Shape Operator in a general basis $\{\mathbf{E}_1, \mathbf{E}_2\}$ is given by

$$[S] = \begin{bmatrix} \kappa(\mathbf{E}_1) & \frac{\Delta\kappa}{2}\sin 2\theta \\ \frac{\Delta\kappa}{2}\sin 2\theta & \kappa(\mathbf{E}_2) \end{bmatrix} \tag{15.20}$$

Note that this general matrix is indeed symmetric, as demanded by (15.9): $[S]^T = [S]$. Also note that if the basis coincides with the principal basis, so that $\theta = 0$, then (15.20) reduces to the diagonal form (15.7), as it should. The general matrix also reduces to this diagonal form if $\theta = (\pi/2)$, for in that case the basis is again aligned with the principal basis, only now $\{\mathbf{E}_1, \mathbf{E}_2\} = \{-\mathbf{e}_2, \mathbf{e}_1\}$. On the other hand, the off-diagonal term is *greatest* ($\mathbf{n}$ swings out of $\Pi_{1,2}$ the fastest) when $\theta = (\pi/4)$, in which case $\Pi_{1,2}$ bisect the principal directions.

15.7 [S] Is Completely Determined by Three Curvatures

Next, observe that the two equal off-diagonal entries are simply the oscillating term in Euler's formula, (10.2), phase-shifted by $\pm(\pi/4)$. Indeed, this formula tells us that the curvature of the normal section in the direction $\mathbf{E}_1 + \mathbf{E}_2$, bisecting the angle between the basis vectors, is given by

$$\kappa(\mathbf{E}_1 + \mathbf{E}_2) = \kappa(\tfrac{\pi}{4} - \theta) = \overline{\kappa} + \tfrac{\Delta\kappa}{2}\sin 2\theta. \tag{15.21}$$

Of course we could equally well work with the orthogonal direction:

$$\kappa(\mathbf{E}_1 - \mathbf{E}_2) = \kappa(\theta - \tfrac{\pi}{4}) = \overline{\kappa} - \tfrac{\Delta\kappa}{2}\sin 2\theta. \tag{15.22}$$

But, choosing (somewhat arbitrarily) to use the first direction, (15.20) may be written

$$[S] = \begin{bmatrix} \kappa(\mathbf{E}_1) & \kappa(\mathbf{E}_1 + \mathbf{E}_2) - \overline{\kappa} \\ \kappa(\mathbf{E}_1 + \mathbf{E}_2) - \overline{\kappa} & \kappa(\mathbf{E}_2) \end{bmatrix}. \tag{15.23}$$

It might seem at first that we still need to know $\kappa_{1,2}$ (or at least their sum) in order to calculate $\overline{\kappa}$. But in fact we do not, and we shall instead see that *the Shape Operator matrix can be expressed purely in terms of the curvatures of the normal sections in three directions:* $\mathbf{E}_1$, $\mathbf{E}_2$, and $(\mathbf{E}_1 + \mathbf{E}_2)$.

Recall that the ***trace*** of a matrix is the sum of its diagonal elements, so for the original diagonal matrix (15.7) the trace is $\mathrm{Tr}\,[S] = \kappa_1 + \kappa_2 = 2\overline{\kappa}$. But we know from Linear Algebra that if $[A]$ and $[B]$ are the matrices of *any* two linear transformation of the plane, then $\mathrm{Tr}\,[A][B] = \mathrm{Tr}\,[B][A]$. Thus,

$$\mathrm{Tr}\,[R_\theta][A][R_{-\theta}] = \mathrm{Tr}\,[A][R_{-\theta}][R_\theta] = \mathrm{Tr}\,[A],$$

in other words,

The trace of any linear transformation is invariant under rotation of the basis vectors. (15.24)

(There is actually a *geometric* reason for this; see Arnol'd (1973, §16.3).)

Therefore, returning to the case at hand, in which $[A] = [S]$, we deduce that even in the case of a general basis, with the matrix (15.23),

$$\kappa(\mathbf{E}_1) + \kappa(\mathbf{E}_2) = \mathrm{Tr}\,[S] = \kappa_1 + \kappa_2 = 2\,\overline{\kappa}.$$

In fact we need not even appeal to the general theorem of Linear Algebra to see this, for in our case we have already explicitly calculated $[S]$, and the truth of our assertion follows immediately from (15.17). Another specific example is obtained by adding (15.21) and (15.22).

Putting this into words, we have a result of interest in its own right:

> *The sum of the curvatures in* any *two perpendicular directions is equal to the sum of the principal curvatures.*

Thus, as claimed, the matrix (15.23) does indeed only depend on the curvatures in three directions, for $\overline{\kappa} = \frac{1}{2}[\kappa(\mathbf{E}_1) + \kappa(\mathbf{E}_2)]$, obviating the need to know the principal curvatures. Therefore, as claimed, the general matrix (15.23) of the Shape Operator can be written explicitly in terms of these three curvatures:

$$[S] = \begin{bmatrix} \kappa(\mathbf{E}_1) & \kappa(\mathbf{E}_1 + \mathbf{E}_2) - \frac{1}{2}[\kappa(\mathbf{E}_1) + \kappa(\mathbf{E}_2)] \\ \kappa(\mathbf{E}_1 + \mathbf{E}_2) - \frac{1}{2}[\kappa(\mathbf{E}_1) + \kappa(\mathbf{E}_2)] & \kappa(\mathbf{E}_2) \end{bmatrix}.$$

Finally, recall from (15.8) that the extrinsic version $\mathcal{K}_{\mathrm{ext}}$ of the Gaussian curvature is the determinant of *any* of these forms of the matrix of $[S]$. For example, (15.23) yields

$$\mathcal{K}_{\mathrm{ext}} = |[S]| = \kappa(\mathbf{E}_1)\kappa(\mathbf{E}_2) - [\kappa(\mathbf{E}_1 + \mathbf{E}_2) - \overline{\kappa}]^2\,.$$

15.8 Asymptotic Directions

Recall that *Dupin's indicatrix* $\mathcal{D}$ (p. 111) arises from the intersection of the surface with a plane $T_p(\epsilon)$ parallel to the tangent plane T_p and distance ϵ away from it, so $T_p(0) = T_p$. As ϵ increases from 0 and $T_p(\epsilon)$ just begins to move away in the normal direction, $\mathcal{D}$ is the nascent conic section of intersection with the surface.

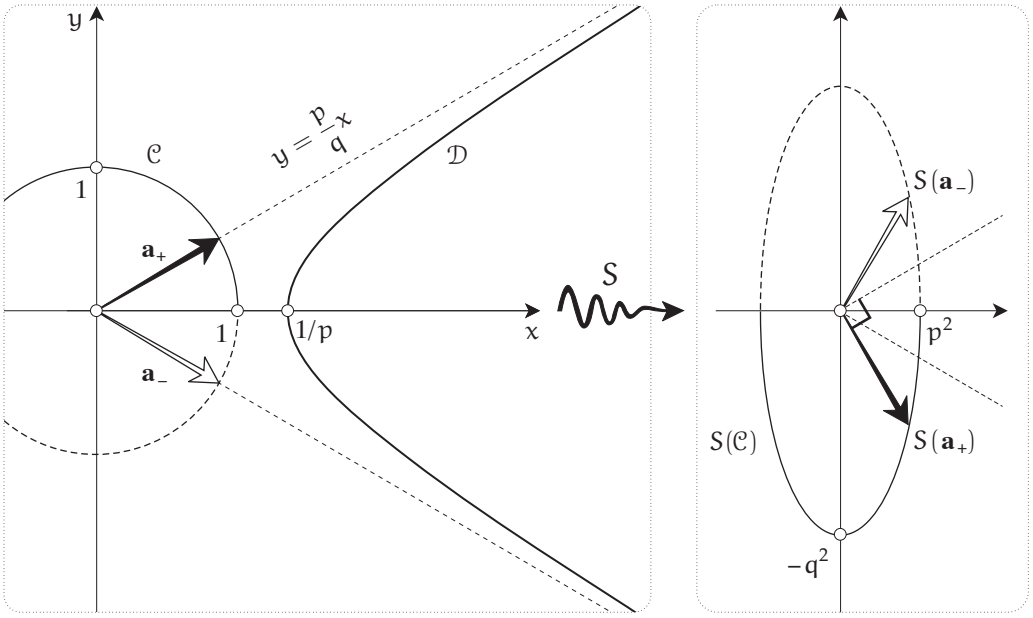
If $\mathcal{K}(p) > 0$ then $\mathcal{D}$ is an ellipse, and p is called elliptic; if $\mathcal{K}(p) = 0$ then $\mathcal{D}$ is a parabola, and p is called parabolic; if $\mathcal{K}(p) < 0$ then $\mathcal{D}$ is a hyperbola, and p is called hyperbolic.

Here we shall focus on the hyperbolic (negative curvature) case, and shall investigate the directions of the asymptotes of the hyperbola $\mathcal{D}$, which are called the ***asymptotic directions***. As we shall now explain, the asymptotic directions have a simple geometric relationship to the Shape Operator S.

As discussed in the previous chapter, if the (x, y)-axes are aligned with the principal directions, the equation of the pair of conjugate hyperbolas $\mathcal{D}$ is

$$\kappa_1 x^2 + \kappa_2 y^2 = \pm 1,$$

the sign being determined by the sign of ϵ, i.e., whether we move the tangent plane up or down. The symmetry axes (in both the elliptic and hyperbolic cases) coincide with the principal/



[15.8] *At a hyperbolic point, the Shape Operator maps each of the asymptotic direction vectors* $\mathbf{a}_\pm$ *of the Dupin indicatrix* $\mathcal{D}$ *into an* orthogonal *direction, and in opposite senses:* S *turns* $\mathbf{a}_\pm$ *through angle* $\mp(\pi/2)$.

coordinate axes, and in the hyperbolic case these bisect the angles between the asymptotic directions.

Let $\kappa_1 = p^2$ and $\kappa_2 = -q^2$ (with p and q positive) and let us choose the plus in the above equation for $\mathcal{D}$. Then,

$$\mathcal{D}: p^2x^2 - q^2y^2 = 1 \qquad \text{and} \qquad [S] = \begin{bmatrix} p^2 & 0 \\ 0 & -q^2 \end{bmatrix}.$$

Thus the asymptotes have equations $y = \pm\frac{p}{q}x$. See [15.8], in which we have added the unit circle $\mathcal{C}$, and its elliptical image $S(\mathcal{C})$, to help visualize the effect of the mapping S, which stretches horizontally and vertically, *and* flips across the x-axis. Thus the vectors in the asymptotic directions, and their images under S, are given by,

$$\mathbf{a}_\pm \propto \begin{bmatrix} q \\ \pm p \end{bmatrix} \qquad \Longrightarrow \qquad S(\mathbf{a}_\pm) \propto \begin{bmatrix} p \\ \mp q \end{bmatrix}.$$

As illustrated in [15.8], this means that athough S maps a principal direction to the *same* direction,

The Shape Operator maps an asymptotic direction to an orthogonal *direction:*

$$\mathbf{a}_\pm \bullet S(\mathbf{a}_\pm) = 0. \tag{15.25}$$

In fact, as we see in [15.8], S turns the two asymptotic directions in opposite senses, $\mathbf{a}_+$ by $-\frac{\pi}{2}$, and $\mathbf{a}_-$ by $+\frac{\pi}{2}$.

But *why* has this happened? We have pictured $\mathcal{D}$ as the magnified intersection curve of the surface with $T_p(\epsilon)$ just as ϵ increases from 0. But, initially, when $\epsilon = 0$ this hyperbola of intersection degenerates into the asymptotes themselves. Since this intersection curve (with directions $\mathbf{a}_\pm$) lies in the tangent plane, it has vanishing normal curvature, so it follows from the geometric interpretation (15.18) of the Shape Operator that

$$\kappa(\widehat{\mathbf{a}}_\pm) = \widehat{\mathbf{a}}_\pm \bullet S(\widehat{\mathbf{a}}_\pm) = 0,$$

thereby explaining the result.

In light of this connection with the normal curvature, the definition of ***asymptotic direction*** has been generalized to mean *any* direction for which the normal curvature vanishes. Thus a parabolic point *also* has one "asymptotic direction" [exercise: what is it?] despite the fact that the parabola $\mathcal{D}$ does not have any asymptotes.

If we choose $\mathbf{E}_1$ to be either one of the asymptotic directions, so that Π_1 is the normal plane in this direction, then as we move off within the surface along $\mathbf{E}_1$, the normal $\mathbf{n}$ must rotate about $\mathbf{E}_1$, swinging straight out of Π_1, its head moving in a direction tangent to the surface and perpendicular to $\mathbf{E}_1$, i.e., in the direction $\pm\mathbf{E}_2$. Thus $S(\mathbf{E}_1) = \pm\tau\,\mathbf{E}_2$, where τ is the *rate* of spinning of $\mathbf{n}$ about $\mathbf{E}_1$, the so-called torsion[5] introduced on page 106. It follows from (15.19) that

> *At a hyperbolic point, the curvature of the surface can be expressed in terms of the torsion τ of an asymptotic curve as*
>
> $$\mathcal{K}_{ext} = |[S]| = -\tau^2. \tag{15.26}$$

According to Stoker (1969, p. 101), this result is due to Beltrami and to Enneper.

15.9 Classical Terminology and Notation: The Three *Fundamental Forms*

When consulting older, classical works on Differential Geometry, you will encounter terms and notation that we do not employ in this book. In particular, you will certainly meet the three so-called ***Fundamental Forms***, denoted I, II, and III.

All three Fundamental Forms are symmetric functions of pairs of tangent vectors, $\mathbf{u}$ and $\mathbf{v}$.

First Fundamental Form:	$\mathrm{I}(\mathbf{u},\mathbf{v})$	$\equiv$	$\mathbf{u}\bullet\mathbf{v}$.
Second Fundamental Form:	$\mathrm{II}(\mathbf{u},\mathbf{v})$	$\equiv$	$S(\mathbf{u})\bullet\mathbf{v}$.
Third Fundamental Form:	$\mathrm{III}(\mathbf{u},\mathbf{v})$	$\equiv$	$S(\mathbf{u})\bullet S(\mathbf{v})$.

The Shape Operator itself does not appear in the classical literature[6]; its mathematical content is instead represented by the Second Fundamental Form. For example, the curvature of the normal section in the direction $\widehat{\mathbf{v}}$, given by (15.18), would, in yesteryears, have been written, $\kappa(\widehat{\mathbf{v}}) = \mathrm{II}(\widehat{\mathbf{v}},\widehat{\mathbf{v}})$.

WARNING: Although we shall not introduce and define ***Differential Forms*** (***Forms*** for short) until Act V, we should immediately caution the reader that the three classical "Forms" are **not** *Forms* at all! While we certainly do not fault modern authors who continue to speak the classical language, *our* reliance on *genuine* Forms impelled us to turn the classical language into a dead language!

---

[5] We hope no confusion arises here from the letter τ serving double duty: we also recently used it to denote the (unrelated) "twist" of an SVD decomposition.

[6] The Shape Operator was first seriously championed, and brought into common usage, by Barrett O'Neill in his groundbreaking introductory text (O'Neill 2006) the first edition of which appeared in 1966.

Chapter 16

Introduction to the Global Gauss–Bonnet Theorem

16.1 Some Topology and the Statement of the Result

The ***Global Gauss–Bonnet Theorem*** is widely considered to be one of the most beautiful results in all of mathematics. Furthermore, it is also *fundamental*, having spawned the discovery of ever-more powerful generalizations, the culmination of which is perhaps (for now) the ***Atiyah–Singer Index Theorem***, discovered in 1963. *This* Theorem, in turn, caused seismic shifts in other areas of mathematics and in theoretical physics. These subsequent developments are far beyond the scope of the present work, and sadly beyond the competence of the present author, but the statement of the original form of GGB (as we shall henceforth abbreviate it) is surprisingly simple and easily understood. We need just a few preliminaries before we can state the result.

First, the ***total curvature*** $\mathcal{K}(\mathsf{P})$ of a region P of a curved surface is defined (naturally enough) to be,

$$\mathcal{K}(\mathsf{P}) \equiv \iint_{\mathsf{P}} \mathcal{K}\, \mathrm{d}\mathcal{A}.$$

(NOTE: Even in modern times, this concept is occasionally still referred to by its old Latin name, ***Curvatura Integra***.) For example, if P is a flat piece of paper with an arbitrary simple curve as boundary, then $\mathcal{K}(\mathsf{P}) = 0$. If we now bend the paper into a different shape $\widetilde{\mathsf{P}}$ (without stretching it), say, into a portion of a cylinder or cone, then $\mathcal{K}(\widetilde{\mathsf{P}}) = 0$, by virtue of the *Theorema Egregium*.

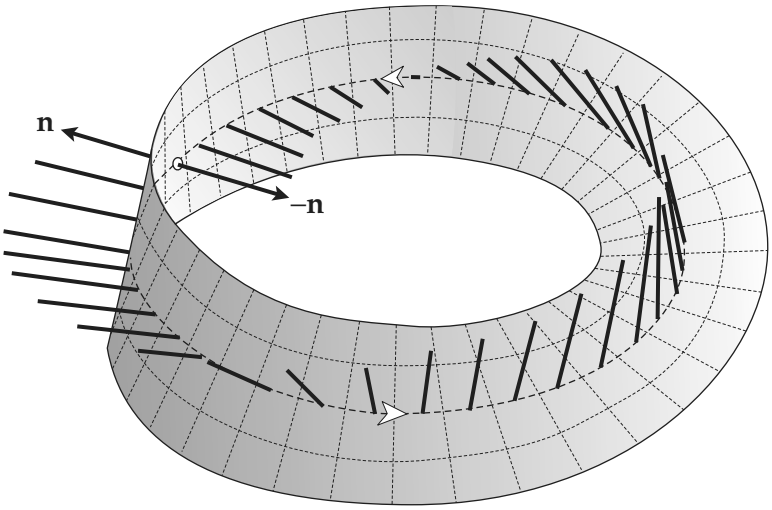
But now suppose instead that P is made of a highly *stretchable* material, like rubber. If we stretch P over the surface of a sphere, then $\mathcal{K}(\widetilde{\mathsf{P}}) > 0$. Indeed, by choosing the sphere to be as small as we please, we can make its curvature as large as we please, and with it $\mathcal{K}(\widetilde{\mathsf{P}})$ becomes as large as we please. If we instead stretch R into the form of a portion of a pseudosphere, then $\mathcal{K}(\widetilde{\mathsf{P}}) < 0$, and again this can be made as negative as we please.

Such a continuous, one-to-one stretching, $\mathsf{P} \mapsto \widetilde{\mathsf{P}}$, which does not preserve lengths or even angles, is called a ***topological mapping*** or ***topological transformation***, or ***homeomorphism***. Just as ancient Greek geometry sought out properties of figures that are *invariant* under rigid, distance-preserving mappings, so in the nineteenth century a new area of mathematics arose, called ***topology***, wherein properties were sought that were invariant under topological transformations.

Clearly, the concept of curvature does *not* belong to topology: stretching the surface in the vicinity of a point p changes the value of $\mathcal{K}(\mathsf{p})$. Indeed, we have just seen that $\mathcal{K}(\widetilde{\mathsf{p}})$ can be made to take on any value we please, either positive or negative. And likewise the more primitive concepts of length and angle do not belong to topology. Thus, at first sight, it might appear that topology would be a rather trivial or barren area of study: how can *anything* interesting or subtle survive the arbitrarily complicated and extreme distortions of a topological mapping?!

Remarkably, nothing could be further from the truth. Out of the kindling embrace of its principal parents (Riemann and Poincaré), topology rapidly grew up to become a powerful yet beneficent Hydra, explaining and unifying phenomena in disparate and distant realms of thought.

In order to introduce our first topological invariant, we restrict our attention to *closed* surfaces that are also *orientable*. Any such surface may be pictured as the boundary of a solid object in $\mathbb{R}^3$. Such a surface is automatically ***orientable***, meaning that one may consistently decide which of the



[16.1] *A* **Möbius band** *is* nonorientable*: carrying the normal* **n** *along a full circuit of the centre line, it returns to its starting point as* $-\mathbf{n}$.

two opposite choices of **n** is *the* normal to the surface—let it point *out* of the solid object into empty space.

But isn't *every* surface orientable? No! As Möbius and Listing independently discovered in 1858, taking a paper strip and giving it a half-twist, and then gluing the ends of the strip together to form a loop, produces a surface with only *one side*! As illustrated in [16.1], starting with an arbitrary initial choice of **n**, then continuously carrying **n** along a full circuit of the centre line, we return it to its starting point as $-\mathbf{n}$. Thus this so-called ***Möbius band***[1] is *non*orientable.

Returning to ordinary, orientable, closed surfaces, the fundamental feature that topologically distinguishes one from another is the number of *holes* it contains. This is called the ***genus*** g of the surface, and its value is illustrated in [16.2] for a few surfaces. Each pair of surfaces with the same genus is ***topologically equivalent***, or ***homeomorphic***, meaning that one may be changed into the other by a topological mapping (aka ***homeomorphism***). But two surfaces of different genus *cannot* be topologically equivalent, since g is invariant under topological mappings.

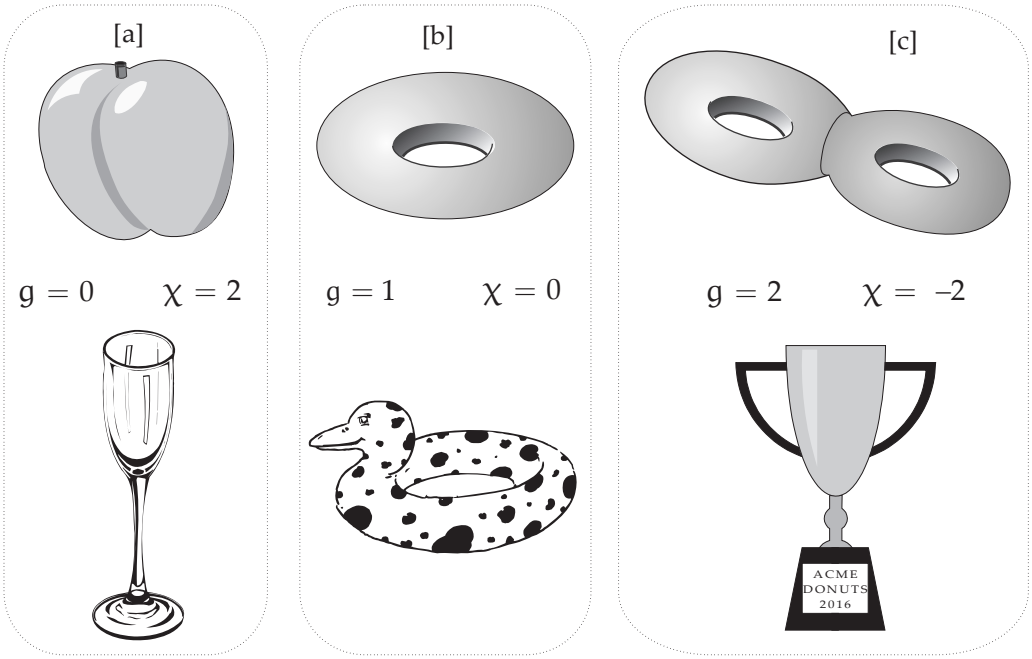
The genus was defined more precisely by Riemann, in 1851,[2] as the *maximum* number of cuts along closed, nonintersecting loops on the surface that can be performed without splitting the surface into two disconnected pieces. For example, cutting the sphere along any closed loop will split it into two, so the genus is zero. On the torus, we can cut along just *one* loop without splitting the torus into two pieces: e.g., cut along either an equator that encircles the axis of symmetry, or along a circle that goes through the hole (whose plane contains the axis of symmetry). But if we now cut the resulting surface along any loop that avoids the first, it will split it into two, so the genus of the torus is one. Try (in your mind) making loop cuts on a two-holed doughnut, as seen in [16.2c], and verify that $g = 2$.

As Möbius realized in 1863, every closed, orientable surface is topologically equivalent to a g-holed torus. We shall accept such visually plausible statements without proof, but the reader in search of precise definitions and proofs should consult the excellent topology texts we recommend in *Further Reading*, at the end of this book.

---

[1] Also commonly called a ***Möbius strip***.

[2] See Stillwell (1995), page 58.



[16.2] *Each pair of surfaces is topologically the same, distinguished from the other pairs by its number of holes, the* genus, g. *Also shown are the corresponding values of the* Euler characteristic, $\chi = 2 - 2g$.

We can now state the stunning result:

> **Global Gauss–Bonnet Theorem (GGB).** *The total curvature of a closed, orientable surface* $\mathcal{S}_g$ *depends* only *on its topological genus* g, *and is given by*

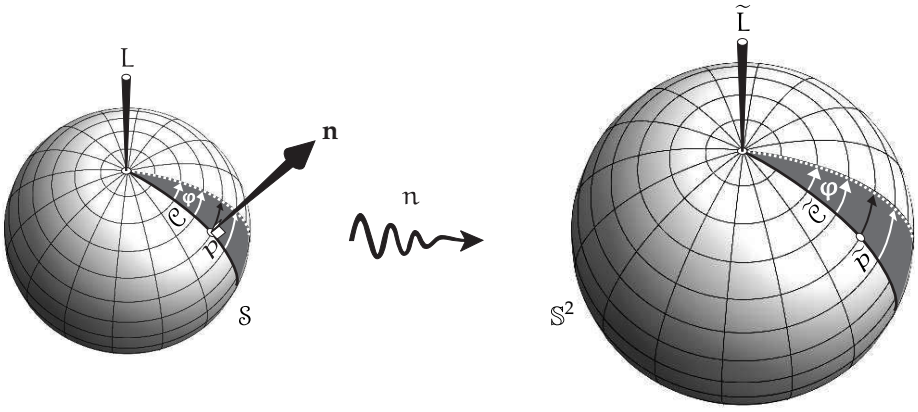
$$\boxed{\mathcal{K}(\mathcal{S}_g) = 4\pi(1-g) = 2\pi\chi(\mathcal{S}_g).} \tag{16.1}$$

Here the quantity

$$\boxed{\chi(\mathcal{S}_g) \equiv 2 - 2g} \tag{16.2}$$

serves (for now) as merely an alternative means of labelling the surfaces of different genus: see [16.2] for examples. This quantity $\chi(\mathcal{S}_g)$ is called the ***Euler characteristic*** of the surface; it arises naturally in many topological results, and actually has a meaning all its own, which will be explained in Section 18.1.

Pause for a moment to let the surprise and the beauty of the result sink in. If we take any given $\mathcal{S}_g$ (such as a simple doughnut with $g = 1$) made out of rubber or Play-Doh, then stretch it, twist it, squeeze it, and deform it in any way we please, *every resulting increase in curvature at one point of the surface must be instantaneously cancelled out by an exactly opposite decrease in curvature somewhere else on the surface.* In the case of a topological doughnut (aka, ***torus***), this total curvature remains exactly zero throughout our manipulations.



[16.3] *As the semicircle $\mathcal{C}$ rotates about L through angle φ, its spherical image $\widetilde{\mathcal{C}}$ rotates on $\mathbb{S}^2$ at the same rate, generating a lune of equal angle φ. As φ increases to 2π, $\widetilde{\mathcal{C}}$ sweeps out the entire surface of $\mathbb{S}^2$, so $\mathcal{K}(\mathcal{S}) = 4\pi$.*

16.2 Total Curvature of the Sphere and of the Torus

16.2.1 Total Curvature of the Sphere

If a surface $\mathcal{S}$ is topologically equivalent to a sphere (i.e., with $g = 0$), the prediction of GGB is that should have total curvature 4π. In the case of a *geometric* sphere $\mathcal{S}$, this is easily verified in multiple ways. We provide three proofs, the last of which will be shown to generalize beyond spheres.

First, each normal section of the sphere of radius R is itself a circle of radius R, i.e., of curvature $(1/R)$. So $\mathcal{K} = \kappa_1\kappa_2 = (1/R^2)$, and the total curvature

$$\mathcal{K}(\mathcal{S}) = \iint_{\mathcal{S}} \mathcal{K}\, d\mathcal{A} = \iint_{\mathcal{S}} \frac{1}{R^2}\, d\mathcal{A} = \frac{1}{R^2} 4\pi R^2 = 4\pi.$$

A second, more enlightening explanation is based on [12.2], page 132: if P is a region on $\mathcal{S}$ and $\widetilde{P} = n(P)$ is its spherical image of P, then

$$\mathcal{K}(P) = [\text{area of spherical image of } P \text{ on } \mathbb{S}^2] = \mathcal{A}(\widetilde{P}). \tag{16.3}$$

Now, take the image unit sphere $\mathbb{S}^2$ to have the same centre as $\mathcal{S}$, then the spherical map becomes a radial projection, as illustrated in [12.3], page 132. This makes it crystal clear that $\widetilde{\mathcal{S}} = n(\mathcal{S}) = \mathbb{S}^2$, so (16.3) implies $\mathcal{K}(\mathcal{S}) = \mathcal{A}(\mathbb{S}^2) = 4\pi$.

Third, and finally, we picture the sphere $\mathcal{S}$ as the surface of revolution obtained by rotating a semicircle $\mathcal{C}$ of radius R about the diameter L through its ends, as depicted in [16.3].

Although we may generate the entire sphere by rotating $\mathcal{C}$ a full 2π, the figure depicts the process partway through, after $\mathcal{C}$ has rotated through angle φ, generating a so-called ***lune***.

The key observation is this, and it applies to an *arbitrary* surface of revolution, generated by rotating an *arbitrary* plane curve $\mathcal{C}$ about an *arbitrary* line within its plane:

> *If a plane curve $\mathcal{C}$ **and** its unit normal vector $\mathbf{n}$ are **together** rotated about an arbitrary line in the plane of $\mathcal{C}$, then the rotated $\mathbf{n}$ is the normal to the surface of revolution generated by $\mathcal{C}$.*

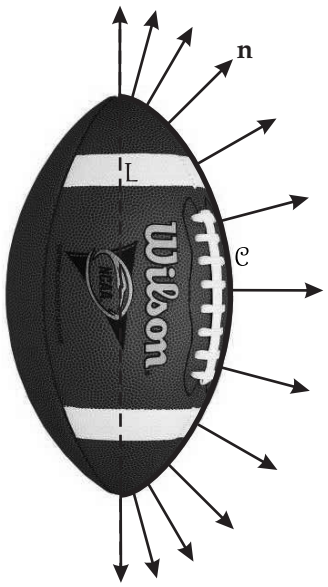
The explanation is readily understood from the special case in the figure. The tangent plane to the surface at a typical point p is spanned by the following two directions: (1) the direction of the rotated $\mathcal{C}$; and (2) the direction in which p moves as it rotates (i.e., perpendicularly to the plane of $\mathcal{C}$). But **n** was initially perpendicular to these two directions, so it *remains* perpendicular to (1) and (2) as $\mathcal{C}$ and **n** rotate together. Since **n** is perpendicular to two directions that span the tangent plane of $\mathcal{S}$, it *is* the normal to $\mathcal{S}$.

By virtue of (16.3), this immediately implies the following:

> *Let $\mathcal{C}$ be a plane curve, and L be a line in that plane, and let $\widetilde{L}$ be the line through the centre of S^2 that is parallel to L. Then, as $\mathcal{C}$ rotates about L, its spherical image $\widetilde{\mathcal{C}} = n(\mathcal{C})$ rotates at the same rate about $\widetilde{L}$, and the total curvature of the surface swept out by $\mathcal{C}$ is equal to the total (signed) area swept out by $\widetilde{\mathcal{C}}$ on S^2.* (16.4)

In particular, as illustrated, the spherical image of this lune is simply a lune of equal angle on S^2, and if $\mathcal{C}$ sweeps out all of $\mathcal{S}$, then its spherical image sweeps out all of S^2, once again confirming that $\mathcal{K}(\mathcal{S}) = 4\pi$.

The advantage of this last point of view is that we now gain our first real insight into GGB itself. Consider the American football depicted in [16.4], which is generated by the rotation of the curve $\mathcal{C}$ about the line L. Here we have drawn the normals at equal angular increments, which makes clear the variation in curvature: greatest around the poles, and least around the equator. Nevertheless, since the football is topologically spherical, the *total* curvature should be 4π, and we can now *see* that is. For the spherical image of $\mathcal{C}$ is the *same* semicircle $\widetilde{\mathcal{C}}$ as before (connecting the poles of S^2), and therefore as $\mathcal{C}$ rotates to generate the football, $\widetilde{\mathcal{C}}$ rotates to sweep out all of S^2, just as before, so the total curvature of the football is indeed 4π!



[16.4] *The spherical image of an American football is the whole of S^2, so $\mathcal{K}(\text{football}) = 4\pi$, in accordance with GGB.*

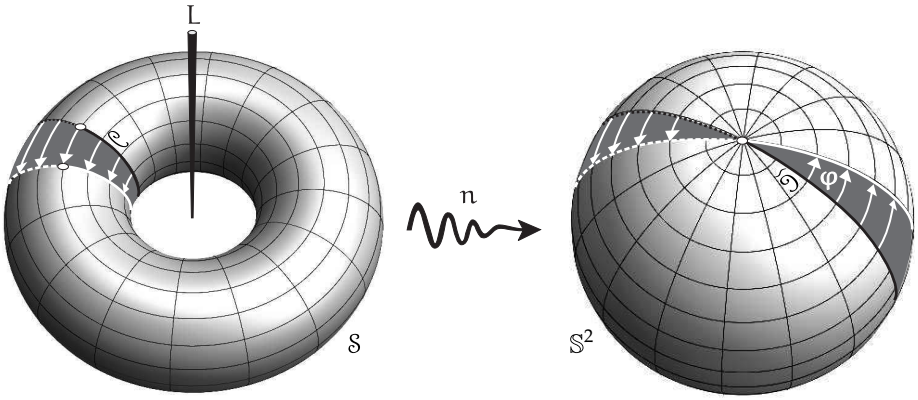
Clearly, the essential point is that spherical image of the football covers all of S^2. In the next chapter we will elaborate on this idea to produce our first (heuristic) proof of GGB.

16.2.2 Total Curvature of the Torus

As we observed in (10.13), page 114, if a section of $\mathcal{C}$ is concave *away* from L, then, as it rotates about L, it generates a portion of the surface that has *negative* curvature. In this case, the area generated by $\widetilde{\mathcal{C}}$ on S^2 should be *subtracted* from the total curvature. We now illustrate this with the torus, which of course has regions of both positive and negative curvature.

According to GGB, the doughnut (torus) should have vanishing total curvature, and in Exercise 23, page 89, you actually verified this by brute force, integrating the curvature formula over the entire surface. We are now in a position to provide a real *explanation*, in fact establishing a (superficially) stronger, local (as opposed to global) result.

Suppose we wish to eat only part of the doughnut: we could cut out a wedge—the darkly shaded region in [16.5]—as one would a wedge of cake, by making two slices through the axis of symmetry. We will now apply (16.4) to [16.5] to show that this wedge of doughnut must have vanishing total curvature. The global result then follows by greed: we cut a larger and larger wedge of doughnut until we eat the whole thing!



[16.5] *As the circle $\mathcal{C}$ rotates about* L *through angle φ, its (great circle) spherical image $\widetilde{\mathcal{C}}$ rotates on $\mathbb{S}^2$ at the same rate, generating two lunes of angle φ. As we see, the spherical map preserves orientation on the outer half of the doughnut, and reverses it on the inner half. The total curvature of the wedge of doughnut is the sum of the equal and opposite (signed) areas of the two lunes, and therefore it vanishes. Thus the total curvature of the doughnut vanishes, also.*

As the circle $\mathcal{C}$ rotates about L through angle φ, its (great circle) spherical image $\widetilde{\mathcal{C}}$ rotates on $\mathbb{S}^2$ at the same rate, generating two lunes of angle φ. As we see, the spherical map preserves orientation on the outer half of the doughnut, and reverses it on the inner half. To highlight this reversal of orientation on the negatively curved inner half of doughnut, we have placed the label $\mathcal{C}$ there, so it is mapped to a *backwards* $\widetilde{\mathcal{C}}$.

The total curvature of the wedge of doughnut is the sum of the equal and opposite (signed) areas of the two lunes, and therefore it *vanishes*. Letting φ increase to 2π, we see that the image of the outer half of the doughnut completely covers $\mathbb{S}^2$ once *positively*, while the image of the inner half completely covers $\mathbb{S}^2$ once *negatively*, so that the total is $\mathcal{K}(\textit{whole doughnut}) = 4\pi + (-4\pi) = 0$.

Observe something else. The entire circle at the top of the doughnut (which divides the outer and inner halves of opposite curvature) is mapped to a single point, namely, the north pole of $\mathbb{S}^2$. Likewise the lowest circle, upon which the doughnut would rest if it were placed on a plate, is likewise mapped to the south pole. These points of $\mathbb{S}^2$, where the two separate layers that cover $\mathbb{S}^2$ join together, are called ***branch points***.

Our next task is to find ways of visualizing the total curvature of surfaces that have *more* than one hole.

16.3 Seeing $\mathcal{K}(\mathcal{S}_g)$ via a Thick Pancake

Imagine pouring a very dense pancake batter into a frying pan to make a large, thick pancake. Before it can start to cook through and set, we quickly take a cylindrical biscuit cutter and remove g cylindrical discs from the interior of the pancake, leaving g holes. The batter starts to ooze back into the holes—thereby producing shapes like the inner halves of doughnuts—then starts to cook and harden, producing a genus g, thick pancake of the form [16.6].

The spherical map sends the entire flat bottom of the pancake to the south pole of $\mathbb{S}^2$, and it likewise sends the flat top of the pancake to the north pole. As expected, these yield no area on the sphere, so no contribution to the total curvature.

On the other hand, the outer edge of the pancake is like the outer half of a torus, and so its spherical image covers the entire sphere once, positively, contributing 4π to the total curvature. But each of the g rims to the holes in the pancake is like the inner half of a torus, so each one

[16.6] *Here $\mathcal{S}_g$ takes the form of a thick pancake with g holes. The top and bottom are flat, and so contribute no curvature, but the outer rim has $\mathcal{K} = +4\pi$, and the rim of each of the g holes has $\mathcal{K} = -4\pi$. The total curvature is therefore $4\pi(1-g)$.*

has a spherical image that covers the entire sphere once, negatively, contributing -4π to the total curvature. Thus (16.3) allows us to *see*[3] that

$$\mathcal{K}(\mathcal{S}_g) = 4\pi + (-4\pi)g = 4\pi(1-g) = 2\pi\chi(\mathcal{S}_g).$$

16.4 Seeing $\mathcal{K}(\mathcal{S}_g)$ via Bagels and Bridges

We now descend momentarily into the first person to relate a personal but pertinent anecdote.

One day, while waiting to be served at a bagel shop near my University of San Francisco campus, I observed that only a half-dozen plain bagels remained in the display case, still joined together in a neat row in the baking tin—exactly the number I hoped to purchase! While I continued to wait, I passed the time with two thoughts: (1) A fervent (albeit atheistic) "prayer" that the three people ahead of me in the queue didn't like plain bagels and (2) the happy certainty that, even in the absence of any detailed knowledge of the geometry of the 6-holed object of my desire, GGB absolutely *guaranteed* that its total curvature was exactly -20π.

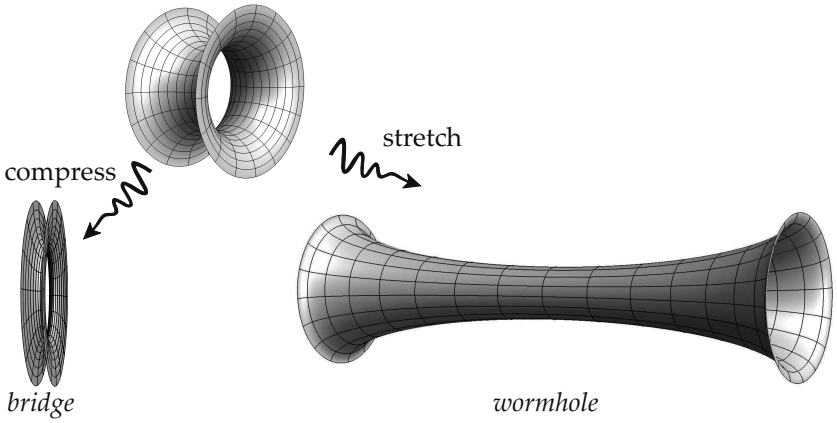
My prayer (1) was granted, and the server gently tore each of the six bagels from the row and successively dropped them into a brown paper bag, which he then handed to me. As I left the shop it struck me that while each of the bagels in my bag *appeared* totally untampered with, their original collective total curvature of -20π had evaporated to ZERO! Seemingly without leaving a trace of his crime, the nefarious server had robbed me of *all* the negative curvature promised to me by GGB!

Clearly, the only change in the geometry had been right *at* the very small joins (or ***bridges***, as we shall call them) between the bagels, where they were torn apart, so these must have been the culprits for the dramatic -20π curvature heist. Since the six bagels were joined by *five* bridges, each one presumably added 4π when it was torn asunder. In that case, each bridge must have originally stored -4π of curvature.

We shall now confirm this theory, thereby explaining the mystery, allowing us to *see* $\mathcal{K}(\mathcal{S}_g)$ in a new way (and, incidentally, absolving the server of malfeasance!).

---

[3] Although I have never seen this idea written down, I know (via private conversation) that my friend Professor Tom Banchoff hit upon the same idea, long before I did, right down to thinking in terms of pancakes!



[16.7] *The surface at the top is the inner half of a torus, and has $\mathcal{K} = -4\pi$. This total curvature is not altered if the surface is compressed to create a bagel* ***bridge*** *(left), or stretched to create a* ***wormhole*** *(right).*

Consider [16.7]. At the top is a surface that looks like the rim of one of the holes in our thick pancake, though now turned on its side; alternatively, think of the inner half of a torus. As we we have discussed, its spherical image covers S^2 once negatively, so its total curvature is $\mathcal{K} = -4\pi$.

Now suppose that we dramatically compress this surface horizontally to yield the surface on the left, resembling a bridge between two joined bagels in the baking tin. If we follow the evolution of the spherical image as the surface undergoes this compression, we see that as a the region around the narrow throat of the bridge contracts, its spherical image actually expands. Nevertheless, while some parts of the spherical image expand and others contract, the compressed bridge surface, taken as a whole, has the *same* spherical image as the uncompressed original, and its total curvature is therefore still $\mathcal{K} = -4\pi$, as anticipated.

For future use, we also note that, likewise, the total curvature does not change if we instead *stretch* the surface to produce the ***wormhole*** (as we shall call it) on the right of [16.7]; it too has $\mathcal{K} = -4\pi$.

As [16.8] makes clear, the consequence of each bridge having $\mathcal{K} = -4\pi$ is that

$$\mathcal{K}(\mathcal{S}_g) = g \cdot \mathcal{K}(\text{bagel}) + (g-1) \cdot \mathcal{K}(\text{bridge}) = g \cdot 0 + (g-1) \cdot (-4\pi) = 4\pi(1-g),$$

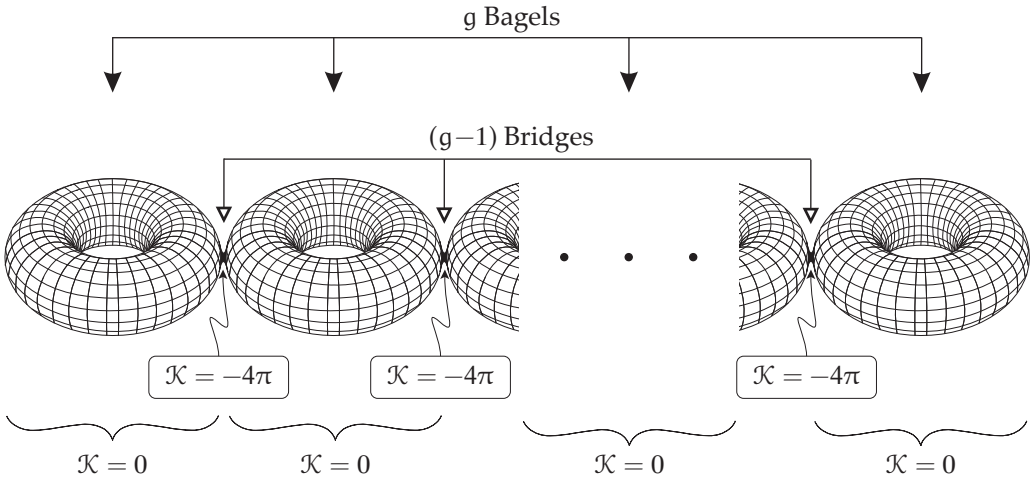
in accord with GGB.

PUZZLE: Place three bagels at the vertices of a large triangle, then connect them together with *three* wormholes along the edges of the triangle, forming a single closed surface $\mathcal{S}$ with $\mathcal{K}(\mathcal{S}) = 3\,\mathcal{K}_{\text{wormhole}} = -12\pi$. Doesn't this violate GGB?! (Further examples along these lines can be found in Ex. 22.)

16.5 The Topological Degree of the Spherical Map

By now we hope that our examples have made it clear that we will have understood GGB *if* we can understand this:

> *Regardless of the form that the surface $\mathcal{S}_g$ takes, $n(\mathcal{S}_g)$ will always cover almost every point of the sphere $(1-g)$ times, provided we count the layers* algebraically, *taking into account positive and negative orientation.* (16.5)



[16.8] *Here $\mathcal{S}_g$ takes the form of g bagels, still joined together in the baking tin by $(g-1)$ bridges. Each bagel has zero total curvature, while each bridge has $\mathcal{K} = -4\pi$. The total curvature is therefore $(-4\pi)(g-1) = 2\pi\chi$.*

In order to clarify this statement, this section will introduce the concept of ***topological degree***,[4] which does the precise algebraic counting of the number of times each point of $\mathbb{S}^2$ is covered. The reason we say "almost every point" is to allow for branch points, as mentioned in the case of the torus.

We begin by noting that (16.5) does *not* mean that there are only $|1-g|$ sheets covering $\mathbb{S}^2$, in all. Indeed, in the case of the thick pancake in [16.6], the covering $n(\mathcal{S}_g)$ of $\mathbb{S}^2$ took the form of one positive covering, and g negative coverings: a total of $(g+1)$ coverings. In other words, for any given point $\tilde{p} = n(p)$ on $\mathbb{S}^2$, there are $(g+1)$ places p_i on the pancake where the normal points in the direction of $\tilde{p}$, i.e., $n(p_i) = \tilde{p}$, for $i = 1, 2, \ldots, (g+1)$.

Let $\mathcal{P}(\tilde{p})$ denote the number of these points p_i on the surface $\mathcal{S}_g$ at which the curvature $\mathcal{K}(p_i)$ is *positive*, so that n is orientation-*preserving*, and a neighbourhood of the image $\tilde{p}$ on the sphere is covered *positively*. Likewise, let $\mathcal{N}(s)$ denote the number points at which $\mathcal{K}(p_i)$ is *negative*, so that n is orientation-*reversing*, and $\tilde{p}$ is covered *negatively*. For example, in the case of the thick pancake [16.6] above, $\mathcal{P}(s) = +1$, independently of $\tilde{p}$; and $\mathcal{N}(s) = g$, also independently of $\tilde{p}$. (Here, and in what follows, we exclude points for which $\mathcal{K} = 0$, as these make no contribution to covering $\mathbb{S}^2$.)

We can now define the ***topological degree*** (or, more commonly, just "degree") of the spherical map, as follows:

Given a closed, oriented surface $\mathcal{S}_g$ of genus g, and a point $\tilde{p}$ on $\mathbb{S}^2$, the **topological degree** *of the spherical map—written $\deg[n(\mathcal{S}_g), \tilde{p}]$—is the algebraic count of the number of times $n(\mathcal{S}_g)$ covers $\tilde{p}$, taking account of orientation:*

$$\deg\left[n(\mathcal{S}_g), \tilde{p}\right] \equiv \mathcal{P}(\tilde{p}) - \mathcal{N}(\tilde{p}), \tag{16.6}$$

where $\mathcal{P}(\tilde{p})$ is the number of preimages of $\tilde{p}$ for which $\mathcal{K} > 0$, and $\mathcal{N}(\tilde{p})$ is the number of preimages of $\tilde{p}$ for which $\mathcal{K} < 0$.

[4] Also called the ***Brouwer degree***, after the Dutch topological pioneer, Brouwer (1881–1966), who was the first to systematically exploit the concept.

In the case that $\mathcal{S}_g$ is the thick pancake, $\mathcal{P}$ and $\mathcal{N}$ are independent of $\widetilde{p}$, and therefore so is the degree:

$$\deg[n(g\text{-holed thick pancake})] = \mathcal{P} - \mathcal{N} = 1 - g.$$

In the case that $\mathcal{S}_g$ is made up of the bridged bagels, [16.8], $\mathbb{S}^2$ is covered $2g$ times by the images of the g bagels, and a further $(g-1)$ times by the images of the bridges. Thus each point of $\mathbb{S}^2$ is covered by $(3g-1)$ layers. Again in this example, the number of coverings is independent of $\widetilde{p}$, and therefore so is the algebraic count of these coverings:

$$\deg[n(g \text{ bagels, bridged})] = \mathcal{P} - \mathcal{N} = g - [g + (g-1)] = 1 - g,$$

just as before.

The key to understanding GGB (at least from the current point of view) is to be able to see that this recurring result is no coincidence, but rather that the degree truly is *topological* in nature—*every* $\mathcal{S}_g$ must satisfy the same equation:

$$\boxed{\deg[n(\mathcal{S}_g)] = \mathcal{P} - \mathcal{N} = (1-g) = \tfrac{1}{2}\chi(\mathcal{S}_g).} \tag{16.7}$$

In short, if we can prove (16.7) then we will have proved GGB, (16.1).

16.6 Historical Note

While Gauss and Bonnet certainly paved the road to GGB, neither one of them was ever even aware of this extraordinary result, let alone stated it!

But the name has stuck. Even those who grasp the name's historical inaccuracy dare not touch it now: it would seem that it matters more that we all agree what a name *means*, than that the name itself be historically accurate.

As we have discussed, in 1827 Gauss announced his discovery of the *Local* Gauss–Bonnet Theorem, (2.6), page 23, stating that if Δ is a geodesic triangle on a general surface, then its angular excess equals its total curvature:

$$\mathcal{E}(\Delta) = \mathcal{K}(\Delta).$$

In fact the ***Local Gauss–Bonnet Theorem*** refers to a generalization of Gauss's original result (by Bonnet in 1848) to the case where the sides of the triangle are no longer required to be geodesics. This adds to the right-hand side of the above equation a term representing the total geodesic curvature of the sides.[5] However, neither of these gentlemen said anything at all about closed surfaces.

It would appear that the honour of discovering GGB actually belongs, in two distinct steps, to Leopold Kronecker and to Walther Dyck.[6] First, in 1869 Kronecker introduced the concept of degree—later clarified and exploited by Brouwer—and proved that $\mathcal{K}(\mathcal{S}_g) = 4\pi \deg(n)$. Second, in 1888 Dyck proved that $\deg(n) = \frac{1}{2}\chi$, thereby completing the proof of GGB in its modern form, (16.1).

[5] This will proved at the end of Act IV: Exercise 6, page 336.

[6] See Hirsch (1976). Occasionally, as in Berger (2010, page 380), Werner Boy is credited with GGB. However, Boy (1903) explicitly gave credit for GGB to Kronecker and to Dyck, while Boy himself generalized GGB to nonorientable surfaces.

Chapter 17

First (Heuristic) Proof of the Global Gauss–Bonnet Theorem

17.1 Total Curvature of a Plane Loop: Hopf's *Umlaufsatz*

To begin to understand the topological nature of the degree of the spherical map (i.e., (16.7)), and with it GGB (i.e., (16.1)), we shall drop down a dimension. That is, in place of the spherical/normal map n of a 2-dimensional surface $\mathcal{S}$ in $\mathbb{R}^3$ to the 2-dimensional sphere $\mathbb{S}^2$, we shall instead consider the normal map N of a 1-dimensional curve $\mathcal{C}$ in $\mathbb{R}^2$ to the 1-dimensional circle $\mathbb{S}^1$.

Since GGB involves *closed* surfaces, we shall correspondingly restrict attention to the case where $\mathcal{C}$ is a *closed* curve. Furthermore, let us begin with the most elementary case, in which $\mathcal{C}$ is not only closed, but is also ***simple***—i.e., without any self-intersections. Let us call such a simple, closed curve a ***loop***.

Although we shall ultimately be concerned with smooth curves that everywhere have a well-defined tangent and normal, let us begin with a triangle, Δ, with internal angles θ_i and external angles φ_i, so that $\theta_i + \varphi_i = \pi$. See [17.1a].

If we imagine a particle travelling counterclockwise round Δ once, then it is clear that it's velocity vector $\mathbf{v}$ executes one full revolution, and this is indeed equivalent to the fundamental fact that the angles of Δ sum to π:

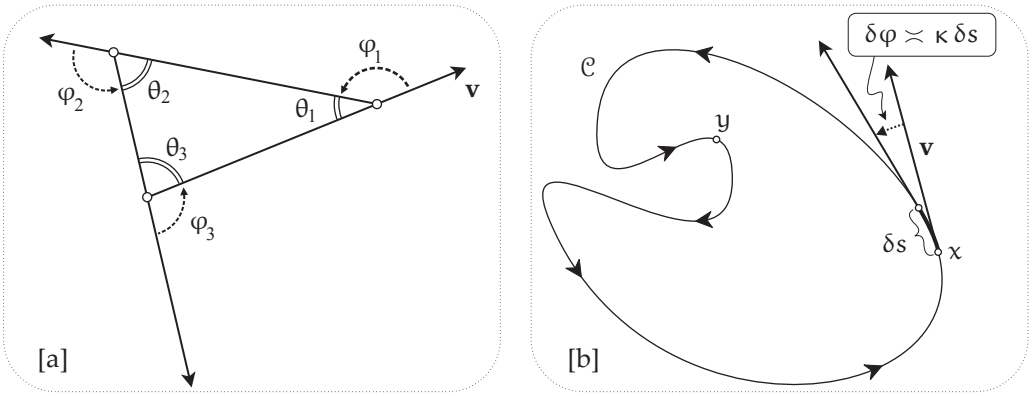
$$\text{net rotation} = \varphi_1 + \varphi_2 + \varphi_3 = 2\pi \quad \Longleftrightarrow \quad \theta_1 + \theta_2 + \theta_3 = \pi.$$

Note that the net rotation statement is somehow the simpler and more fundamental of the two, in the sense that it remains the same if we generalize to n-gons, whereas the equivalent statement in terms of internal angles is now dependent upon n:

$$\text{net rotation} = \sum_{i=1}^{n} \varphi_i = 2\pi \quad \Longleftrightarrow \quad \sum_{i=1}^{n} \theta_i = (n-2)\pi.$$

The concept of *net* rotation becomes important when not all the rotation is in the same direction. This is vividly illustrated by the movement of a nut along a bolt. Suppose you observe the initial position of the nut, then shut your eyes while a friend spins the nut in a complicated combination of positive and negative rotations, moving the nut back and forth along the bolt. When you open your eyes, you have no idea exactly what rotations your friend performed, yet you *do* know what the *net* rotation has been: it is simply measured by how far the nut has moved from its starting position.

Figure [17.1b] considers this net rotation of the velocity $\mathbf{v}$, but now for a smooth loop $\mathcal{C}$, instead of a polygon. Whereas the direction of $\mathbf{v}$ executed sudden jumps of φ_i at the vertices of Δ, its direction now changes smoothly along $\mathcal{C}$, and for the illustrated curve it sometimes turns positively (counterclockwise), as at x, and sometimes turns negatively (clockwise), as at y. Nevertheless, it seems intuitively clear that as the particle travels along $\mathcal{C}$, its velocity $\mathbf{v}$ rocking back and forth along the way, but after one full positive orbit of $\mathcal{C}$, the net effect is that $\mathbf{v}$ has executed one full positive revolution. Again, in speaking of this *net* rotation of $\mathbf{v}$, we mean that negative rotation is allowed to *cancel* positive rotation.



[17.1] *Hopf's* **Umlaufsatz***: As a particle traces a simple loop, its velocity executes one positive revolution. In* **[a]**, $\varphi_1 + \varphi_2 + \varphi_3 = 2\pi$; *in* **[b]**, $\oint_{\mathcal{C}} d\varphi =$ *net rotation of* $\mathbf{v} = 2\pi$.

[17.2] *Heinz Hopf (1894–1971). Photograph by Ernst Ammann, CC BY-SA 4.0*

The fact that **v** executes one full revolution in one orbit of a loop is a theorem called Hopf's ***Umlaufsatz*** (from the German, "Umlauf" (circulation), and "Satz" (theorem)). And while you may not doubt its truth for an uncomplicated curve like that in [17.1b], is it really so obvious for the "simple" curve in [17.3]?! Furthermore, the result does *not* apply to curves that intersect themselves. For example, what is the net rotation [exercise] for a figure-eight curve?

While this result was in some sense known since antiquity, Watson (1917) seems to have been the first to articulate it clearly, while Hopf (1935) was the first to provide a purely topological proof. Hopf's ingenious geometric idea is described in Exercise 23. Here, however, we wish to provide a different, heuristic proof that will serve as an exact model for a corresponding proof of GGB.

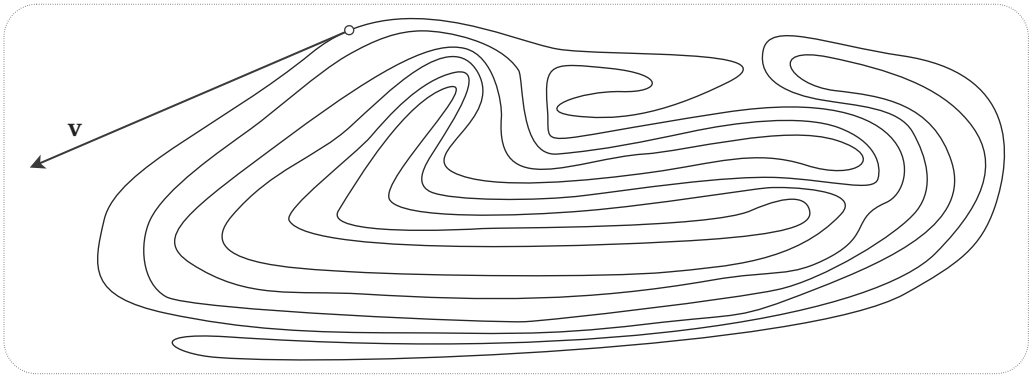
We begin by making explicit connections with our previous discussion of GGB. First, recall that the curvature κ is simply the rate of rotation of **v** with arc length s (or time, if the particle travels at unit speed). Thus, the *Umlaufsatz* can be rephrased in terms of the topological invariance of the *total curvature* of the loop:

$$\oint_{\mathcal{C}} \kappa \, ds = \oint_{\mathcal{C}} d\varphi = \text{net rotation of } \mathbf{v} = 2\pi, \tag{17.1}$$

independently of the shape of $\mathcal{C}$. This bears a striking family resemblance to GGB!

To relate this to the spherical/normal map, let the unit normal to $\mathcal{C}$ be **N**, and, as in [12.1], page 131, view this as a mapping N from the point p of the plane curve $\mathcal{C}$ to the the point $\widetilde{p} = N(p)$ on the unit circle S^1 that lies at the tip of $\mathbf{N}_p$.

While Hopf stated his result in terms of the rotation of the tangent, we may equally well say that that it is **N** that makes one net revolution after orbiting $\mathcal{C}$ once. Equivalently, we may say that, algebraically, $N(\mathcal{C})$ covers S^1 once.



[17.3] *Is it really so obvious that* (net rotation of $\mathbf{v}$) $= 2\pi$*?*

Over the segment δs of $\mathcal{C}$, the directions of the normal vectors are spread over angle $\delta\varphi$, and therefore the tips of these normal vectors fill an arc of length $\delta\widetilde{s} = \delta\varphi$ on S^1. Thus, as we first discussed on page 131,

$$\kappa = \text{local length magnification factor of the N map} \asymp \frac{\delta\widetilde{s}}{\delta s},$$

and therefore,

$$\boxed{\oint_{\mathcal{C}} \kappa \, ds = 2\pi \, [\text{Number of times } N(\mathcal{C}) \text{ covers } S^1].} \tag{17.2}$$

As with our discussion of GGB, we next introduce the ***degree*** of N, to clarify how to algebraically count the the number of times $N(\mathcal{C})$ covers S^1.

Even if $\mathcal{C}$ is not a simple loop, but is allowed self-intersections, we can define the degree of the spherical/normal map N in exactly the same way as before. If $\kappa(p) > 0$ then $N(p)$ is rotating positively (counterclockwise) round S^1 as the particle travels through p on $\mathcal{C}$. Likewise, if $\kappa(p) < 0$ then $N(p)$ is instead rotating negatively (clockwise) round S^1. Then,

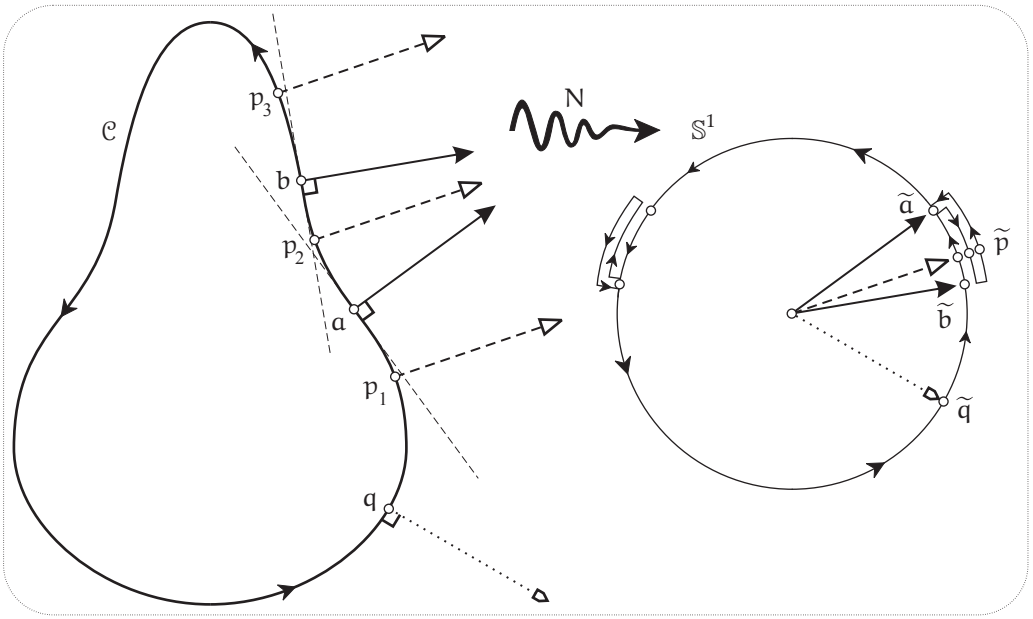
Given a closed curve $\mathcal{C}$, traced counterclockwise, and a point $\widetilde{p}$ on S^1, the degree *of the spherical/normal map* N *is the algebraic count of the number of times* $N(\mathcal{C})$ *covers* $\widetilde{p}$*, taking account of orientation:*

$$\deg\,[N(\mathcal{C}), \widetilde{p}] \equiv \mathcal{P}(\widetilde{p}) - \mathcal{N}(\widetilde{p}),$$

where $\mathcal{P}(\widetilde{p})$ is the number of preimages of $\widetilde{p}$ for which $\kappa > 0$, and $\mathcal{N}(\widetilde{p})$ is the number of preimages of $\widetilde{p}$ for which $\kappa < 0$. (17.3)

With this more precise definition in place, the key fact is that the degree is independent of the choice of $\widetilde{p}$, so that (17.2) takes the form,

$$\boxed{\oint_{\mathcal{C}} \kappa \, ds = 2\pi \deg[N(\mathcal{C})].} \tag{17.4}$$



[17.4] *The* **degree** *of* N *is the algebraic count of how many times* $N(\mathcal{C})$ *covers* S^1. *The sign of* κ *changes at the inflection points* a *and* b, *causing the image on* S^1 *to reverse direction at* $\widetilde{a}$ *and* $\widetilde{b}$. *This in turn can be thought of a* folding *of the orbit back on itself. Since* $\widetilde{p}$ *is traversed three times, twice with positive motion, and once with negative motion, the* net *number of coverings of* $\widetilde{p}$ *is* $2-1=1$.

Hopf's *Umlaufsatz* states that $\deg[N(\text{simple loop})] = +1$, in which case (17.4) reduces to (17.1).

17.2 Total Curvature of a Deformed Circle

Let us illustrate and clarify these ideas with a concrete example. If $\mathcal{C}$ is a circle, then as x orbits $\mathcal{C}$ once, its image $\widetilde{x} = N(x)$ orbits S^1 once, with matching angular speed. Clearly, $\deg[N(\mathcal{C})] = +1$. Now suppose that we gradually and symmetrically deform the circle, so that $\mathcal{C}$ takes the form shown on the left of [17.4], resembling the cross section of a pear.

For the illustrated point $\widetilde{q}$ on S^1 there is precisely one preimage q on $\mathcal{C}$. But for point $\widetilde{p}$ there are instead three preimages p_1, p_2, p_3.

A helpful mental device in locating these preimages of $\widetilde{p}$ is to first observe that the tangent to S^1 at $\widetilde{p}$ (not shown) must be parallel to the tangent to $\mathcal{C}$ at each preimage of $\widetilde{p}$. Now imagine taking this tangent line at $\widetilde{p}$ and letting it move parallel to itself towards $\mathcal{C}$, ultimately sweeping across all of $\mathcal{C}$. Note each time the moving line touches $\mathcal{C}$: these include all the preimages of $\widetilde{p}$, but they also include the preimages of the antipodal point $-\widetilde{p}$. Where [exercise] are the preimages of $-\widetilde{p}$ in [17.4]?

Restricting attention to the right hand side of $\mathcal{C}$, we observe that there are precisely two inflection points, a and b, distinguished by the fact that $\mathcal{C}$ *crosses* the tangent line at these points, and only at these points.

To see that the inflection points play a crucial role in relation to the spherical/normal map N, imagine x starting at q and travelling up the right side of $\mathcal{C}$. Then $\widetilde{x} = N(x)$ travels forward along S^1 from $\widetilde{q}$ through $\widetilde{b}$ and $\widetilde{p}$ until it hits $\widetilde{a}$. But at $\widetilde{a}$ it bounces and travels *backwards*, passing through $\widetilde{p}$ for a *second* time before hitting $\widetilde{b}$. Now it bounces again and resumes its forward motion along S^1, passing through $\widetilde{p}$ a *third* time. Next it arrives at $\widetilde{a}$ for a second time [exercise: where is x at this

moment?] and this time it passes right through $\widetilde{a}$ and keeps going. We strongly suggest that you follow x in your mind as it completes a full orbit of $\mathcal{C}$, with $\widetilde{x}$ performing another back-and-forth motion as x traverses the left-hand side of $\mathcal{C}$.

Now comes a crucial mental leap in the visualization of this motion: *think of the motion of* $\widetilde{x}$ *as the orbit of a bead travelling along a continuous, unbroken thread.* Thus, as illustrated, when $\widetilde{x}$ first arrives at $\widetilde{a}$ and starts to move backwards on S^1, *it can only be because the thread is folded back on itself.* Likewise, when the bead next arrives at $\widetilde{b}$ and reverses course again, resuming its forward motion along S^1, it can only be because the thread has folded back on itself a *second* time. Thus, as illustrated, as x passes through p_1, then p_2, and finally p_3, $\widetilde{x}$ passes through $\widetilde{p}$ three times, first forward, then backward along the folded thread, then forward again along the twice-folded thread.

In reality, all three folded segments of the thread between $\widetilde{a}$ and $\widetilde{b}$ are plastered right down on top of each other on S^1, but we have lifted them away slightly in order to reveal the folds and the three *different* places on the thread that correspond to the single point $\widetilde{p}$ of S^1.

The positive curvature κ at q gave rise to positive motion through $\widetilde{q}$. Thus, since q is the *only* preimage of $\widetilde{q}$, $\mathcal{P}(\widetilde{q}) = 1$ and $\mathcal{N}(\widetilde{q}) = 0$, and therefore $\deg[N(\mathcal{C}), \widetilde{q}] \equiv \mathcal{P}(\widetilde{q}) - \mathcal{N}(\widetilde{q}) = 1$, as anticipated.

Likewise, the positive curvature κ at p_1 and p_3 gave rise to positive motion through $\widetilde{p}$, while the negative value of κ at p_2 gave rise to negative motion through $\widetilde{p}$. Thus $\mathcal{P}(\widetilde{p}) = 2$ and $\mathcal{N}(\widetilde{p}) = 1$, and therefore $\deg[N(\mathcal{C}), \widetilde{p}] \equiv \mathcal{P}(\widetilde{p}) - \mathcal{N}(\widetilde{p}) = 1$, as before.

17.3 Heuristic Proof of Hopf's *Umlaufsatz*

Let us now apply what we have learned from this example to the general case. Imagine beginning with a circle, and allowing it to continuously deform and evolve into the most general simple closed curve. Here we are thinking of the form of the curve as a function $\mathcal{C}(t)$ of time, with time t going from 0 to 1, say, and with $\mathcal{C}(0)$ being the initial circle, and $\mathcal{C}(1)$ being the final curve, such as $\mathcal{C}$ in [17.4].

But we restrict this evolution to exclude self-intersections, and to be such that κ is well-defined everywhere on the intermediate curves: this is necessary if we wish our mapping N to be continuous on each $\mathcal{C}(t)$. If a sharp corner were to develop, for example, then N would have a jump discontinuity there.

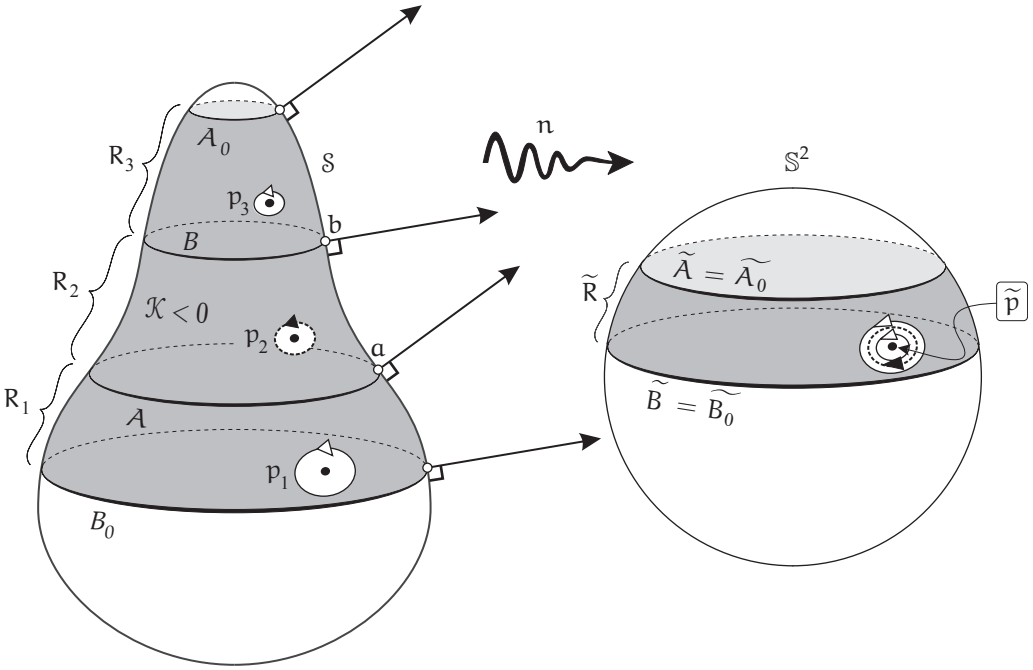
As $\mathcal{C}(t)$ evolves continuously, so too does the orbit of $\widetilde{x}$, as embodied and visualized as a stretchable, unbreakable, foldable string wrapped around S^1. Whenever an inflection point develops in $\mathcal{C}(t)$, signaling a change in the sign of κ there, the image string develops a fold.

If $\widetilde{x}$ stays away from such folds, both $\mathcal{P}$ and $\mathcal{N}$ remain separately constant, and therefore so too does their difference, $\deg[N[\mathcal{C}(t)], \widetilde{x}]$. However, if $\widetilde{x}$ crosses a point of folding, we either gain two new layers of opposite orientation, or we lose two layers of opposite orientation. But, this means that $\mathcal{P}$ and $\mathcal{N}$ both increase by 1 or they both decrease by 1. In either event, the crucial observation is that,

> *Changes in the sign of* κ *on* $\mathcal{C}$ *cause folding of* $N(\mathcal{C})$ *on* S^1*, but this folding has no effect on the number of layers (counted algebraically) that cover* S^1.

That is, $\deg[N\{\mathcal{C}(t)\}] = \mathcal{P}(\widetilde{x}) - \mathcal{N}(\widetilde{x})$ remains constant as we cross a fold. Thus $\deg[N\{\mathcal{C}(t)\}]$ is a well-defined property of the curve as a whole, independent of $\widetilde{x}$.

The value of $\deg[N\{\mathcal{C}(t)\}]$ must vary continuously with time, *and* remain an integer: thus it cannot change at all. But, initially, $\deg[N\{\mathcal{C}(0)\}] = 1$. Therefore, since the final value of the degree is the same as its initial value, $\deg[N\{\mathcal{C}(1)\}] = 1$, as Hopf asserted.



[17.5] *The* **degree** *of* n *is the algebraic count of how many times* $n(\mathcal{S})$ *covers* $\mathbb{S}^2$. *The sign of* $\mathcal{K}$ *changes as we cross* A *and* B, *causing the image on* $\mathbb{S}^2$ *to reverse orientation as we cross* $\widetilde{A}$ *and* $\widetilde{B}$. *Since,* $\widetilde{p}$ *is covered three times, twice with positive orientation, and once with negative orientation, the* net *number of coverings of* $\widetilde{p}$ *is* $2-1=1$.

17.4 Total Curvature of a Deformed Sphere

In [17.4] we likened $\mathcal{C}$ to a cross section of a pear. We now turn our attention to the surface of the pear itself!

That is, let us take the right half of $\mathcal{C}$ and rotate it 2π about the vertical axis through its ends to generate the pear-like surface of revolution $\mathcal{S}$ shown in [17.5]. As is visually evident (and as was previously noted in (10.13), p. 114) the segments of $\mathcal{C}$ with $\kappa > 0$ generate elliptic regions of $\mathcal{S}$ with $\mathcal{K} > 0$, and the segments of $\mathcal{C}$ with $\kappa < 0$ generate hyperbolic regions of $\mathcal{S}$ with $\mathcal{K} < 0$. The inflection points a and b rotate to generate two circles A and B (both parabolic; i.e., with $\mathcal{K} = 0$) that separate the positively and negatively curved parts of the pear's surface.

But the figure also illustrates the fact that there are two other circles, A_0 and B_0, that have the same images under the spherical map as A and B: $\widetilde{A} = \widetilde{A_0}$ and $\widetilde{B} = \widetilde{B_0}$. These four circles divide $\mathcal{S}$ into a top region, a bottom region, and the three regions in between, which we have shaded grey and labelled R_1, R_2, R_3. All three of these regions R_i are mapped to the *same* region $\widetilde{R}$ on $\mathbb{S}^2$.

Since $\mathcal{K} > 0$ on R_1 and R_3, the spherical map is orientation-preserving in those regions, and a point in those regions is therefore mapped to a point on $\mathbb{S}^2$ that is covered *positively*. On the other hand, since $\mathcal{K} < 0$ on R_2, the spherical map is orientation-reversing in that region, and a point in that region is is therefore mapped to a point on $\mathbb{S}^2$ that is covered *negatively*.

In particular, the figure illustrates the three preimages p_1, p_2, p_3 of a point $\widetilde{p}$ in $\widetilde{R}$. Thus, since $\widetilde{p}$ is covered three times, twice with positive orientation, and once with negative orientation, the *net* number of coverings of $\widetilde{p}$ is $2-1=1$.

Let us not lose sight of original formulation of GGB in terms of curvature, and our quest to understand why the total curvature of a closed surface is topologically invariant. In the case of our deformed sphere $\mathcal{S}$, the total curvature residing in the unshaded top of the pear is the area of the

northern polar cap of S^2; likewise, the total curvature in the unshaded bottom of S is likewise the area of the larger unshaded southern region. Each of the three shaded regions contains the *same* amount of curvature, namely, the area of $\widetilde{R}$. But since $K < 0$ on R_2, n is orientation reversing there, and therefore its total curvature is the *negative* of the area of $\widetilde{R}$. Integrating the curvature over all of S therefore yields the whole area of S^2, once: i.e., 4π.

17.5 Heuristic Proof of the Global Gauss–Bonnet Theorem

In the lower-dimensional case, we gained a much more intuitive understanding of the degree by thinking of the spherical image of the curve as being a stretchable, foldable thread covering S^1. In the present case, it is likewise very helpful to *imagine* $n(S)$ *as being a highly stretchable, unbreakable,* foldable *membrane covering* S^2—perhaps think of very thinly rolled out pizza dough. But understand that this special mathematical dough can not only be stretched out in whatever manner we require, but it can also *contract* just as easily, as needed.

On the left of [17.5], think of the entire surface of the pear S as being covered with such a thin layer of pizza dough, loosely sticking to the surface. Let us not worry for now about the *geometry* of the spherical map, and instead focus solely on the *topology* of how it maps the grey region $R_1 \cup R_2 \cup R_3$ on S to the grey region $\widetilde{R}$ on S^2. See [17.6].

The arrows in this diagram indicate the sequence of topological transformations to be carried out. First, suppose we tie two pieces of string around A and B_0 to keep them in place, and then stretch B radially outward till it is the same size as B_0, and likewise stretch A_0 radially outward till it is the same size as A, stretching R_2 and R_3 into two illustrated portions of cones.

Next, leaving new A_0 in place, let us move the new B vertically downward and stick it onto B_0, *folding* R_2 on top of R_1 in the process. Note, crucially, how this folding causes the positive circulation around p_2 to be *reversed*. The fate of the large letter "T" in R_2 serves to stress this point.

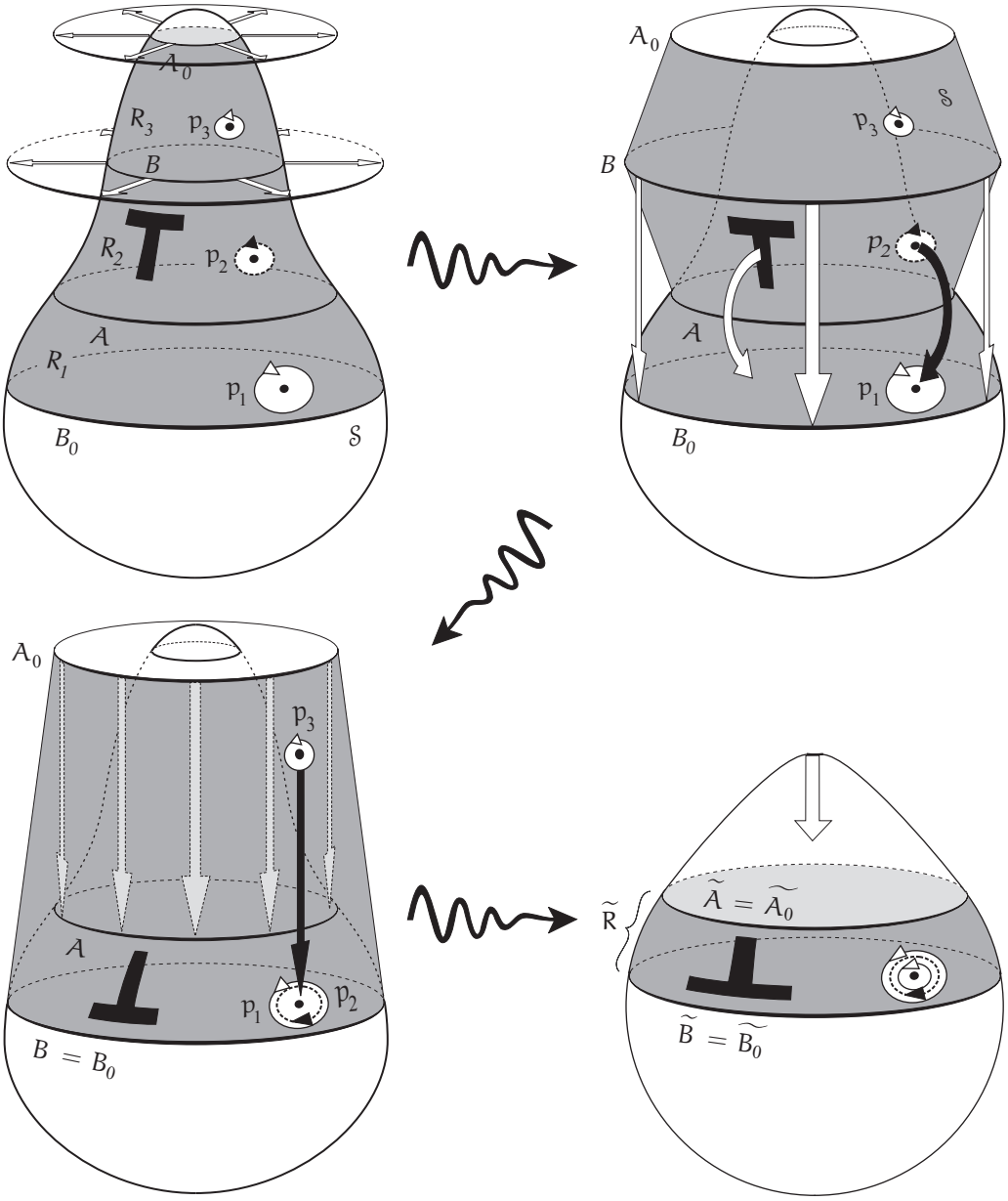
Now move the new A_0 vertically downward and stick it onto A, folding R_3 on top of the new R_2 (and R_1) in the process.

At this point we have correctly embodied the topology of the spherical image, and we can imagine doing a final geometric housekeeping, pushing the top of the pear downward to make it spherical (and of unit radius), and stretching each of the three folded grey layers to get the *geometry* of the spherical map correct, so that, in particular, $\widetilde{p}_1$, $\widetilde{p}_2$, and $\widetilde{p}_3$ will all end up on top of each other, but with their orientations having been reversed between successive layers by the folding.

The details of how we arrived at this final state are irrelevant; our specific intermediate transformations were chosen merely to make this net transformation easier to imagine and draw. Also, this example involved a slight sleight of hand: if $g \neq 0$ then $n(S_g)$ must have multiple layers, and we cannot imagine starting with a single layer of dough covering S_g and manipulating it to obtain $n(S)$.

At this point we can instead try to imagine directly (without intermediate steps) the effect of the spherical map on the evolving surface S. If we start with a spherical S_0 and gradually deform it into the pear-shaped surface in [17.5], then we can imagine the mathematical dough $n(S_0)$ covering S^2 following a corresponding evolution. As negative curvature emerges on the deformed S_0 at parabolic curves, so the spherical image flows backwards over itself at the images of these parabolic curves, producing folds there. But as we cross such a fold in $n(S_0)$ on S^2 we always gain or lose *two* new layers of *opposite* orientation, so the algebraic sum of the number of coverings is unaltered.

This same reasoning applies if we instead start with some S_g (where $g \neq 0$) and allow it to evolve. For example, suppose that initially S_g takes the form of the g bridged bagels of [16.8]. Now suppose we press inward with our thumb on the outer (positively curved) surface of one of the bagels, creating a depression centred at a point p. As we do so, $n(S_g)$ will undergo a



[17.6] *The topological consequence of a change of sign of $\mathcal{K}$ is* folding *of the spherical image. As we cross the boundary between regions of opposite curvature, the spherical image either gains or loses* two layers of opposite orientation*, so the algebraic count of the number of coverings of* $\mathbb{S}^2$ *(i.e., the degree) remains constant, and the total curvature is therefore a topological invariant.*

corresponding evolution in the vicinity of $\widetilde{p}$, lying on one of the g sheets that positively covers $\mathbb{S}^2$. That single sheet, out of the total of $(3g-1)$ sheets, will undergo an evolution like that described in the previous paragraph. This single sheet containing $\widetilde{p}$ will fold over on itself, creating two new layers of opposite orientation in a topological annulus centred at $\widetilde{p}$, but (algebraically) this folded region will continue to cover $\mathbb{S}^2$ once positively, and therefore the net number of coverings of $\mathbb{S}^2$ will remain $(1-g)$, even in the vicinity of $\widetilde{p}$.

This concludes our first (heuristic) explanation of GGB.

Chapter 18

Second (Angular Excess) Proof of the Global Gauss–Bonnet Theorem

18.1 The Euler Characteristic

Thus far we have introduced the Euler characteristic χ as merely a convenient alternative means of labelling a closed, orientable surface, $\mathcal{S}_g$, of a given genus g. However, χ actually has a definition and meaning all its own, which applies to a much wider class of objects than closed, orientable surfaces. Once we have explained this deeper meaning of χ, we may apply it to $\mathcal{S}_g$, in particular, and it is then an important *theorem* (not a definition) that $\chi(\mathcal{S}_g) = 2 - 2g$.

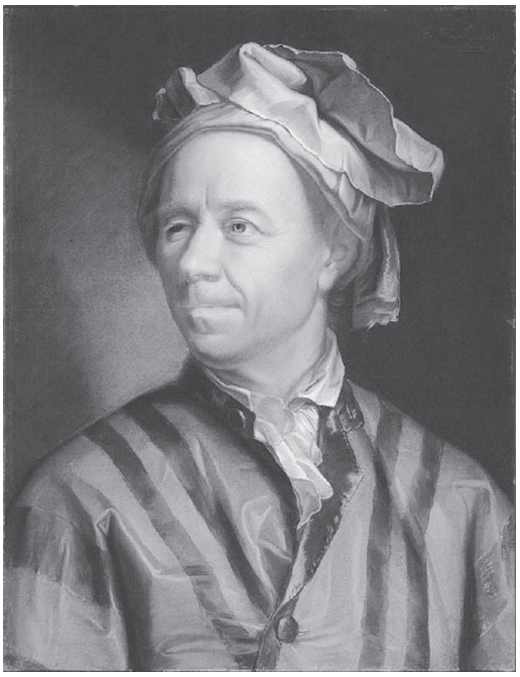
18.2 Euler's (Empirical) Polyhedral Formula

The story of Euler's characteristic begins on the 14th of November, 1750, not with smooth surfaces, but rather with polyhedra. On that date, Euler—pictured in [18.1]—wrote a letter to Christian Goldbach,[1] outlining a remarkable *empirical* discovery, which Euler finally succeeded in proving two years later.[2]

The discovery rested on a realization that may seem strangely obvious to modern eyes: Euler was the first to clearly recognize the very existence of a polyhedron's ***vertices***, ***edges***, and ***faces***. As he wrote in the letter,

> Therefore three kinds of bounds are to be considered in any kind of body, namely 1) points, 2) lines, and 3) surfaces, with the names specifically used for this purpose.

Before Euler, mathematicians focused instead on the magnitude of the solid



[18.1] *Leonhard Euler (1707–1783)*

[1] Goldbach is best known for a letter he wrote *to* Euler in 1742. In it, he conjectured that *every even number greater than* 2 *is the sum of two primes*; this is now known as ***Goldbach's Conjecture***. While it is believed to be true, and (as of 2017) has been verified by computer up to $4,000,000,000,000,000,000$, it remains unproven after almost 300 years.

[2] For a masterful, mathematically accurate, yet riveting account of this history and the connected mathematical ideas, see Richeson (2008). Also, see *Further Reading* at the end of this book.

angle contained between the faces that met at a vertex, rather than on the vertex itself. If we gradually change the dimensions of a polyhedron, the solid angles will vary continuously, but the discrete count V of the *number* of vertices will remain constant. This discrete count V is what Euler now seized upon.

Likewise, before Euler, nobody had given much thought to what Euler described as "the junctures where two faces come together along their sides, which, for lack of an accepted term, I call ***edges***." Euler then went on to count the *number* E of these edges, and the *number* F of faces. This shift from looking at polyhedra in terms of continuously varying lengths and angles, to looking instead at discrete topological features, was yet one more instance of Euler's genius.

Having made this subtle but profound leap, it did not take Euler long to detect a remarkable empirical pattern. Figure [18.2] illustrates the five Platonic solids, and next to each are listed the values of V, F, $(V+F)$, and E. As is now obvious, and as Euler was the first[3] to announce, the $(V+F)$ entries exceed the E entries by 2. In all cases, as Euler wrote to Goldbach, we have

$$\textbf{\textit{Euler's original formula}}: \qquad \boxed{V+F=E+2.}$$

While we have merely verified this for the five Platonic solids, Euler tested many more non-Platonic examples, and in his letter to Goldbach in 1750 he stated his conviction that it was a *universal* truth about polyhedra, though he confessed he had no idea how to prove it at the time. The result is now known as ***Euler's Polyhedral Formula***. Only gradually did mathematicians discover the precise *limits* of its universality: in fact the formula only applies if the polyhedron is *topologically spherical*.

The above formula is not, however, the modern form of the result. If we move the E to the other side of the equation, we obtain a result that is trivially equivalent, algebraically speaking. However, this seemingly trivial step represents a highly *non*trivial conceptual advance (first taken by Poincaré in 1895), for the left hand side of the equation is now *a single integer that characterizes the polyhedron* $\mathcal{P}$, and *this* is our new definition of its

$$\textbf{\textit{Euler characteristic}}\textbf{:} \qquad \boxed{\boxed{\chi(\mathcal{P}) \equiv V-E+F.}} \tag{18.1}$$

Armed with this concept, here is the modern form of ***Euler's Polyhedral Formula***:

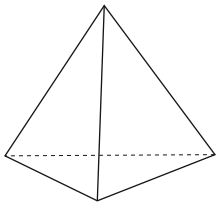
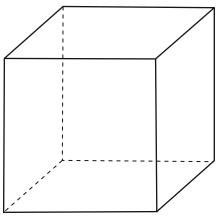
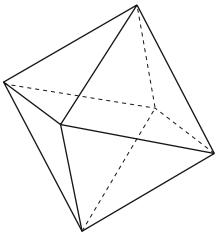
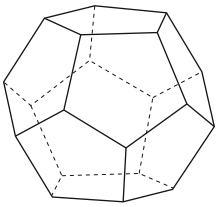
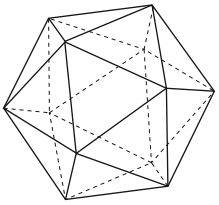
$$\boxed{\boxed{\chi(\text{topologically spherical polyhedron}) = 2.}} \tag{18.2}$$

This is in accord with our previous definition of $\chi(\mathcal{S}_g) = 2-2g$, with $g=0$, but the key difference is that now we will be able to *prove* that (18.2) follows from (18.1). In fact, we shall present two quite different proofs.

Euler himself did ultimately succeed in providing the very first proof of his own formula, and in some ways his proof is more topologically natural than the two that we shall instead present, but there are subtle obstacles that must be overcome in order to make his argument completely convincing. (See Richeson 2008, Ch. 7, for details.)

---

[3] In 1860, more than a century after Euler announced his wonderful discovery, a long-lost manuscript miraculously surfaced, *two centuries* after it had been written. It revealed that *Descartes* had made essentially the same discovery as Euler (but in a different form) as early as 1630, more than a century before Euler! The fascinating story is well told in both Stillwell (2010, p. 469) and Richeson (2008, Ch. 9).

Platonic Solid	Name	V	F	V + F	E
	Tetrahedron	4	4	8	6
	Cube	8	6	14	12
	Octahedron	6	8	14	12
	Dodecahedron	20	12	32	30
	Icosahedron	12	20	32	30

[18.2] *Euler's Polyhedral Formula, empirically verified for all five Platonic solids:* $V + F = E + 2$.

Although we have not yet provided *any* proof of (18.2), we end this section by noting that it has many consequences, one of the most striking of which concerns the five Platonic solids shown in [18.2]. First, recall that Euclid's *Elements*[4] provided a *geometric* proof (outlined in Ex. 24) that these five are the *only* "regular polyhedra" that can exist. A regular polyhedron is one for which all the faces are congruent regular polygons, and the same number of these regular polygons meet at each vertex.

[4] Although Euclid is responsible for publishing the proof in his *Elements*, the proof itself is believed to be due to Theaetetus of Athens (a friend of Plato), dating from around 400 BCE.

Now suppose that we completely relax the *geometrical* constraints, and only insist that each "face"—which we now imagine to be curved, bendable, and stretchable—has the same *number* of wavy edges, and that an equal number of these irregular, bendy faces meet at each vertex. Exercise 25 uses (18.2) to demonstrate the remarkable fact that it is *still* true that there are only five *topological* possibilities, each one being topologically indistinguishable from one of the five Platonic solids of antiquity. Thus the mesmerizing *geometrical* beauty and regularity of the five Platonic solids turns out to have been an extraordinary, 2000-year-long red herring!

18.3 Cauchy's Proof of Euler's Polyhedral Formula

18.3.1 Flattening Polyhedra

From the time of the Ancient Greeks, all the way through the eighteenth century, polyhedra were viewed as solid objects; indeed, we still speak of Platonic *solids*. Cauchy, in 1813, appears to have been the first to make the conceptual leap of peeling the polyhedron off the solid it bounds, viewing it as a hollow surface in its own right.

His next crucial step towards proving Euler's Polyhedral Formula (18.2) was to *flatten* this hollow surface onto the plane. Cauchy was somewhat vague about precisely how this was to be accomplished, and Richeson (2008, Ch. 12) does an admirable job of elucidating both Cauchy's approach, and subsequent clarifications by other mathematicians.

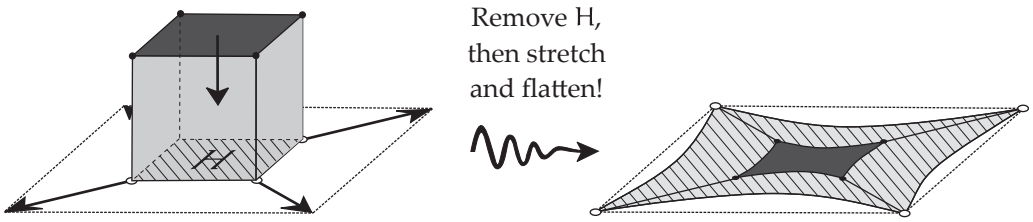
One thing is certain, though: Cauchy's thinking was still firmly planted in *geometry*, and his polyhedra necessarily had straight edges and plane faces, and their flattened versions were also required to have straight edges. Additionally, his proof required the polyhedron to be ***convex***: this stringent geometrical requirement means that if your eye is *anywhere* inside the polyhedron, you can see the *entire* surface—there's nowhere to hide inside a convex polyhedron! But convexity is clearly *not* a topological condition, and therefore it cannot be essential to the validity of Euler's Polyhedral Formula. We therefore choose to sidestep all these historical artifacts, providing instead a more modern, purely topological version of Cauchy's argument, but one that remains faithful to his brilliant original insight.

To this end, imagine that the faces of the polyhedron are made of a bendable, stretchable *rubber sheet*, and picture the vertices and edges as nothing more than dots and connecting curves drawn in pen on this closed, topologically spherical, polyhedral balloon. Even if we start with a classical, rectilinear polyhedron, we may now continuously deform it (without cutting or joining any of its parts) and the resulting curved surface will have the same number of vertices, edges, and faces drawn upon it as did the original.

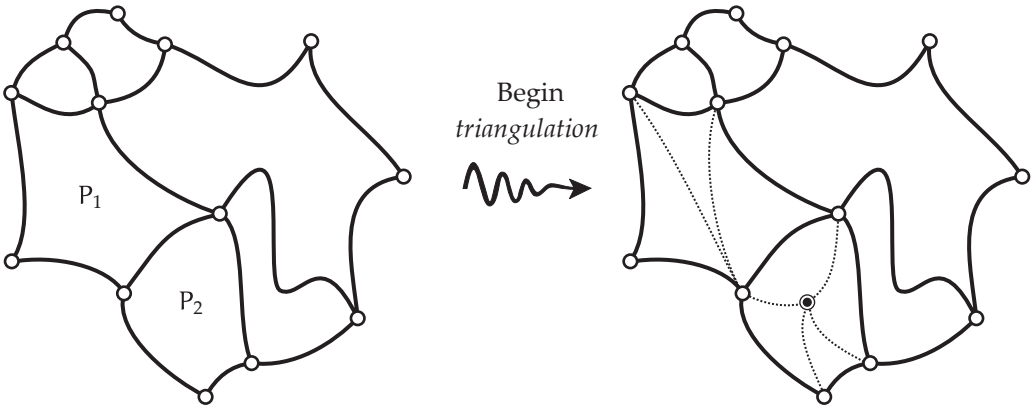
To flatten the polyhedron, imagine once again that we start with the classical rigid surface with plane faces. Now cut out and discard one of these faces, H, and place the polyhedron on a flat surface, H-side down. The left-hand side of [18.3] illustrates this for a cube. Next, return to picturing the polyhedron as a rubber sheet (now with the hole H), the vertices being dots drawn on its surface. Take the vertices that bound the hole H, and pull each of them radially outward within the plane, so that the boundary of the hole gets larger and larger, and the rest of the polyhedron is pulled down and ultimately stretched out flat within the expanding hole H. See the right-hand side of [18.3].

After flattening the polyhedron in this manner, we may further deform the edges (while keeping them in the plane) so that they take on any shape we please, and we do so with topological impunity, for such a deformation will not change V, E, or F. We shall call the resulting figure a ***polygonal net***[5].

[5]More traditional names are ***network*** or ***graph***, but the usage of these terms allows for "polygons" with only two sides, whereas we shall insist that our polygons have at least three sides.



[18.3] *To flatten the cube, we remove its bottom face* H*, then stretch out the resulting hole (hatched) until we have pulled the remaining faces down into the plane.*



[18.4] *The Euler characteristic of a polygonal net is unaltered by triangulation.*

18.3.2 *The Euler Characteristic of a Polygonal Net*

Figure [18.4] illustrates a more typical example of a polygonal net, obtained by flattening a non-Platonic polyhedron.

As Cauchy reasoned, in the process of flattening the polyhedron, we have removed one face, and we have not altered the number of vertices or edges:

$$V \rightsquigarrow V, \quad E \rightsquigarrow E, \quad F \rightsquigarrow F-1 \qquad \Longrightarrow \qquad \chi \rightsquigarrow \chi - 1.$$

In short, flattening the polydedron has reduced its Euler characteristic by 1.

It follows that to prove Euler's Polyhedral Formula (18.2), it suffices to prove *this* elegant result:

$$\boxed{\chi(\text{polygonal net}) = 1.} \tag{18.3}$$

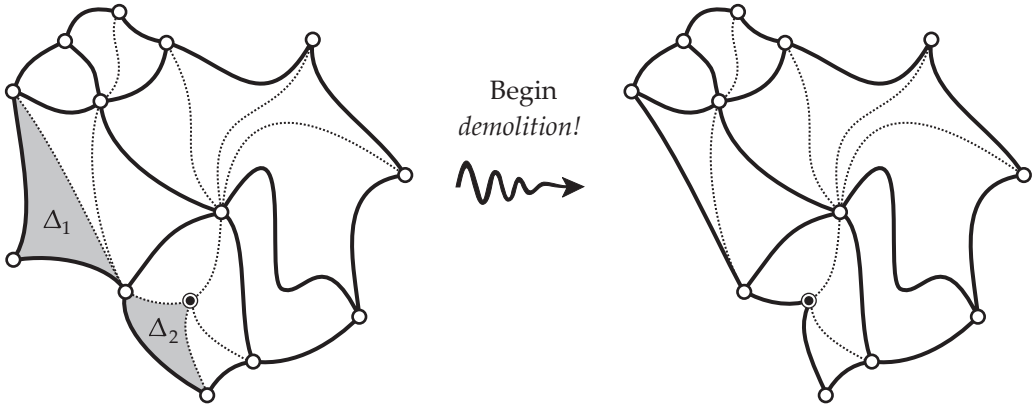
You may care to check that this is true for the specific example we have drawn.

To show that this is always true, we begin by observing that we may dissect each n-gon (with $n > 3$) into triangles, and that, crucially, this process of ***triangulation*** does not alter the value of χ.

The figure illustrates two approaches to the dissection. In polygon P_1 (which we suppose to be an n-gon) we draw in the $(n-3)$ curvilinear "diagonals" from one vertex to all the others, splitting this one face into $(n-2)$ triangles, for a *net* increase (pardon the pun!) of $(n-3)$ faces. Thus,

$$\text{in case } P_1\text{:} \quad V \rightsquigarrow V, \quad E \rightsquigarrow E + (n-3), \quad F \rightsquigarrow F + (n-3) \qquad \Longrightarrow \qquad \chi \rightsquigarrow \chi.$$

In P_2 (which we suppose to be an m-gon) we have instead added a vertex somewhere inside, and have then joined this to the m vertices, creating m new edges. This splits this one face into m



[18.5] *The Euler characteristic of a triangulation is unaltered by its demolition.*

triangles, for a net increase of $(m-1)$ faces. Thus,

$$\text{in case } P_2\text{:}\quad V \rightsquigarrow V+1,\quad E \rightsquigarrow E+m,\quad F \rightsquigarrow F+(m-1) \qquad \Longrightarrow \qquad \chi \rightsquigarrow \chi.$$

The left-hand side of [18.5] shows one possible way of completing the triangulation, and we have just proved that its Euler characteristic must be the same as that of the original polygonal net.

Having gone to all the trouble of constructing this triangulation, the final step in the proof of (18.3) is to *demolish* it! That is, we shall gradually erase each of the triangles, one by one, until only one triangle remains standing. In order to reduce the number of cases that need be considered, we shall agree to only nibble away triangles around the edges; we shall *not* burrow through the middle of the net, which could split it into two distinct, disconnected islands.

With this agreement in place, there are only two cases to consider: either a boundary triangle shares *one* (dotted) edge with the interior of the net (e.g., Δ_1), or else it shares *two* (dotted) edges (e.g., Δ_2).

First, suppose that we

$$\text{erase } \Delta_1\text{:}\quad V \rightsquigarrow V-1,\quad E \rightsquigarrow E-2,\quad F \rightsquigarrow F-1 \qquad \Longrightarrow \qquad \chi \rightsquigarrow \chi.$$

Second, suppose that we

$$\text{erase } \Delta_2\text{:}\quad V \rightsquigarrow V,\quad E \rightsquigarrow E-1,\quad F \rightsquigarrow F-1 \qquad \Longrightarrow \qquad \chi \rightsquigarrow \chi.$$

Ultimately, only one triangle Δ will remain standing, and its Euler characteristic will be

$$\chi(\Delta) = V - E + F = 3 - 3 + 1 = 1.$$

Since χ is invariant under the triangulation process [18.4], and under the demolition process [18.5], its initial value must equal its final value, 1, so we have proved (18.3). And with this, we have also completed Cauchy's proof of Euler's Polyhedral Formula.

18.4 Legendre's Proof of Euler's Polyhedral Formula

Recall Harriot's beautiful result (1.3), page 8, relating the angular excess $\mathcal{E}$ of a geodesic triangle on the sphere to its area $\mathcal{A}$. As you proved in Exercise 5, page 83, this can be generalized to geodesic

n-gons. A Euclidean n-gon has angle sum $(n-2)\pi$, and therefore the *angular excess* $\mathcal{E}$ of a geodesic n-gon on a curved surface is

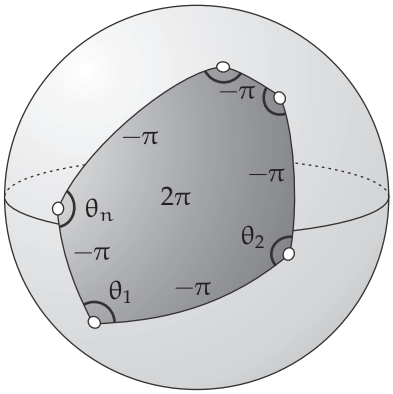
$$\mathcal{E}(n\text{-gon}) = [\text{angle sum}] - (n-2)\pi. \tag{18.4}$$

From the proven additivity of $\mathcal{E}$, Harriot's result then yields,

$$\frac{1}{R^2}\mathcal{A}(n\text{-gon}) = \mathcal{E}(n\text{-gon}) = [\text{angle sum}] - n\pi + 2\pi, \tag{18.5}$$

where R is the radius of the sphere.

Figure [18.6] provides a valuable visualization[6] of the expression for the angular excess on the right hand side of this formula. *Inside* the n-gon, (A) mark the interior angles $\theta_1, \theta_2, \ldots, \theta_n$; (B) write "$-\pi$" next to each edge; (C) write "2π" in the middle. Then $\mathcal{E}$ is the sum of everything in the picture: (A) + (B) + (C).



[18.6] *The value of $\mathcal{E}$ for the geodesic n-gon can be visualized as the sum of everything in the picture.*

Legendre presented an ingenious proof in 1794, the first step of which was to project the polyhedron onto a sphere. The specific way in which he carried out this projection (described in Ex. 26) required him to assume that the polyhedron was *convex*, just as Cauchy would also do, 20 years later. But, as we have noted in the context of Cauchy's proof, convexity cannot actually be relevant to a topological result. Therefore, once again, we shall sidestep this historical artifact and present a more blatantly topological version of the argument—one that does not hinge on convexity, but that stays true to Legendre's essential insight.

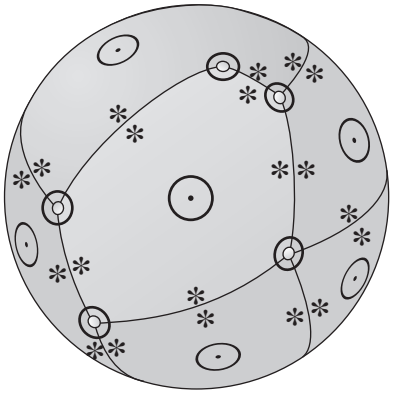
As before, imagine our polyhedron $\mathcal{P}$ to be a curved, topologically spherical, rubber membrane, with dots (vertices) and connecting curves (edges) drawn simply on its surface. Instead of collapsing the polyhedron, as we did in Cauchy's proof, let us this time inflate it like a balloon! We thereby arrive at a polygonal net covering the surface of an ultimately spherical balloon. Note that the edges that result from this are *not* geodesics, but, as we now explain, we can easily make them so without altering the Euler characteristic.

Imagine this sphere to be rigid, with small nails driven in at each of the vertices we have just constructed. Next, picture the current, *non*geodesic edges as being made of stretched elastic strings, temporarily held in place and attached to the nails (vertices) at their ends. Further imagine that there is no friction between these elastic strings and the sphere: they are free to slide over the surface without resistance. Now let us hammer the nails down flat onto the surface, so that any string that is not attached to this particular nail can sweep over its location without getting caught on it.

Finally, *release* the strings! They will automatically contract to create the shortest (= geodesic = great circle) routes between the nails. We have thus arrived at a *geodesic* polygonal net that completely covers the sphere, and its Euler characteristic must the same as the original value, $\chi(\mathcal{P})$.

Now we come to the second part of Legendre's ingenious idea: sum both sides of the formula (18.5) over *all* the geodesic polygons P_j that make up this geodesic, sphere-covering net, an example of which is shown in [18.7]. Let us look first at the sum of the right-hand side. The efficacy of the visualization [18.6] is now evident, for it enables us to *see* the value of this sum for the complete net, as follows.

[6] We hit upon this idea independently, but have since found it published by Richeson (2008, p. 95).



[18.7] $\sum_j \mathcal{E}(P_j)$ *is the sum of everything in this picture. Here,* $* = -\pi$, *and* $\odot = 2\pi$.

To avoid clutter, [18.7] employs two visual abbreviations: $* = -\pi$, and $\odot = 2\pi$. Focus attention on any one vertex: we are summing the interior angles θ_j of *all* the polygons, so, in particular, we are summing all the angles that surround this vertex, yielding 2π. Since each vertex contributes 2π, the total contribution of the interior angles is $2\pi V$. Next, each edge has a $*$ (i.e., $-\pi$) written *on either side* of it, yielding -2π per edge. The total contribution of the $-\pi$'s is therefore $-2\pi E$. Finally, each face carries a $\odot$ (i.e., 2π), yielding a total of $2\pi F$. Thus,

$$\boxed{\sum_j \mathcal{E}(P_j) = 2\pi\,[V - E + F] = 2\pi\,\chi(\mathcal{P}).} \tag{18.6}$$

Since the polyhedral net completely covers the surface of the sphere, summing the left-hand side of (18.5) yields

$$\frac{1}{R^2}\sum_j \mathcal{A}(P_j) = \frac{1}{R^2}\,[\text{area of the sphere}] = 4\pi.$$

Finally, (18.5) equates these two quantities:

$$2\pi\,\chi(\mathcal{P}) = 4\pi \qquad \Longrightarrow \qquad \chi(\mathcal{P}) = 2,$$

thereby completing Legendre's proof of Euler's Polyhedral Formula.

Despite its evident beauty, Legendre's proof feels morally wrong, for it achieves success (perversely) by means of continuously varying geometrical angles that are *meaningless* within topology. But from the point of view of Differential Geometry, it is perhaps the *right* kind of proof, for it appears to link geometry and topology in a surprising way that might help to explain GGB. As we shall see shortly, this optimism is justified.

18.5 Adding Handles to a Surface to Increase Its Genus

In order to connect these latest ideas with GGB, our next step is to show that our new definition (18.1) of the Euler characteristic does indeed imply (16.2) as a *theorem*:

$$\boxed{\chi(\mathcal{S}_g) = 2 - 2g.} \tag{18.7}$$

We do not believe that this result has an agreed upon name, which permits *us* to call it the ***Euler–L'Huilier formula***, in honour of Simon Antoine Jean L'Huilier, who in 1813 was the first person to state this generalization of Euler's result; see Richeson (2008, Ch. 15) for details.

We have just given two proofs of this theorem in the topologically spherical case, $g = 0$, so now we need to understand the effect of putting *holes* in our surface. Clearly, we should begin by trying to understand the simplest case, namely, the torus/doughnut with one hole, $g = 1$.

[18.8] *A coffee mug is topologically equivalent to a doughnut.* Topology Joke, *by Keenan Crane and Henry Segerman; see Segerman (2016 p. 101). Photograph provided by, and used with permission of, Professor Segerman.*

We shall tackle this problem by means of a *joke*, and it's a bad joke, at that—"probably the first time such a [thing] has ever been used for constructive purposes."[7] The joke goes like this: *A topologist is a person who cannot tell a coffee mug from a doughnut.*

The truth behind the joke is illustrated in [18.8]. Halfway through the transformation of the mug into the doughnut, the body of the mug has coalesced into a blob, still with a handle attached to it. If we now imagine this blob growing to become a sphere, we arrive at the important observation first made by Felix Klein in 1882:[8]

A torus is topologically equivalent to a sphere with a handle attached. More generally, $\mathcal{S}_g$ *is topologically equivalent to a sphere with* g *handles attached.* (18.8)

(See *Further Reading* (at the end of this book) for rigorous statements and proofs of all our plausible but unproven claims regarding the topological classification of surfaces.)

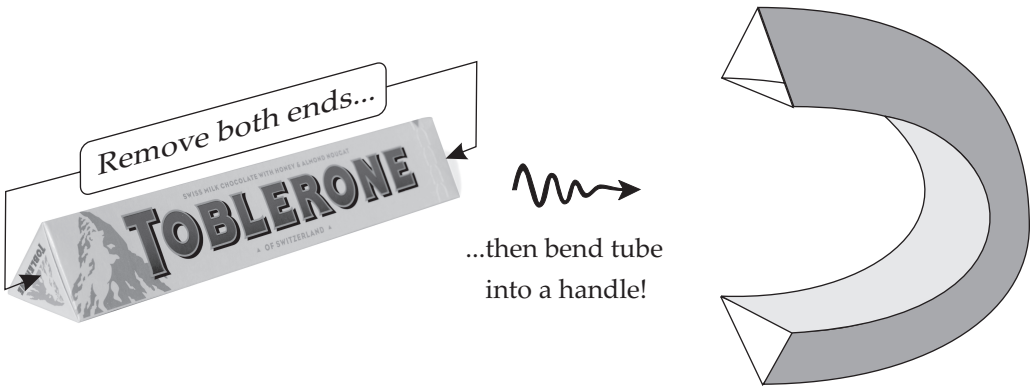
To prove (18.7), it now suffices to prove that

Adding a handle to a closed surface $\mathcal{S}$ *reduces* $\chi(\mathcal{S})$ *by* 2. (18.9)

Granted this, the addition of g handles clearly reduces χ by $2g$. But we have already established that a topological sphere has $\chi = 2$, so adding g handles therefore reduces this to $\chi(\mathcal{S}_g) = 2 - 2g$, thereby proving (18.7).

---

[7] Captain Kirk, addressing Mr. Spock, at the conclusion of *The Doomsday Machine*. The original quote was "such a weapon."
[8] See Stillwell (1995, p. 60).



[18.9] *To create a simple handle, remove the ends from the triangular prism (represented here by a Toblerone® box), then bend the tube to create the handle (with vanishing Euler characteristic).*

To prove (18.9), we will first construct a handle, then glue it to the given surface, $\mathcal{S}$. Figure [18.9] shows the construction of an especially simple handle out of a Toblerone® box, i.e., a triangular right prism. We begin by *removing the two triangular end faces* (so $F \rightsquigarrow F-2$), thereby creating a hollow tube. By Euler's Polyhedral Formula, or [exercise] by simple counting,

$$\chi(\text{Toblerone box}) = 2 \qquad \Longrightarrow \qquad \chi(\text{hollow tube}) = 0.$$

Finally, as illustrated, we imagine the tube to be fashioned out of rubber, and we bend it into the form of a hollow handle. This does not alter its Euler characteristic, so

$$\boxed{\chi(\text{handle}) = 0.}$$

Next, as illustrated in [18.10], we create two holes in our surface $\mathcal{S}$, to which the ends of this handle will be glued. To do so, suppose that we have triangulated[9] $\mathcal{S}$. Next, remove two of these triangular faces, so $F \rightsquigarrow F-2$, and $\chi(\mathcal{S}) \rightsquigarrow \chi(\mathcal{S})-2$.

Figure [18.10] depicts the handle moving towards the now two-holed $\mathcal{S}$, moments before it gets glued to the holes in $\mathcal{S}$. The ends of the handle have six vertices and six edges, and the same is true of the holes. But *after* they are glued together, only six vertices and six edges remain: a net reduction of six vertices and six edges. But this means that *the total Euler characteristic is unaltered by the act of gluing on the handle.* (NOTE: Clearly, this is still true if the ends of the handle are n-gons (with $n > 3$), glued to matching n-gon holes.)

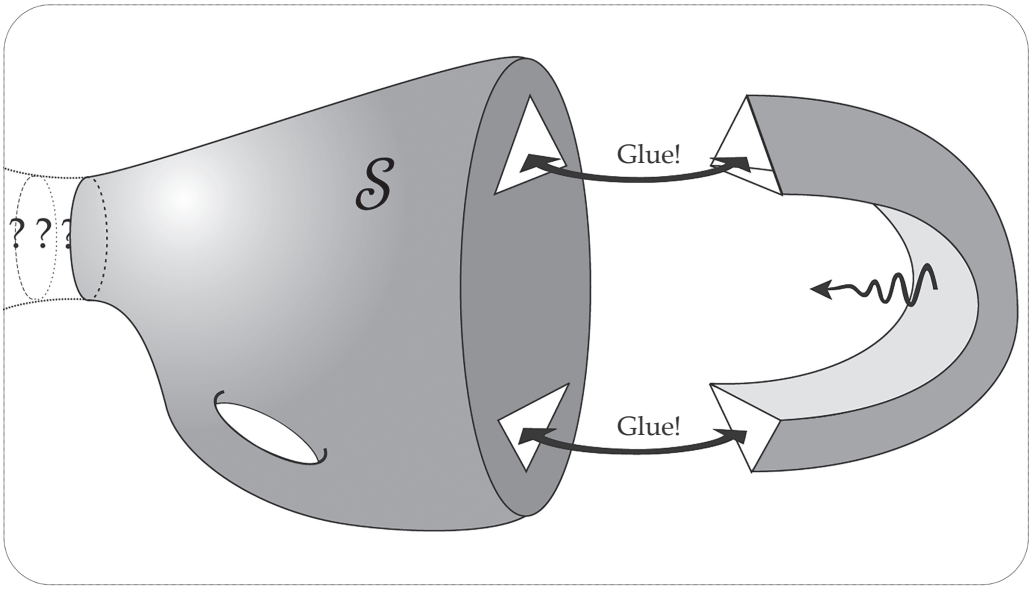
Thus, if $\widetilde{\mathcal{S}}$ is the new surface obtained by gluing on the handle,

$$\begin{aligned} \chi(\widetilde{\mathcal{S}}) &= \chi(\text{two-holed } \mathcal{S} + \text{glued on handle}) \\ &= \chi(\text{two-holed } \mathcal{S}) + \chi(\text{handle prior to gluing}) \\ &= \chi(\mathcal{S}) - 2 + 0, \end{aligned}$$

thereby proving (18.9), and therefore also (18.7).

For an elegant alternative approach, due to Hopf, see Exercise 27.

[9] The triangulation is not shown, but note that the argument does *not* require that we construct the triangulation out of *geodesic* triangles.



[18.10] *The general closed surface $\mathcal{S}$ extends to the left in an unknown manner, indicated by ???. We cut out two triangles from the triangulation (not shown) of $\mathcal{S}$, reducing $\chi(\mathcal{S})$ by 2, then glue on the handle, which has no effect on $\chi(\mathcal{S})$. Thus gluing on a handle reduces $\chi(\mathcal{S})$ by 2.*

18.6 Angular Excess Proof of the Global Gauss–Bonnet Theorem

Let us return to Legendre's proof of Euler's Polyhedral Formula, and let us take one step back from the result upon which it hinged: the generalization of Harriot's result, (18.5). Focus for now only on the right hand side, the "angular excess," given by (18.4):

$$\mathcal{E}(n\text{-gon}) = [\text{angle sum}] - n\pi + 2\pi.$$

Of course the whole point of this expression is to measure the excess of the interior angles of a *geodesic* n-gon over and above the prediction of Euclidean Geometry. If we drop the requirement that the sides of the n-gon are geodesics, then this interpretation ceases to apply. Furthermore, the expression is *always* devoid of *topological* meaning, for it involves angles. Nevertheless, the *expression itself* remains meaningful even if the sides of the n-gon are *not* geodesics.

Despite the untopological nature of this expression, our first crucial step on the road to a new explanation of GGB is the realization that the *sum* of this expression over the entire polygonal net *does* have topological meaning. Indeed, even if the edges are not geodesics, we see that [18.7] and its conclusion, (18.6), *remain valid*:

$$\boxed{\textstyle\sum_j \mathcal{E}(P_j) = 2\pi\,[V - E + F] = 2\pi\,\chi(\mathcal{P}).}$$

For all that was needed in the proof of this result was the fact that the angles in [18.7] that surround a vertex sum to 2π.

By the same token, this part of Legendre's proof does not depend on the topologically spherical polyhedron being inflated into a perfectly round *geometric* sphere. Indeed, and this is the next crucial point, it does not even depend on the surface being *topologically* spherical: *(18.6) remains valid on a surface $\mathcal{S}_g$ of arbitrary genus g.*

As topologically cavalier as we have been (and shall continue to be!) we would be remiss if we did not point out that we cannot cast a *completely* arbitrary polygonal net over $\mathcal{S}_g$. Instead, we must exercise a modicum of caution: *we must not allow our polygons to join onto themselves.* On the torus, for example, we can certainly imagine (but must avoid) a polygon that stretches through the hole and returns to bite its own tail!

With our topological conscience now clear(er), we return to our effort to explain GGB, and to the key equation in Legendre's proof, namely, the generalized Harriot result, (18.5), for an arbitrary geodesic polygon Pj in the net:

$$\frac{1}{R^2}\,\mathcal{A}(P_j) = \mathcal{E}(P_j).$$

While the sum of the right-hand side has just been seen to be to be topological in nature—and always equal to $2\pi\chi(\mathcal{S}_g)$—equality with the left-hand side is *only* valid if we (i) inflate the polyhedron into a perfect sphere of radius R and (ii) ensure that all the edges of the polygonal net are *geodesic* arcs of great circles on this sphere.

Although Harriot could not have recognized it in 1603, two centuries later Gauss taught us that the sphere has constant curvature $\mathcal{K} = (1/R^2)$, and that the left-hand side of the above equation should in fact be viewed as the *total curvature* $\mathcal{K}(P_j)$ residing within P_j:

$$\mathcal{K}(P_j) = \mathcal{K}\,\mathcal{A}(P_j) = \frac{1}{R^2}\,\mathcal{A}(P_j).$$

Gauss's 1827 form of the Local Gauss–Bonnet Theorem, ((2.6), p. 23), was a tremendous generalization of this result: the angular excess of a geodesic triangle Δ on a *general* curved surface is likewise given by the total curvature within. And just as Harriot's original result generalizes to geodesic polygons, so too does Gauss's result, and in exactly the same manner:

$$\mathcal{E}(\Delta) = \iint_\Delta \mathcal{K}\,\mathrm{d}\mathcal{A} \qquad \Longrightarrow \qquad \mathcal{E}(P_j) = \iint_{P_j} \mathcal{K}\,\mathrm{d}\mathcal{A} = \mathcal{K}(P_j).$$

Finally, summing over all the geodesic polygons, and using (18.6) and (18.7), we have arrived at our second proof of GGB:

$$\mathcal{K}(\mathcal{S}_g) = \textstyle\sum_j \mathcal{K}(P_j) = \sum_j \mathcal{E}(P_j) = 2\pi\chi(S_g) = 2\pi(2-2g).$$

Chapter 19

Third (Vector Field) Proof of the Global Gauss–Bonnet Theorem

19.1 Introduction

We end Act III (the traditional "climax" of our drama) with a beautiful link between geometry, topology, and *vector fields*—a climax within a climax! Here we shall sketch only those ideas that are needed to achieve a new understanding of GGB; for a much fuller account of vector fields, linked to Complex Analysis and to Physics, see chapters 10, 11, and 12 of VCA. See also *Further Reading*, at the end of this book.

19.2 Vector Fields in the Plane

Imagine a thin layer of fluid flowing over a horizontal plane. It will be to our advantage to think of this plane as the *complex* plane, $\mathbb{C}$. At each point z of $\mathbb{C}$ we therefore have a velocity vector, or rather, a complex number, $V(z)$, which we draw emanating from z. This flow $V(z)$ is called a ***vector field*** on $\mathbb{C}$.

We shall suppose that our vector field $V(z)$ is extremely well behaved, being continuous and differentiable[1] at all but a finite number of isolated points. Thus, at a normal or *regular* point, a small movement in any direction results in a correspondingly small, ultimately proportional, change in the direction and length of V.

In contrast to this, a ***singular point*** s is an exceptional place where the vector field suffers a *discontinuity*: infinitesimal movements away from s in different directions result in V pointing in completely different directions or having completely different lengths.

Figure [19.1] illustrates a typical vector field, in which the singular points a, b and c are instantly identifiable to the untrained eye, marked with black dots, for clarity.

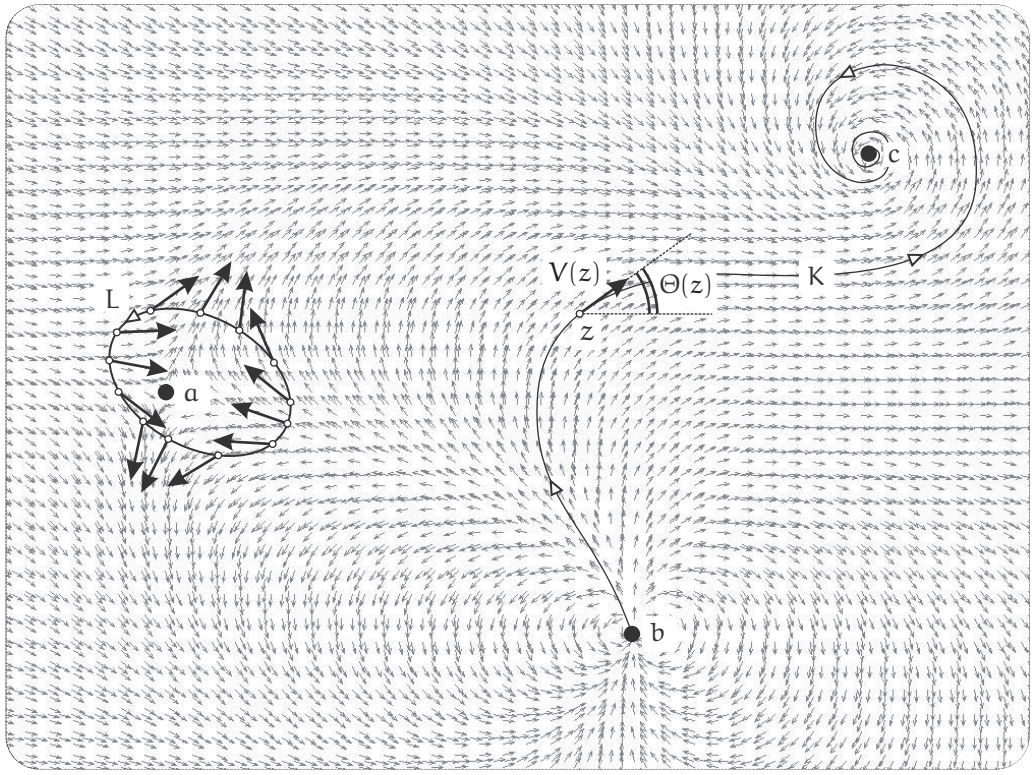
If we follow the path of an individual particle of fluid, we obtain a ***streamline*** K (or ***integral curve***) of the flow, such that V is always tangent to K. If we trace *all* such streamlines, as we shall do in examples below, then we obtain a vivid depiction of the vector field as a whole, called the ***phase portrait***. Note that the phase portrait completely encapsulates the information about the direction $\Theta(z) = \arg[V(z)]$ of $V(z)$ at every point.

In [19.1] the vectors all have equal length, but this is certainly not true in general, as we shall illustrate in the examples that follow. This is a serious deficiency of the phase portrait, for it fails[2] to illustrate the *magnitude*, $|V(z)|$, and this magnitude represents vital information: the speed of a fluid flow, or the strength of a magnetic field, for example. Nevertheless, from the point of view of *topology*, the phase portrait is the *only* thing that matters.

---

[1] In the weaker real sense, *not* the complex analytic sense of being an amplitwist.

[2] In the case of the most physically important vector fields, there *is* in fact a special way to draw the streamlines so that the strength of the flow becomes visible, as the *crowding together* of the streamlines. See VCA, pages 494–502.



[19.1] *A typical vector field,* $V(z)$. *Clearly visible are three singular points:* a *is called a* **saddle point** *(or* **crosspoint***);* b *is called a* **dipole***; and* c *is called a* **vortex** *(or* **focus***). Also shown is a* **streamline**, K. *The* **index** $\mathcal{I}_V(a)$ *is the number of revolutions executed by* $V(z)$ *as* z *travels counterclockwise once around* a*, along a loop such as* L. *Since* V *executes one clockwise (i.e., negative) revolution as we loop around* a*, we see that* $\mathcal{I}_V(a) = -1$.

19.3 The Index of a Singular Point

Just as the locations of the singular points are obvious, so too are their dramatically different *characters*. If we imagine the flow pattern [19.1] drawn on a rubber sheet, which is then stretched this way and that, we intuit that the character of a is quite different from that of b or c, and that this distinction is invariant under the stretching. In other words, it seems clear that the character of a singular point s is not so much a geometric feature as it is a *topological* feature of the flow.

Indeed, these different characters lead to different *names* for the various types of singular points. For example, in [19.1], a is called a ***saddle point*** (or ***crosspoint***), for this is the way water flows when it is poured over a horse's saddle; b is called a ***dipole***, for this is the way the magnetic field lines stream between the two poles of a short bar magnet; and c is called a ***vortex*** (or ***focus***), for this is the way water swirls around and down the drain of a kitchen sink.

We now explain how we may crystalize this vague concept of the "character" of a singular point s into a single, topologically invariant *integer* $\mathcal{I}_V(s)$, called the ***index*** of s. (NOTE: Absent any ambiguity regarding the vector field V, we may drop it and abbreviate the notation to $\mathcal{I}(s)$.)

Let us immediately state the definition, although it may not be immediately evident that it is even well defined:

If s *is a singular point of a vector field* $V(z)$*, and* L *is any simple loop containing* s *(and no other singular points) then the* **index** $\mathcal{I}_V(s)$ *of* s *is* the net number of revolutions *that* $V(z)$ *executes as* z *travels once round* L*, counterclockwise.* (19.1)

Let us immediately illustrate this with the singular points in [19.1]. Since the index only cares about the direction of the vectors on L, we have taken the liberty of enlarging these specific vectors for clarity. As we traverse L, it is clear that $V(z)$ undergoes one clockwise (i.e., negative) revolution, so

$$\mathcal{I}(\text{saddle point}) = -1.$$

Try your own hand at this by examining the singular points at b and c, and verify that

$$\mathcal{I}(\text{dipole}) = +2 \qquad \text{and} \qquad \mathcal{I}(\text{vortex}) = +1.$$

Next, let us make our definition (19.1) a bit more precise. Let $\Theta(z)$ be the angle that $V(z)$ makes with the horizontal. Of course there are infinitely many choices for this angle, differing by multiples of 2π, but suppose we choose one at the point z. Provided we insist that $\Theta(z)$ vary *continuously*, then $\Theta(z)$ is now *uniquely determined* as z moves along a directed curve J, say, that does not pass through any singular points, for then $V(z)$ varies continuously on J.

We can now define $\delta_J\Theta$ to be the net change in the angle $\Theta(z)$ as z travels along J, from Start to Finish:

$$\delta_J\Theta \equiv \Theta(\text{Finish}) - \Theta(\text{Start})$$

Note that this is the *signed* change in the angle, so if we reverse the direction of J (denoted $-J$), thereby swapping Start and Finish, the sign of $\delta_J\Theta$ is also reversed:

$$\delta_{-J}\,\Theta = -\delta_J\,\Theta. \tag{19.2}$$

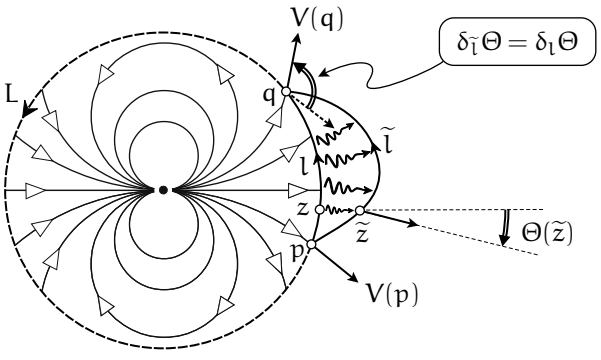
With this understanding and notation in place, we can restate our definition (19.1) of the index as follows:

$$\mathcal{I}_V(s) = \frac{1}{2\pi}\,\delta_L\Theta. \tag{19.3}$$

By now you probably already intuit that this definition is indeed well defined, but to *prove* that it is, consider [19.2], which illustrates the general argument with an (initially) circular loop L encircling a dipole.

Focus attention on the illustrated segment l of L, running from $p =$ Start to $q =$ Finish. As illustrated,

$$\delta_l\Theta = \Theta(q) - \Theta(p).$$



[19.2] *The circular loop* L *encircles a dipole. If the segment* l *of* L *is deformed into* $\tilde{l}$*, carrying* z *to* $\tilde{z}$*, the angle* $\Theta(z)$ *varies continuously, so* $\Theta(\tilde{z})$ *will be close to its original value,* $\Theta(z)$*. It follows that the change in* Θ *along this segment is* invariant *under the deformation:* $\delta_{\tilde{l}}\,\Theta = \delta_l\Theta$.

Suppose we now continuously deform l a small amount into neighbouring $\widetilde{l}$, without l crossing any singular points in the process.

As a point z on l evolves into the new point $\widetilde{z}$ of $\widetilde{l}$, the angle $\Theta(z)$ will also vary continuously, so $\Theta(\widetilde{z})$ will be close to its original value, $\Theta(z)$. It follows that as $\widetilde{z}$ travels along $\widetilde{l}$, the change $\delta_{\widetilde{l}}\Theta$ must also be close to the change $\delta_l\Theta$ as z travels along l.

Now comes the archetypal topological argument. If $\delta_{\widetilde{l}}\Theta$ were not identically equal to $\delta_l\Theta$, it would have to differ from it by a multiple of 2π, but this is impossible, because $\delta_{\widetilde{l}}\Theta$ is close to $\delta_l\Theta$. Thus,

$$\delta_{\widetilde{l}}\,\Theta = \delta_l\Theta.$$

We may similarly deform any other segment of L, or all of it at once, so we immediately deduce that the index is indeed independent of the size and shape of L:

> *If* s *is a singular point of a vector field* $V(z)$*, and* L *is any simple loop containing* s *(and no other singular points) then the* **index** $\mathcal{I}_V(s)$ *does not change as* L *continuously deforms into any other such loop, so long as it does not cross any singular points in the process. In short,* $\mathcal{I}_V(s)$ *is independent of* L*, and is a property of* V *that can be attached to* s *itself.* (19.4)

If we instead view the vector field as a complex mapping $z \mapsto V(z)$ from one complex plane to another, taking the point z in the first plane to the *point* $V(z)$ in the second, then [exercise] $\mathcal{I}_V(s)$ may instead be viewed as the number of times that the image loop $V(L)$ winds around 0. This is called the ***winding number*** of $V(L)$, and it plays a fundamental role in Complex Analysis; see VCA, Chapter 7.

19.4 The Archetypal Singular Points: Complex Powers

Why did we take our plane to be the *complex* plane? There are several reasons, but here is an important one: the archetypal singular points arise naturally from the *powers* of the complex variable z:

$$P_m(z) \equiv z^m = [r\,e^{i\theta}]^m = r^m\,e^{im\theta},$$

where m is assumed to be an integer.

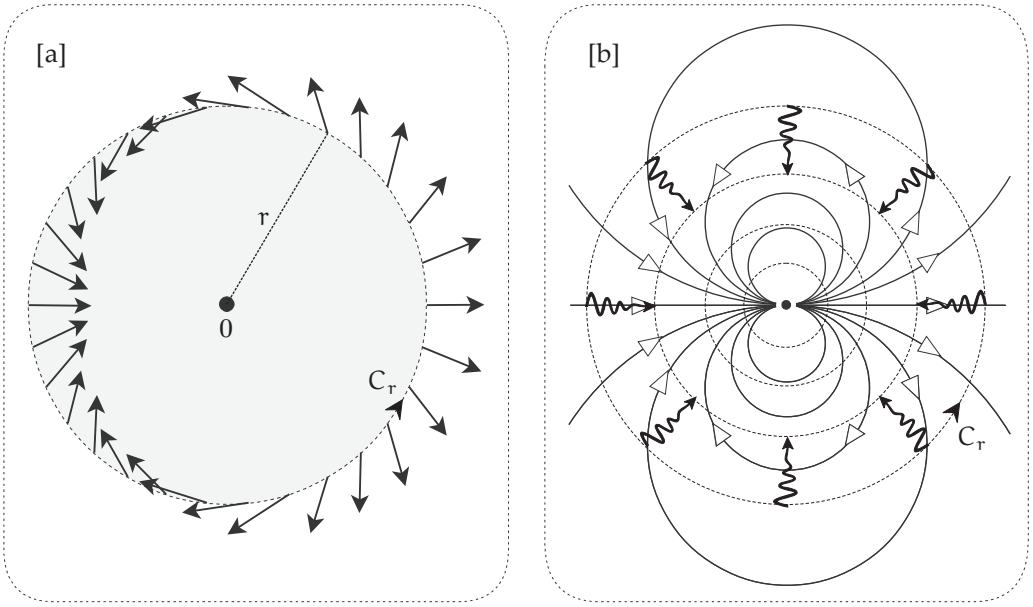
The only (finite)[3] singular point of $P_m(z)$ is at the origin, and it is easy to determine its index. Everywhere along the ray in the direction θ, we see that P_m points in the direction $m\theta$; only the length $|P_m(z)| = r^m$ varies as we move along the ray. Thus, if we traverse any loop that goes once round the origin counterclockwise, so that θ goes from 0 to 2π, then the angle of P_m goes from 0 to $2\pi m$. In other words, P_m executes m (positive or negative) revolutions, so $\mathcal{I}(0) = m$.

For example, reconsider the "dipole" we encountered in [19.1] at b. Figure [19.3] illustrates that this dipole field is none other than $P_2(z) = z^2$: [19.3a] shows the vector field itself on a circular loop C_r of radius r, and [19.3b] shows the phase portrait of streamlines. Make sure you can see that both these diagrams are at least qualitatively correct. (For bonus points, prove that the streamlines in [19.3b] are indeed perfect circles!)

Although the definition (19.3) of the index makes essential use of the loop L enclosing the singular point, we have already seen in (19.4) that the shape and size of L is actually a red herring: the index is actually a property of the vector field *at* the singular point itself.

To make this idea more vivid, [19.3b] illustrated the loop C_r shrinking towards the singular point. Clearly, *it is the behaviour of the vector field in an infinitesimal neighbourhood of the singular point*

[3] There is actually a second singular point at infinity; see VCA, page 492.



[19.3] *As we traverse the circle in* **[a]** *once (positively) round the singular point at* 0, $P_2(z) = z^2$ *executes two positive revolutions, so* $\mathcal{I}(0) = +2$. *The same result can be obtained by inspecting the phase portrait (i.e., streamlines) of the flow in* **[b]**. *As we shrink* C_r *down towards the singular point, the index does not change: the index characterizes the vector field in an infinitesimal neighbourhood of the singular point.*

that determines the index of the point:

$$\mathcal{I}_V(s) = \frac{1}{2\pi} \lim_{r\to 0} \delta_{C_r} \Theta.$$

To sum up, if m is any integer,

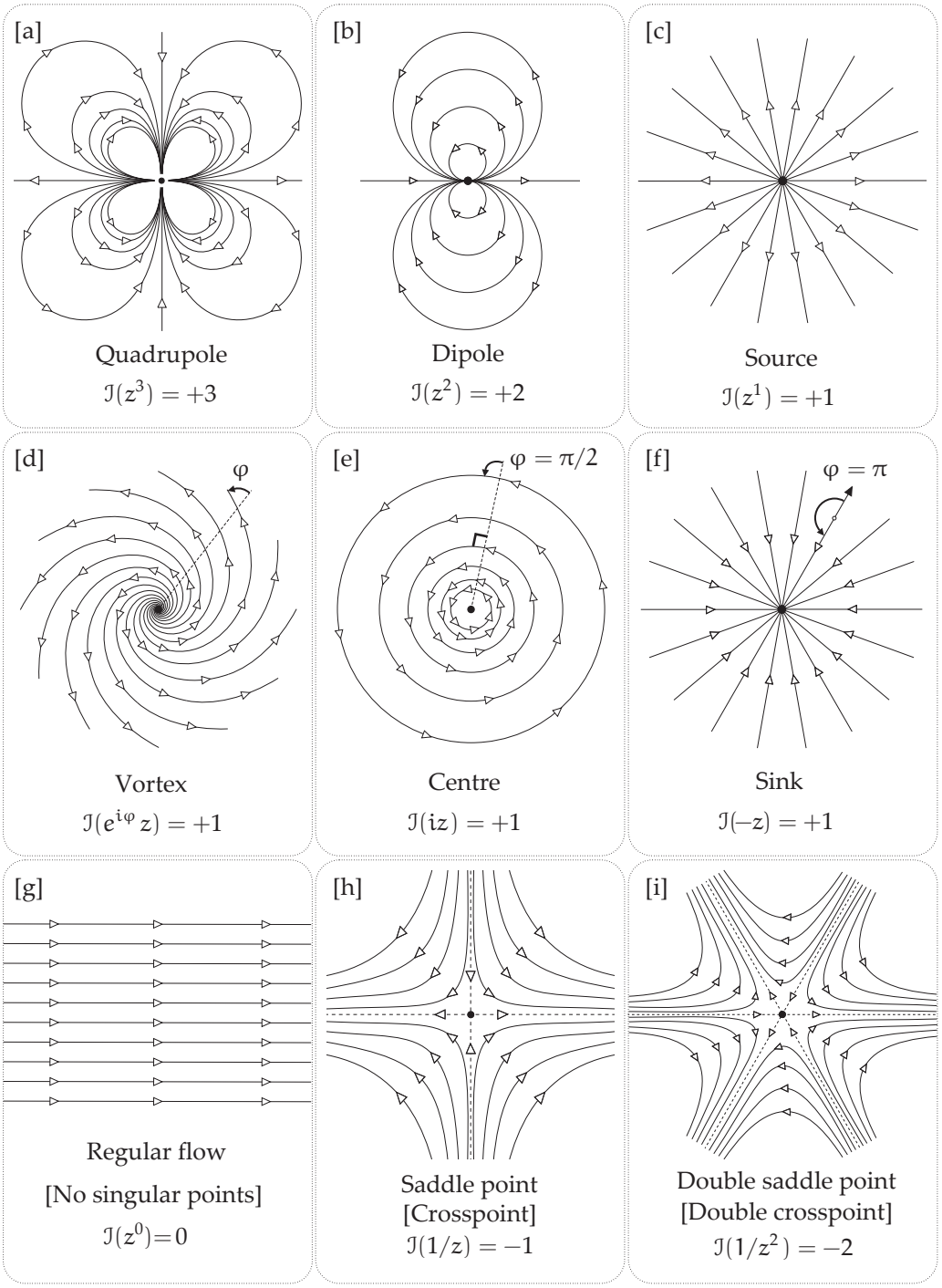
$$P_m(z) = z^m \qquad \Longrightarrow \qquad \mathcal{I}(0) = m.$$

Figure [19.4] shows these fields for $m = +3, +2, +1, 0, -1, -2$. Please take a moment to visually verify, for each field, that (i) it is indeed what it purports to be and (ii) that the index is indeed equal to the power m.

While we started our discussion of singular points with the intuitive idea of their "character," the more precise concept of index has moved us in a new direction. For example, a source, a vortex, and a sink are, from the physical point of view, all quite different in character, but they are nevertheless *indistinguishable* from the topological point of view, for all three have index $+1$.

This is explained by the following fact: multiplying a complex function $f(z)$ by a complex *constant* $k = Re^{i\varphi}$ has no effect on its indices. For $f(z) \rightsquigarrow Re^{i\varphi} f(z)$ stretches the vectors by R and *rotates them by the fixed angle* φ. Thus $kf(z)$ executes the same number of revolutions as $f(z)$ as z traverses any closed loop, and therefore the index remains the same. In particular, if we let $f(z) = z$ and let φ gradually increase from 0 to $(\pi/2)$, and then from $(\pi/2)$ to π, the field of the source in [19.4c] evolves like this: [c] $\rightsquigarrow$ [d] $\rightsquigarrow$ [e] $\rightsquigarrow$ [f].

We end this section with some general observations. First, note that *reversing the direction of any flow has no effect on the location or indices of its singular points.* As we traverse a loop around a singular point, V and $-V$ undergo exactly equal rotations, like the two ends of a compass needle, so the common singular point of the two vector fields has the same index for both.



[19.4] *The vector fields in* $\mathbb{C}$ *of the integer powers of the complex variable* z, $P_m(z) = z^m$, *for* $m = +3, +2, +1, 0, -1, -2$. *The index* $\mathcal{I}$ *of the singular point at the origin is equal to* m. *The middle row shows the vector fields of* $e^{i\varphi} z$, *all with* $\mathcal{I} = +1$, *for* **[d]** *a* **Vortex** *with a general value of* φ; **[e]** *the* **centre**, $\varphi = \pi/2$; **[f]** *the* **sink**, $\varphi = \pi$.

ALTERNATIVE NAMES: *sink* = **stable node**; *source* = **unstable node**; *vortex* = **(stable or unstable) focus—[d] is unstable**; *saddle point* = **crosspoint**.

Next, note that if we define the complex conjugate vector field by

$$\overline{P}_n(z) = \overline{z^n} = \overline{z}^n,$$

then the streamlines of $P_m(z)$ are identical to those of $\overline{P}_{-m}(z)$, so their common singular point at the origin is a zero for both fields and has the same index in both fields.

More generally, for any complex function $f(z)$, the vector field $\overline{f(z)}$ has become known as the ***Pólya vector field of*** $f(z)$, in honour of George Pólya (1887–1985). Clearly the Pólya vector field has the same singular points as $f(z)$, but since $\overline{f(z)}$ has the opposite angle, these common singular points have opposite indices. The principal virtue of the Pólya vector field is that it allows for simple, intuitive, *visual and physical* interpretations of the complex contour integral of $f(z)$. For details, see Chapter 11 of VCA.

Finally, observe that if we take the reciprocal of $f(z)$ then we obtain a new vector field $1/f(z)$ that goes to infinity (the north pole of the Riemann sphere) at a singular point where $f(z) = 0$, and the index of $1/f(z)$ at such a singular point is [exercise] the *negative* of that of $f(z)$: this is particularly clear in the case of $P_m(z)$, for then $1/P_m(z)$ points in the same direction as $P_{-m}(z)$. If we now consider the Pólya vector field of this reciprocal, namely $1/\overline{f(z)}$, the indices are reversed a second time, and are therefore the *same* as those of $f(z)$.

NOTES ON THE PHYSICAL TERMINOLOGY. We have borrowed the terms, ***source***, ***vortex***, ***sink***, ***dipole***, and ***quadrupole***, from physics, to describe the streamlines of z, z^2, and z^3. While this is topologically correct, it is *not* physically correct: the physical fields corresponding to these terms are actually the Pólya vector fields $\overline{(1/z)}$, $\overline{(1/z^2)}$, and $\overline{(1/z^3)}$.

As explained in the previous paragraph, each of these fields points in the same direction as the corresponding positive power, so they have the same singular points and indices, thereby justifying our use of these terms in the current topological setting. However, the *magnitudes* of the physical fields are the *reciprocals* of those of the positive powers to which we have attached the same names.

In an attempt to clear our conscience, let us explain this in the simplest case of a *source*. Suppose water is supplied at a constant rate s through a very narrow tube to a point on the surface of a horizontal plane, which we shall take to be $\mathbb{C}$. The water will flow radially outward, symmetrically, from this *source* (now used in the *physical* sense), which we suppose to be at the origin.

While the radial velocity field $P_1(z) = z = re^{i\theta}$ depicted in [19.4c] has magnitude r, *speeding up* as we move away from the origin, it is intuitively clear that the velocity v of the radial flow from a physical source must *slow down* the further away the water travels. More precisely, the total amount of water flowing out across a circle of radius r (per unit time) must equal the amount being supplied at the origin, namely, s. Thus,

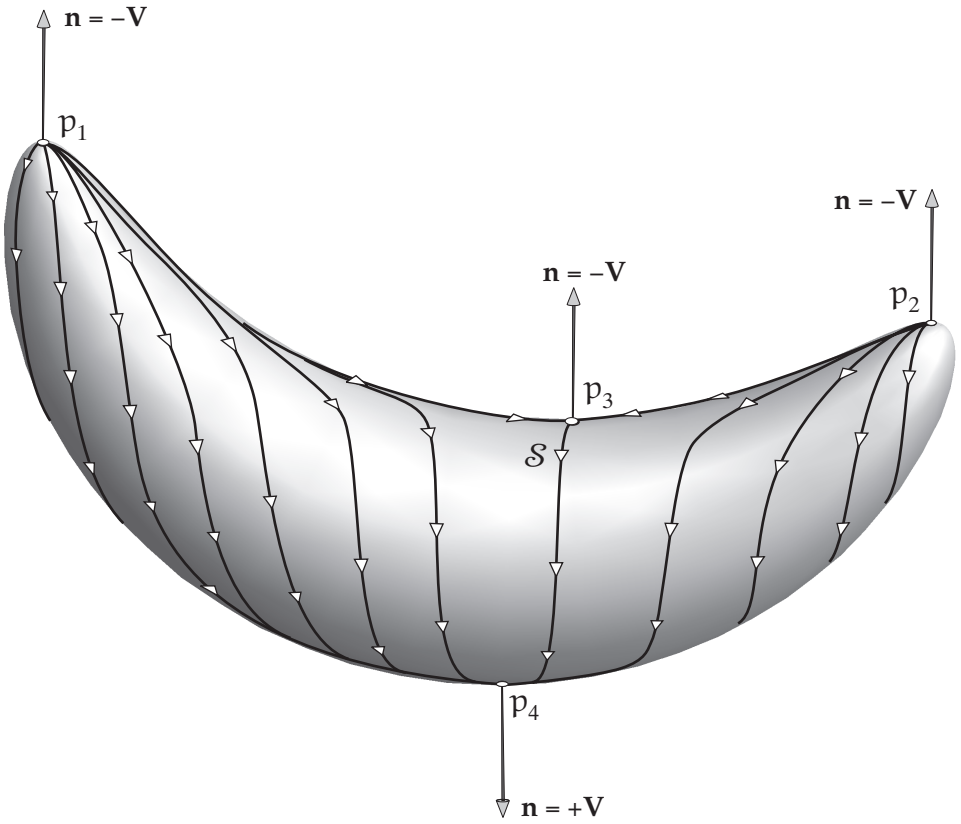
$$2\pi r\, v = s \quad \Rightarrow \quad v = \frac{s}{2\pi r} \quad \Rightarrow \quad \text{physical source} = v\, e^{i\theta} = \frac{s}{2\pi}\left[\frac{e^{i\theta}}{r}\right] = \frac{s}{2\pi}\overline{\left[\frac{1}{z}\right]}.$$

For much more on the physical interpretations of such fields, and their symbiotic relationship with Complex Analysis, see VCA, Chapters 10, 11, and 12.

19.5 Vector Fields on Surfaces

19.5.1 The Honey-Flow Vector Field

It is easy enough to imagine a vector field on a surface, instead of in the plane. Picture any smooth object left out in a rainstorm: rain water flows over the surface $\mathcal{S}$ of the object, down towards the



[19.5] The Honey-Flow. *Honey rains down in the gravitational direction* $\mathbf{V}$ *onto the surface* $\mathcal{S}$ *of a fried banana. The honey then flows down over its surface, creating the "honey-flow" velocity vector field* $\mathbf{v}$ *on* $\mathcal{S}$. *The singular points of* $\mathbf{v}$ *occur when the outward normal* $\mathbf{n}$ *of* $\mathcal{S}$ *is* $\mathbf{n} = -\mathbf{V}$ *(at* $p_{1,2,3}$*), or* $\mathbf{n} = +\mathbf{V}$ *(at* p_4*). Visual intuition suggests that* p_1 *and* p_2 *are sources, with* $\mathcal{I}(p_1) = \mathcal{I}(p_2) = +1$, p_3 *is a saddle point, with* $\mathcal{I}(p_3) = -1$, *and* p_4 *is a sink,* $\mathcal{I}(p_4) = +1$. *These intuitions are confirmed by the precise definition, (19.5).*

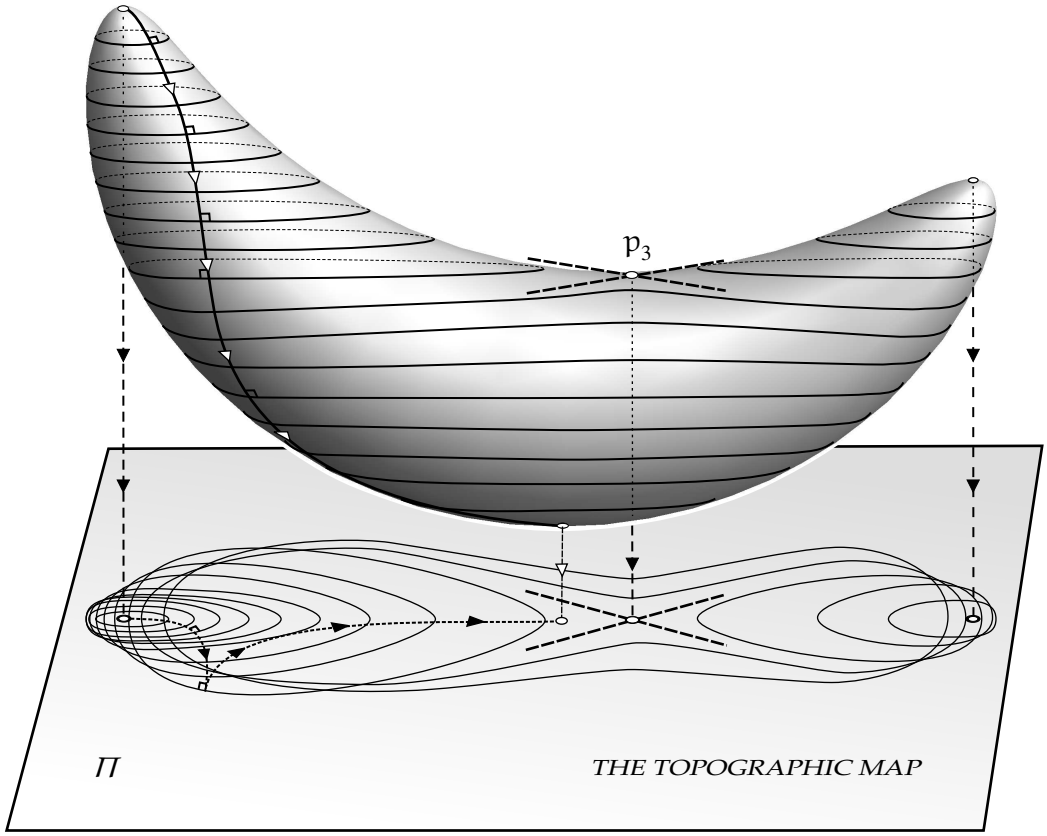
ground, and its velocity $\mathbf{v}(p)$ at each point $p \in \mathcal{S}$ is then an example of a ***vector field on the surface***, the vectors themselves being everywhere tangent to the surface.

While this rain flow $\mathbf{v}(p)$ is but one example of a vector field on $\mathcal{S}$, we shall see shortly that this *specific* one is a powerful theoretical tool, and the key to a new proof of GGB.

Figure [19.5] illustrates this flow on a particular surface, the direction of the gravitation pull being $\mathbf{V}$. In reality, the rain water flowing down over this surface might well be pulled *off* the surface, falling straight down to the ground as soon as it can, i.e., once the tangent plane becomes vertical, containing $\mathbf{V}$.

In order to allow physical intuition to continue to inform mathematical intuition, we must ensure that the fluid *adheres* to $\mathcal{S}$ and always flows over it. We therefore perform a biblical miracle: we turn water into *honey!* (By way of an ancillary miracle, let us also turn [19.5] into a fried banana!)

Accordingly, the flow $\mathbf{v}$ (resulting from this honey being dragged over the surface by a force $\mathbf{V}$ in space) will henceforth be referred to as the ***honey-flow vector field*** associated with $\mathbf{V}$. (WARNINGS: You may be shocked to learn that "honey-flow" is *not* (yet) standard mathematical terminology! Also, we shall only pay poetic homage to physics, faithfully keeping track of the *direction* of the honey-flow, but *not* its speed.)



[19.6] The Topographic Map *of a surface is obtained by intersecting it with equally spaced horizontal planes, then projecting the resulting* **level curves** *vertically downward onto the map plane,* Π. *(We have not mapped the bottom half of the surface, to avoid clutter.) The level curve through the saddle point* p_3 *is a figure eight, of which only its tangents at* p_3 *are shown (dashed lines). The streamlines of the honey-flow on the surface are orthogonal to the level curves, and they* remain *orthogonal after projection down to* Π *(despite the fact that the map is* not *conformal).*

19.5.2 Relation of the Honey-Flow to the Topographic Map

We will give a concise, purely mathematical definition of the honey-flow **v** in Section 19.7, but, for now, let us continue to think physically, in order to gain a fresh insight. To that end, recall that we can represent the shape of the fried banana in [19.5], or any other surface, using a flat ***topographic map***, of the type used in geography.

To construct this map, we take many equally spaced horizontal planes and look at the curves on the surface where these planes intersect it, the so-called ***level curves***. Finally, we project these level curves vertically downward onto the horizontal map plane Π below, thereby creating the topographic map.

This process is illustrated in [19.6], but only for the *upper* part of the fried banana; the map of the bottom half would clutter the diagram, and is therefore best drawn in a separate figure (not shown). If the level curves are imagined to exist on a transparent surface, the topographic map is what you would see if you were to look straight down at the surface from high above. In the case of the banana in [19.6], you can experimentally verify the map by taking a long knife and slicing up the banana thinly, then examining the shape of each slice.

Note that when the intersecting plane passes through the saddle point p_3, the level curve is a figure eight (not shown); the figure does show the tangents (dashed lines) to this figure eight

as it passes through p_3. Recall that these are the asymptotic directions of the surface, $\kappa = 0$, and that the pattern of hyperbolas in the vicinity of the saddle point is the Dupin Indicatrix there. See Section 15.8, page 162.

Note also that where the level curves are closest together, the surface of the banana is steepest, and where they are furthest apart, the surface is shallowest. Moving along a level curve corresponds to moving sideways on the surface, at constant height. Moving at *right angles* to the level curves therefore corresponds to travelling in the direction of steepest descent (or ascent) on the surface. Points where (infinitesimally separated) neighbouring curves in the map *intersect* correspond to different heights above a single point, i.e., to a vertical tangent plane.

Since the gravitational force pulling the honey down the surface has no horizontal component (along the level curves), we immediately deduce that the honey-flow on the surface is along the *orthogonal trajectories* of the level curves.

This has implications for the topographic map. Since $\mathbf{v}$ is orthogonal to the tangent $\mathcal{T}$ to the level curve, it must lie within the plane that is orthogonal to $\mathcal{T}$. But since $\mathcal{T}$ is horizontal (by construction) this orthogonal plane is vertical. Since the vertical planes through $\mathcal{T}$ and $\mathbf{v}$ are orthogonal, they intersect the horizontal map plane Π in orthogonal directions. So, as illustrated in [19.6],

> *On both the surface* and *in its topographic map, the streamlines of the honey-flow are the* orthogonal trajectories *of the level curves and of the topographic map.*

Note that the vertical projection from the surface down to the map is *not* conformal: in general, angles on the surface are *not* faithfully represented in the topographic map. However, despite this nonconformality, we have just proved that the right angles between the level curves and the streamlines of the honey-flow *are* preserved.

If two plane vector fields are orthogonal, then they must rotate the same amount as we traverse a curve. A shared singular point of the two fields must therefore have the same index. For example, in the topographic map, the level curves near either peak of the fried banana look like the streamlines of a centre vector field, with $\mathcal{I} = +1$. Since the streamlines of the honey-flow are orthogonal to the level curves, it follows that the (projected) honey-flow must have $\mathcal{I} = +1$, too, as it does.

Likewise, near the saddle point of the fried banana, the level curves look like the streamlines of a saddle point, with $\mathcal{I} = -1$. It follows that the honey-flow also has a saddle point with $\mathcal{I} = -1$; recall that this is *why* it is called a saddle point!

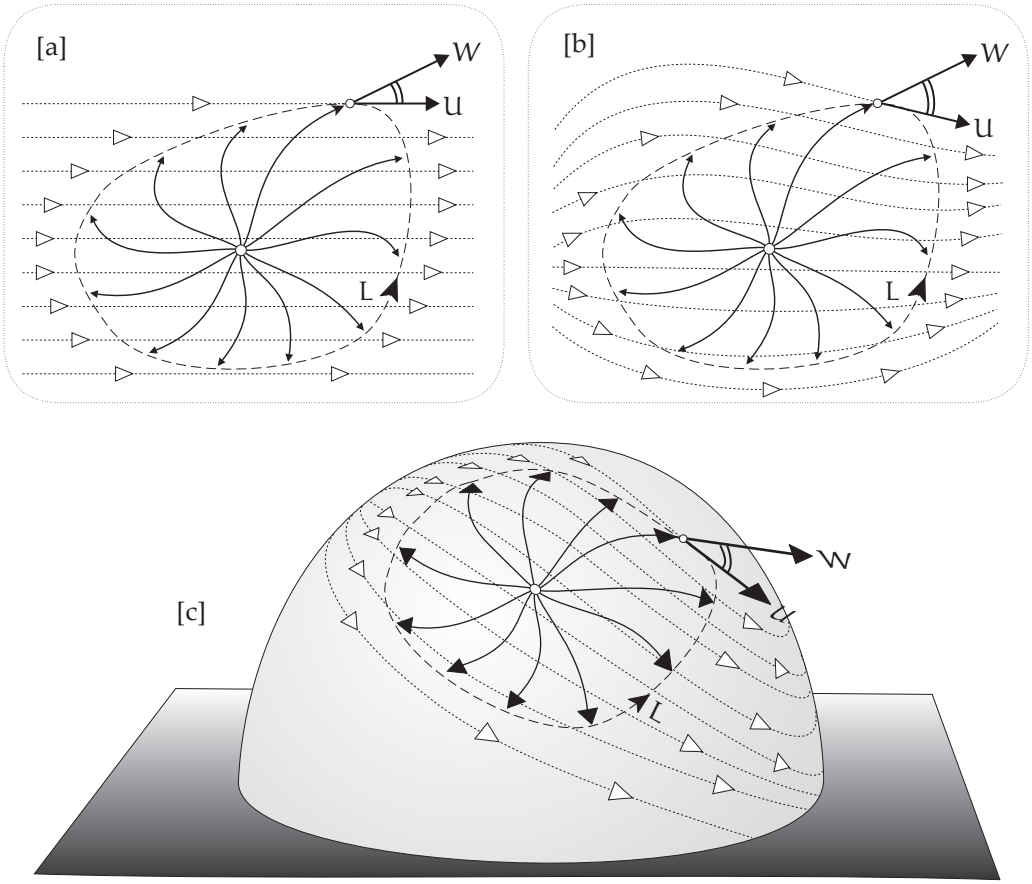
In the case of the honey-flow, we thus have a simple way of ascribing an "index" to each of its singular points, via the surface's topographic map. We now turn to a more general analysis of the concept of the index, directly on the surface itself.

19.5.3 *Defining the Index on a Surface*

Given *any* vector field on a surface, not necessarily the honey-flow, it seems intuitively clear that it must be possible to extend our definition of the index to its singular points.

For example, in [19.5] we immediately identify four singular points of the honey-flow: p_1, p_2, p_3, and p_4. These are the points where the direction of the outward surface normal either coincides with $-\mathbf{V}$, as at $p_{1,2,3}$, or else coincides with $+\mathbf{V}$, as at p_4. Clearly, the "correct" generalization of $\mathcal{I}$ should be such that p_1 and p_2 are sources, with $\mathcal{I}(p_1) = \mathcal{I}(p_2) = +1$, p_3 is a saddle point, with $\mathcal{I}(p_3) = -1$, and p_4 is a sink, with $\mathcal{I}(p_4) = +1$.

To give precise meaning to this new, generalized notion of "index," presumably we should draw a loop round the singular point on the surface, then find the net rotation of the vector field as the loop is traversed. But wait, *rotation relative to what?*



[19.7] Defining the Index on a Surface. *We usually measure the rotation of a plane vector field* W *with respect to a horizontal fiducial (or* **reference***) field* U*, as in* **[a]**. *But we may instead measure the rotation with respect to* any *other field without singular points, such as that in* **[b]**. *Finally, in* **[c]**, *we generalize to surfaces: draw a regular flow* U *across the region containing the singular point; then the index is the count of the revolutions of* W *relative to* U.

To answer this question, we first re-examine the familiar concept of rotation in the plane. Figure [19.7a] illustrates that the rotation of a plane vector field $W(z)$ along a loop L can be thought of as taking place relative to a ***fiducial*** (or ***reference***) vector field U having horizontal streamlines, say $U(z) = 1$. If we define $\angle UW$ to be the angle from U to W, and let $\delta_L(\angle UW)$ be the net change in this angle along L, then our old definition (19.3) of the index can be written as

$$\boxed{\mathcal{I}_W(s) = \tfrac{1}{2\pi}\,\delta_L(\angle UW).} \tag{19.5}$$

If we continuously deform the straight horizontal streamlines of U in [19.7a] to produce the curved ones in [19.7b], then, by the usual reasoning, the right-hand side of (19.5) will not change. Thus we conclude that this formula yields the correct value of the index if we replace U with *any* vector field that is nonsingular on and inside L.

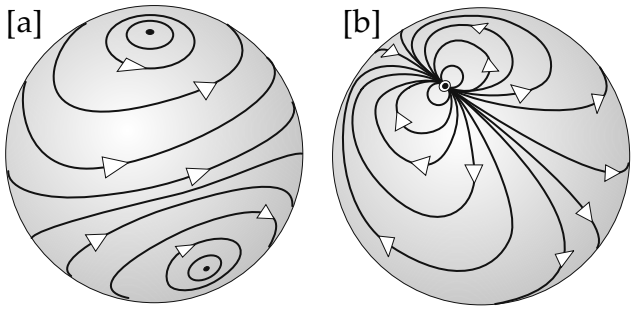
Next, imagine that [19.7b] is drawn on a rubber sheet. If we continuously stretch it into the form of the curved surface in [19.7c], then not only will the right-hand side of (19.5) remain well defined, but its value will not change.

To summarize, if s is a singular point of a vector field W on a surface $\mathcal{S}$, we define its ***index*** as follows. Draw any nonsingular vector field U on a patch of $\mathcal{S}$ that covers s but no other singular points; on this patch, draw a simple loop L going round s; finally, apply (19.5)—that is, count the net revolutions of W relative to U as we traverse L.

19.6 The Poincaré–Hopf Theorem

19.6.1 Example: The Topological Sphere

Figure [19.8] shows the streamlines of two possible flows on the sphere. Notice that both possess singular points: [19.8a] has two centres, while [19.8b] has a dipole. In fact there can be *no* vector field on the sphere that is free of singular points.



[19.8] [a] *Two centres with index sum* 2. **[b]** *A dipole with index* 2.

This is but one consequence of an extremely beautiful and important result, called the ***Poincaré–Hopf Theorem***, which we will state momentarily.

For now, to get a first whiff of the result, note that if we sum all the indices in [19.8a] we obtain

$$\mathcal{I}(\text{centre}) + \mathcal{I}(\text{centre}) = 1 + 1 = 2,$$

while if we do the same for [19.8b] we obtain

$$\mathcal{I}(\text{dipole}) = 2.$$

Next, reconsider the four singular points of the honey-flow depicted in [19.5]. Here the sum of the indices is $(1) + (1) + (-1) + (1) = 2$, *again!*

Try drawing your own streamlines on an orange. For example, consider flow from the north pole along the meridians to the south pole. Summing the indices of the singular points at the poles,

$$\mathcal{I}(\text{source}) + \mathcal{I}(\text{sink}) = 1 + 1 = 2,$$

again! Perhaps this is all some bizarre coincidence?

There *are* no coincidences in mathematics! In the case of the sphere, the Poincaré–Hopf Theorem states that if we sum the indices of *any* vector field on its surface, we will always get 2 for the answer. Indeed, it says that we will get this answer for any surface that is *topologically* a sphere, such as the fried banana in [19.5].

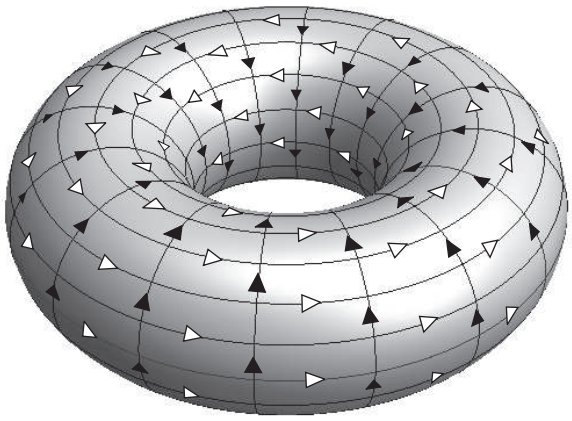
Here is the splendid general result:

> ***Poincaré–Hopf Theorem:*** *If a vector field* $\mathbf{v}$ *on a smooth surface* $\mathcal{S}_g$ *of genus* g *has only a finite number of singular points,* p_i*, then the sum of their indices equals the Euler characteristic of the surface:*

$$\boxed{\sum_i \mathcal{I}_{\mathbf{v}}(p_i) = \chi(\mathcal{S}_g) = 2 - 2g.} \tag{19.6}$$

An immediate consequence of (19.6) is that a vector field without *any* singular points can exist *only* on a surface of vanishing Euler characteristic, i.e., a topological doughnut.

Even then, the theorem does not actually guarantee that such a vector field must exist, it merely demands that the indices of any singular points on a topological doughnut sum to zero. Nevertheless, we can easily *see* that on a doughnut there *do* exist vector fields without any singular points: [19.9] illustrates two such fields, one shown with white arrows, the other with black arrows.



[19.9] *A flow without singular points can* only *exist on a topological doughnut. Here are two such* regular *(topologically distinct) flows on the torus: white and black.*

All the streamlines in the white flow are topologically equivalent to each other, for any one can be continuously deformed into any other. The same is true for the black flow. However, the white and black flows are topologically *distinct* from each other: no white streamline can be deformed into a black streamline.

If we were to "add" these two flows, we could imagine a third, topologically distinct regular flow, such that each streamline went once around the doughnut and once through the hole before joining up again.

More generally, we can construct infinitely many, topologically distinct, regular flows, such that each closed streamline does m (white) circuits around the axis of symmetry, and n (black) circuits through the hole, before closing up on itself. "Adding" such an (m,n)-flow to an (m',n')-flow, by joining the end of one streamline to the start of the other, then yields the $(m+m',n+n')$-flow. This idea leads naturally to what topologists call the ***fundamental group*** of the surface, discovered by Poincaré in 1895. But we digress!

19.6.2 *Proof of the Poincaré–Hopf Theorem*

We can now give a very elegant derivation of the theorem (19.6), due to Heinz Hopf himself (see Hopf 1956, p. 13). The argument proceeds in two steps. First, we show that on a surface of given genus, all vector fields yield the same value for the sum of their indices. Second, we produce a concrete example[4] of a vector field for which the sum equals the Euler characteristic. This proves the result.

Consider the surface $\mathcal{S}$ in [19.10], and suppose that $\mathbf{X}$ and $\mathbf{Y}$ are two different vector fields on $\mathcal{S}$; to avoid clutter, we have only drawn these fields at a single point. If $\bullet$ are the singular points of $\mathbf{X}$, and $\odot$ are those of $\mathbf{Y}$, we must show that

$$\sum_{\bullet} \mathcal{I}_{\mathbf{X}}[\bullet] = \sum_{\odot} \mathcal{I}_{\mathbf{Y}}[\odot].$$

We begin by partitioning $\mathcal{S}$ into curvilinear polygons (dashed curves) such that each one contains at most one $\bullet$ and one $\odot$.

Now concentrate on just one of these polygons (darkly shaded) and its boundary K_j, taken counterclockwise as viewed from outside $\mathcal{S}$. To find the indices of the singular points of $\mathbf{X}$ and $\mathbf{Y}$

[4] We shall actually give a different example than Hopf gave in the reference (Hopf 1956) above.

within K_j, draw any nonsingular fiducial vector field **U** (again only shown at a single point) on the polygon, and then use (19.5). The difference of these indices is then

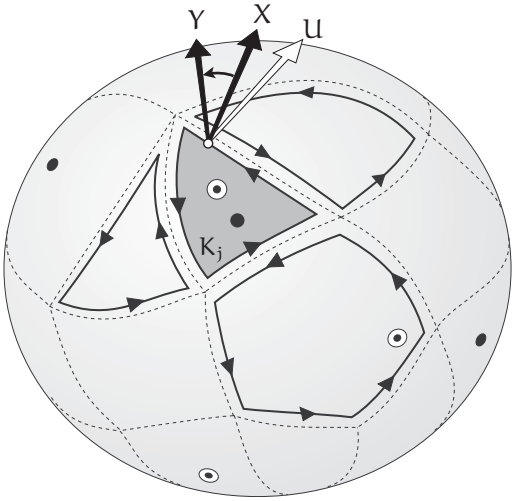
$$\begin{aligned} \mathcal{I}_{\mathbf{Y}}[K_j] - \mathcal{I}_{\mathbf{X}}[K_j] &= \frac{1}{2\pi}\left[\delta_{K_j}(\angle UY) - \delta_{K_j}(\angle UX)\right] \\ &= \frac{1}{2\pi}\,\delta_{K_j}(\angle XY), \end{aligned}$$

which is explicitly independent of the local fiducial vector field **U**; it depends *only* upon the illustrated (signed) angle from **X** to **Y**.

Finally, from this we deduce that

$$\begin{aligned} &\sum_{\odot} \mathcal{I}_{\mathbf{Y}}[\odot] - \sum_{\bullet} \mathcal{I}_{\mathbf{X}}[\bullet] \\ &= \sum_{j} \left(\mathcal{I}_{\mathbf{Y}}[K_j] - \mathcal{I}_{\mathbf{X}}[K_j]\right) \\ &= \frac{1}{2\pi} \sum_{\text{all polygons}} \delta_{K_j}(\angle XY) \\ &= 0, \end{aligned}$$

because (see (19.2)) *every edge of every polygon is traversed once in each direction, producing equal and opposite changes in the (signed) angle* $\angle XY$. We have thus completed the first step: the sum of the indices is independent of the vector field.



[19.10] *The difference of the indices of the singular points of* **X** *(marked •) and of* **Y** *(marked ⊙) inside* K_j *depends only on the rotation of* **Y** *relative to* **X**. *But each edge of* K_j *abuts an oppositely directed edge of a neighbouring polygon, so the sum of these rotations over all polygons must vanish.*

Since the index sum for the examples in [19.8] is 2, we now know that this is the value for *every* vector field on a topological sphere. The second step of the general argument is likewise to produce a concrete example on a surface $\mathcal{S}_g$, of arbitrary genus g, such that the sum is $\chi(\mathcal{S}_g) = (2-2g)$. Figure [19.11] is such an example (here with $g = 3$)—namely, our old friend the honey-flow vector field. As the figure explains, the source at the top and the sink at the bottom each have $\mathcal{I} = +1$, and each of the g holes has two saddle points, each with $\mathcal{I} = -1$. Thus the sum of the indices of this *particular* flow is indeed $2-2g = \chi(\mathcal{S}_g)$, and so this must be the sum of the indices for *every* flow on $\mathcal{S}_g$. Done.

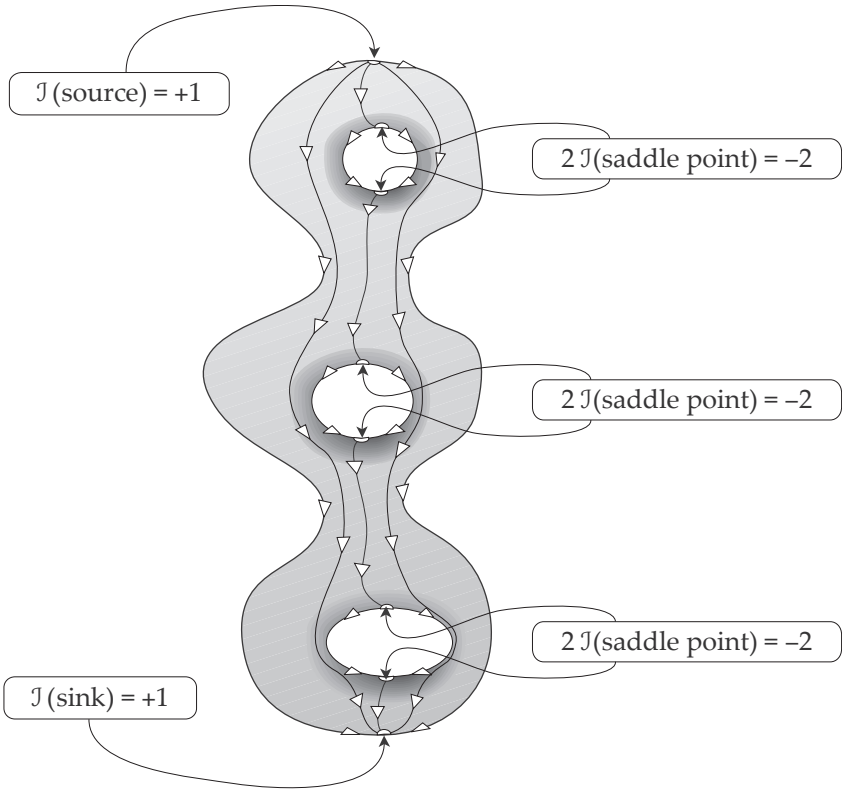
19.6.3 Application: Proof of the Euler–L'Huilier Formula

The Poincaré–Hopf Theorem provides a stunningly immediate proof of the Euler–L'Huilier Formula, (18.7). Recall that the latter states that if we partition a surface $\mathcal{S}_g$ of genus g into polygons, with V vertices, E edges, and F faces, then

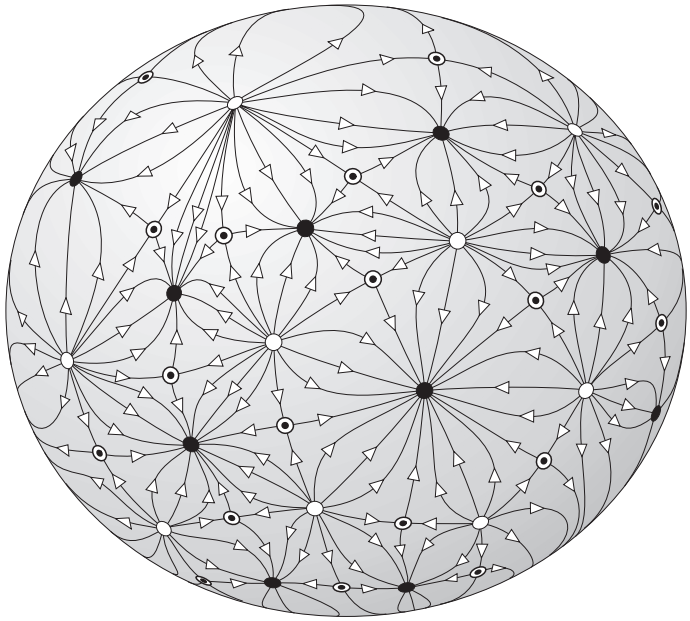
$$V - E + F = 2 - 2g.$$

As [19.12] illustrates, we may construct a consistent ***Stiefel vector field***[5] on $\mathcal{S}_g$ as follows: place a source (marked ◯) at each of the V vertices, a saddle point (marked ⊙) on each of the

[5] According to Frankel (2012, §16.2b), this is due to Hopf's student, Eduard Stiefel (1909–1978).



[19.11] *The honey-flow vector field on a surface of genus* g*. The source (top) and sink (bottom) each contribute* $+1$ *to the index sum, while each of the* g *holes contributes* -2*. Therefore the index sum is* $2-2g=\chi$*.*



[19.12] *Given any polygonal partition of a surface* $\mathcal{S}_g$ *of genus* g*, we may construct a consistent Stiefel vector field on* $\mathcal{S}_g$ *by placing a source* (◯)*—think white hole!—at each of the* V *vertices, a saddle point* (⊙) *on each of the* E *edges, and a sink* (●)*—think black hole!—inside each of the* F *faces. Thus,* $\sum \mathcal{I} = V - E + F$*.*

E edges, and a sink (marked ●) inside each of the F faces. Applying the Poincaré–Hopf Theorem to our constructed vector field, we instantly deduce that

$$2-2g=\sum \mathcal{I}=V\,\mathcal{I}(\bigcirc)+E\,\mathcal{I}(\odot)+F\,\mathcal{I}(\bullet)=V-E+F.$$

Done!

19.6.4 *Poincaré's Differential Equations Versus Hopf's* Line Fields

The 2-dimensional Poincaré–Hopf Theorem, as stated, was actually discovered by Poincaré. However, Hopf's name is correctly attached, for he enormously extended the result in two distinct directions. The first direction was to generalize the result to vector fields on n-dimensional closed manifolds; sadly, to explain this would take us too far afield, but see *Further Reading*. However, the second direction in which Hopf extended the result is quite elementary, and we shall explain it now.

Poincaré was led to investigate the topological behaviour of vector fields by his pioneering work on the *qualitative* theory of differential equations. Such equations became central to physics in 1687, when Newton announced his Second Law of Motion: if a point particle with position vector $\mathbf{X}$, and of mass m, is subjected to a vector force $\mathbf{F}$, then it responds by accelerating in the direction of the force, according to this law:

$$\ddot{\mathbf{X}}=\text{acceleration}=\frac{1}{m}\mathbf{F}. \tag{19.7}$$

Newton's motivation in formulating this law was the determination of the orbits of the planets—part of what is now called ***Celestial Mechanics***. As we discussed in the Prologue, in later life Newton shunned his own youthful discoveries in symbolic calculus, and in the *Principia* he instead employed elegant *geometrical* reasoning to determine the orbit of a planet, given the known (inverse square) law of the Sun's gravitational force acting upon it.

However, from the more traditional, modern point of view, (19.7) is simply a second-order differential equation, which may be solved (*in principle*) by integrating twice: once to find the velocity $\dot{\mathbf{X}}$, and a second time to find the orbit $\mathbf{X}$ itself. In the case of just two gravitating bodies (such as the Earth and Sun) Newton found the exact solution. By introducing the Moon as a *third* body, Newton was able to explain the tides on Earth, and by making *approximations*, he was able to make predictions about the variations of the tides. But Newton and his successors were unable to find an *exact* solution to this 3-body problem.

Finally, more than 200 years later, Poincaré—pictured in [19.13]—made decisive progress by *proving* that the 3-body problem was insoluble.[6] Of course matters only get much worse if we try to analyze the entire solar system, in which each planet is not only held in its orbit by the Sun, but also attracts every *other* planet, and they mutually determine their collective orbits. This is the so-called n-*body problem*.

It is fitting (and perhaps somewhat ironic) that it was Poincaré's return to *geometrical* methods that secured the greatest advance in Celestial Mechanics since Newton. In three volumes (published in 1892, 1893, and 1899), Poincaré published his *New Methods of Celestial Mechanics*,[7] in which he no longer sought to explicitly solve the differential equations (which he had shown to be impossible), but rather to determine the *qualitative behaviour of the solutions* of these equations.

---

[6] To understand what we mean by "insoluble," and for a lively account of Poincaré's voyage of discovery, see Diacu and Holmes (1996).

[7] Poincaré (1899).

To see how this connects with what has gone before, let us consider, as Poincaré did, first-order differential equations in the (x, y)-plane, of the form

$$\frac{dy}{dx} = -\frac{P(x,y)}{Q(x,y)}. \tag{19.8}$$

Poincaré actually wrote these as

$$P(x,y)\,dx + Q(x,y)\,dy = 0.$$

To investigate the integral curves (the solutions), consider an infinitesimal vector $\begin{bmatrix} dx \\ dy \end{bmatrix}$ along the integral curve, and note that the previous equation then implies that

$$\begin{bmatrix} P \\ Q \end{bmatrix} \cdot \begin{bmatrix} dx \\ dy \end{bmatrix} = 0.$$

In other words,

[19.13] *Henri Poincaré (1854–1912)*

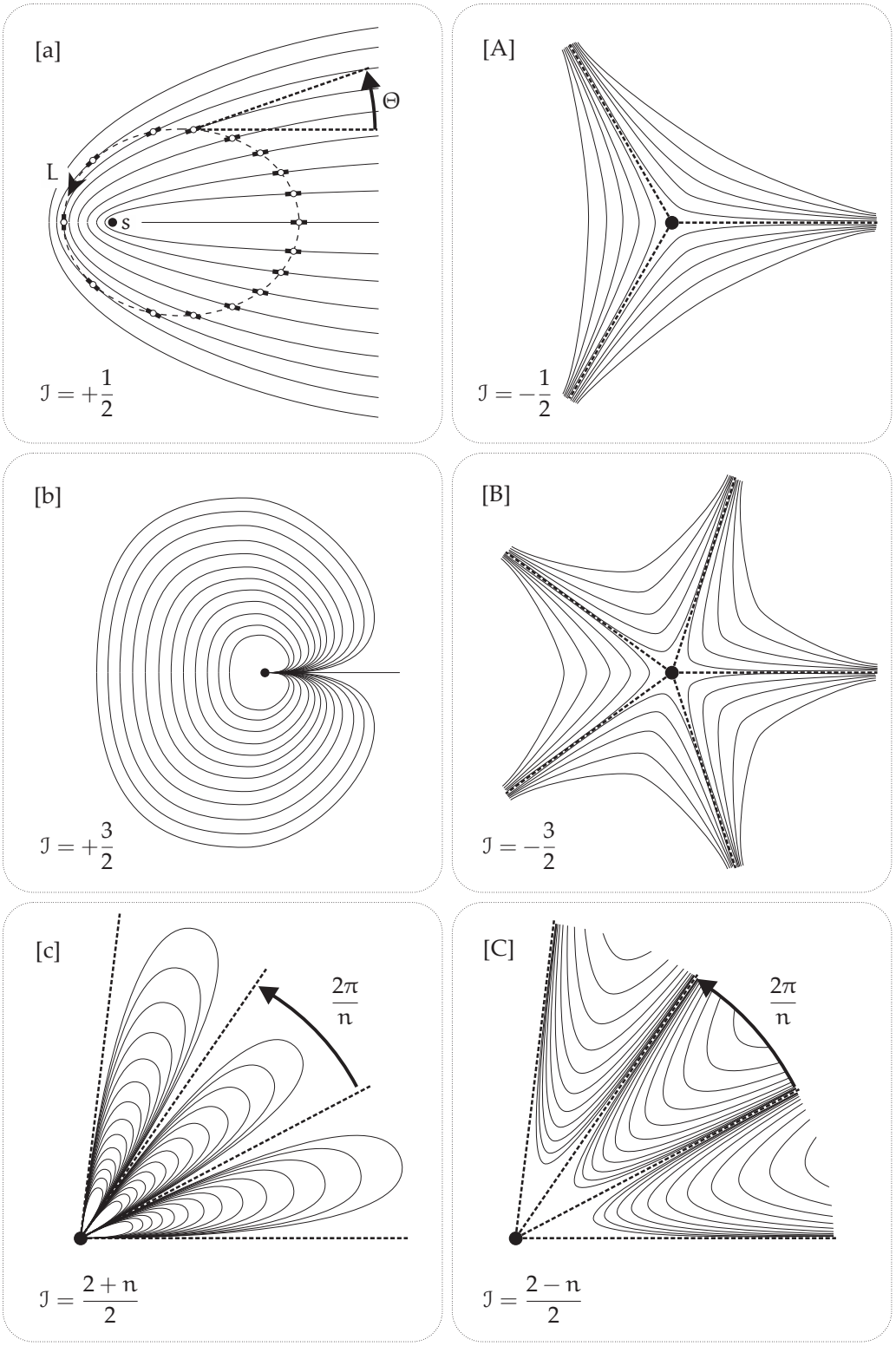
> *The integral curves of (19.8) are everywhere orthogonal to the vector field* $\begin{bmatrix} P \\ Q \end{bmatrix}$.

For example, consider the radial vector field $\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$; this is the source in [19.4c]. Clearly, the orthogonal integral curves are origin-centred circles, so we obtain the centre shown in [19.4e]. Likewise [exercise], $\begin{bmatrix} y \\ x \end{bmatrix}$ yields the saddle point in [19.4h], and $\begin{bmatrix} y \\ -x \end{bmatrix}$ yields either a source or a sink, [19.4c,f].

Now let us turn to Hopf's extensions of Poincaré's discovery. In addition to generalizing to n-dimensional vector fields, which required fundamentally new ideas, Hopf realized that even in 2 dimensions there exist what we might call "directionless flows" that do *not* arise from differential equations of the type considered by Poincaré.

For our first example, consider [19.14a]. We have called such a pattern a "directionless flow," but Hopf actually named this concept a ***field of line elements***. However, the modern terminology, which we shall adopt going forward, is ***line field***, by analogy with *vector field*.

As you see, no arrows are attached to these "streamlines," but we may nevertheless assign a continuously varying angle Θ to the tangent line. If we travel along a simple loop L around the singular point s, the illustrated line element must return to it original position. Previously, with vector fields, the *vector* had to return home pointing in the same direction as when it left, and *therefore it had to undergo a complete number of revolutions*: the rotation was necessarily a multiple of 2π, that multiple being the index.



[19.14] Line Fields *do not have a direction associated with their "streamlines." As a consequence, their singular points can have* fractional *indices, which vector fields cannot have. By repeating the patterns in* **[c]** *and* **[C]**, *we may construct a singular point of arbitrary positive or negative index, respectively. If the positive integer* n *is odd, the index is a fraction, but if* n *is even then we recover the flows of the vector fields shown in [19.4], for which the indices are integers.*

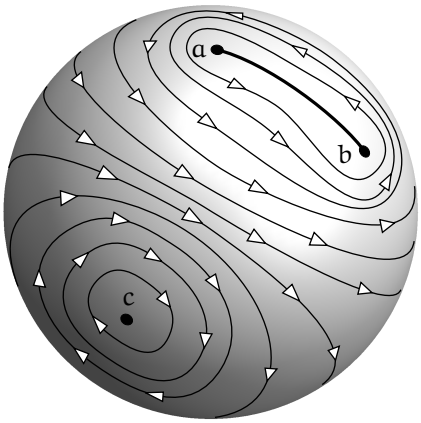
But a *line element* will return to its original orientation if it undergoes a rotation through *any* multiple of π. Indeed, in [19.14a] we see that the net rotation is π. If we retain our original definition of the index, (19.3), as the rotation divided by 2π, then we see that this singular point has a *fractional index*: $\mathcal{I} = +\frac{1}{2}$. Some authors call such fractional indices, ***Hopf indices***.

Please look through all the remaining figures in [19.14] and visually confirm their indices. In particular, [19.14c] and [19.14C] show how to construct a singular point of arbitrary positive or negative index, respectively, simply by repeating the illustrated patterns until they completely surround the singular point. If n is odd, the index is a fraction, but if n is even then we recover the flows of the vector fields shown in [19.4], for which the index is an integer.

Although we cannot interpret a line field with fractional index as the streamlines of an unobstructed flow, we can instead interpret it as *a flow in the presence of 1-dimensional barriers*. For example, in [19.14a], if we think of the horizontal ray to the right of s as such a barrier, then we *can* attach arrows to the curves in a manner that is consistent with a flow around this barrier, flowing to the left above the barrier, say, and to the right below it.

From this perspective, every point of the barrier is a singular point of the vector field, for the direction of the vector flips discontinuously as we cross the barrier. However, for the illustrated, directionless line field, every such point p (other than s) is a *regular* point. For if K is a small loop around p, then you can see that Θ varies continuously on K, so $\mathcal{I}(p) = \frac{1}{2\pi}\delta_K\Theta = 0$.

Now think back over the reasoning that culminated in the Poincaré–Hopf Theorem, concerning the singular points of vector fields. You will discover that it is all *equally applicable* to these more general line fields. In short,



[19.15] *By installing a barrier that connects* a *and* b*, we can create a flow that encircles both points, and that has a third singular point at the centre* c*. The index sum is still equal to the Euler characteristic, despite the* fractional indices *at* a *and* b*.*

> *The Poincaré–Hopf Theorem (19.6) not only applies to vector fields, but also to line fields, with fractional indices.* (19.9)

Figure [19.15] is an example of this result on the sphere. Here we see a line field with three singular points, a, b, and c. We have imagined that the curve connecting a and b is a barrier, and we have *added arrows* to the curves to create a consistent flow that circulates around this barrier. Note that the flow near a and b is the one shown in [19.14a], with $\mathcal{I} = +(1/2)$.

The Poincaré–Hopf Theorem (19.6) predicts that the sum of the indices on a sphere should be $\chi = 2$, and indeed it is:

$$\mathcal{I}(a) + \mathcal{I}(b) + \mathcal{I}(c) = \frac{1}{2} + \frac{1}{2} + 1 = 2.$$

Finally, note that if we shrink the length of the barrier and allow the two $\mathcal{I} = (1/2)$ singular points a and b to approach one another, then they will ultimately coalesce into a single $\mathcal{I} = 1$ centre, creating the flow shown in [19.8a].

THE HISTORY AND FUTURE OF LINE FIELDS: Despite the fact that Hopf gave a wonderfully lucid account of his ideas in Hopf (1956), I am not aware of *any* modern introductory text on Topology, Differential Topology, or Differential Geometry that even mentions this fascinating concept. Indeed, the only example we have found is now more than 50 years old: the classic text by

Stoker (1969, p. 244), who was a doctoral student of Hopf. This is despite the fact that such line fields arise naturally in mathematics. For example, the lines of curvature surrounding an umbilic point on a surface (where $\kappa_1 = \kappa_2$) typically take the form [19.14 a or A]. For a lovely illustration of case [a], see Hilbert (1952, p. 189).

While mathematicians seem to have paid scant attention to the applicability of the Poincaré–Hopf Theorem to line fields with fractional indices, the same is not true of physicists, for Nature herself has thrust such line fields upon them in multiple areas of physics, but especially in optics. The *Further Reading* section at the end of this book seeks to guide you to many of these fascinating new physical applications of Hopf's idea.

19.7 Vector Field Proof of the Global Gauss–Bonnet Theorem

Let V be an arbitrary point on $\mathbb{S}^2$, with unit position vector $\mathbf{V}$. Now imagine a plane with normal $\mathbf{V}$, initially located far from a smooth closed surface $\mathcal{S}$, and with $\mathbf{V}$ pointing away from $\mathcal{S}$. Next, imagine the plane moving towards $\mathcal{S}$. Eventually the plane must make *contact* with $\mathcal{S}$, and at that moment it will become the tangent plane to $\mathcal{S}$ at the point(s) p of contact, and therefore $\mathbf{V}$ must coincide with the surface normal there: $\mathbf{V} = \mathbf{n}(p)$. In other words, V is the spherical image of p: $V = n(p)$. Since this is true for arbitrary $\mathbf{V}$, it must also be true of opposite vector, $-\mathbf{V}$; the plane simply approaches from the other side of $\mathcal{S}$. To sum up,

> *If $\mathcal{S}$ is an arbitrary, closed, smooth surface, then its spherical image covers every antipodal pair of points $\pm V$ of $\mathbb{S}^2$ at least once.* (19.10)

Of course, V and $-V$ may each be covered *more* than once. For example, in [19.5] we see that V is covered once, while $-V$ is covered three times.

Let us now to return to GGB, and let us briefly recap our very first "heuristic" proof in Chapter 17. Look at these coverings of V from the point of view of the total curvature, $\mathcal{K}(\mathcal{S})$. If $\mathcal{K}(p) > 0$ at p, then n preserves the orientation of a small patch of area $\delta\mathcal{A}$ surrounding p, so its spherical image of area $\delta\widetilde{\mathcal{A}} \asymp \mathcal{K}(p)\,\delta\mathcal{A}$ surrounding V on $\mathbb{S}^2$ has the *same* orientation. As we discussed in Section 17.4, we count this as a *positive* covering of V. Likewise, if $\mathcal{K}(p) < 0$, then the spherical map *reverses* the orientation, and the covering is counted as *negative*. The total curvature integral automatically counts these coverings *algebraically*, taking into account orientation. As earlier, let $\mathcal{P}(V)$ and $\mathcal{N}(V)$ denote the number of positive and negative coverings of V, respectively, so that the net number of coverings of V is $[\mathcal{P}(V) - \mathcal{N}(V)]$. Thus,

> *Let V be a point on $\mathbb{S}^2$, and let its preimages (under the spherical map n) be the points p_i of $\mathcal{S}$. Consider a small (ultimately vanishing) patch around V (with area $\delta\widetilde{\mathcal{A}}$) and let the areas of the small preimage patches around p_i be $\delta\mathcal{A}_i$. Then,*
>
> $$\text{Total curvature within the } \delta\mathcal{A}_i \asymp \sum_i \mathcal{K}(p_i)\,\delta\mathcal{A}_i \asymp [\mathcal{P}(V) - \mathcal{N}(V)]\,\delta\widetilde{\mathcal{A}}. \tag{19.11}$$

For example, in [19.5] we see that the curvatures at p_1 and p_2 are positive, and the coverings of $-V$ originating from these points are positive. On the other hand, the negatively curved patch containing p_3 is mapped to an orientation-reversed, negative covering of $-V$. The *net algebraic* count of the coverings of $-V$ is therefore 1. Likewise, the net number of coverings of V is 1.

Recall that this net number of coverings of $\mathbb{S}^2$ is the ***topological degree*** $\deg(n)$ of the spherical mapping, and the fundamental fact is that the degree is the same for all[8] points of $\mathbb{S}^2$. This was the essence of our heuristic explanation of GGB in Chapter 17. The final step of that explanation was the recognition that the degree depends solely on the genus g of $\mathcal{S}_g$. This is (16.7), on page 174, which we restate here:

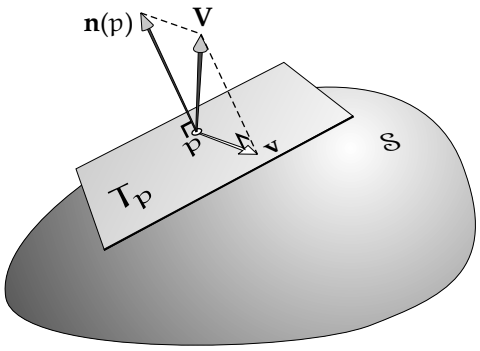
$$\deg[n(\mathcal{S}_g)] = \mathcal{P} - \mathcal{N} = (1-g) = \tfrac{1}{2}\chi(\mathcal{S}_g). \tag{19.12}$$

As we noted in Section 16.6, this was first proved by Walther Dyck in 1888.

If we take it as *given* that the topological degree exists, so that every point of $\mathbb{S}^2$ is covered the same (algebraic) number of times, then the Poincaré–Hopf Theorem can be used to give an elegant proof of Dyck's result, and, with it, GGB. Indeed, this is the approach to GGB taken in such well-known works as Guillemin and Pollack (1974, p. 198).

We too shall reproduce that argument shortly. However, the existence of the topological degree is by no means obvious, so we now present a slightly different argument that does not depend upon it. The argument stills hinges on the Poincaré–Hopf Theorem, but it does *not* assume the existence of the topological degree. That is, if V and W are two points of $\mathbb{S}^2$, we shall *not* assume that they must be covered the same number of times (algebraically); instead, we shall entertain at least the possibility that $\deg[n(\mathcal{S}), V] \neq \deg[n(\mathcal{S}), W]$. Henceforth, we shall take it as understood that we are dealing with the spherical map of $\mathcal{S}$, so this (hypothetical and in fact impossible) nonequality can then be abbreviated to $\deg[V] \neq \deg[W]$.

To prove GGB without assuming the existence of the topological degree, reconsider the honey-flow in [19.5]. The first important observation is that the singular points of this flow occur at precisely those points where the vertical gravitational force $\mathbf{V}$ has zero component within the surface—that is, at points where the vertical force vector points directly into (or out of) the surface—that is, where the outward surface normal $\mathbf{n} = \pm\mathbf{V}$. Of course, mathematically speaking, there is nothing special about this gravitational direction $\mathbf{V}$—we may imagine gravity pulling in an arbitrary direction; alternatively, if we do not wish to tamper with the Earth's gravitational field, we may simply rotate the surface!

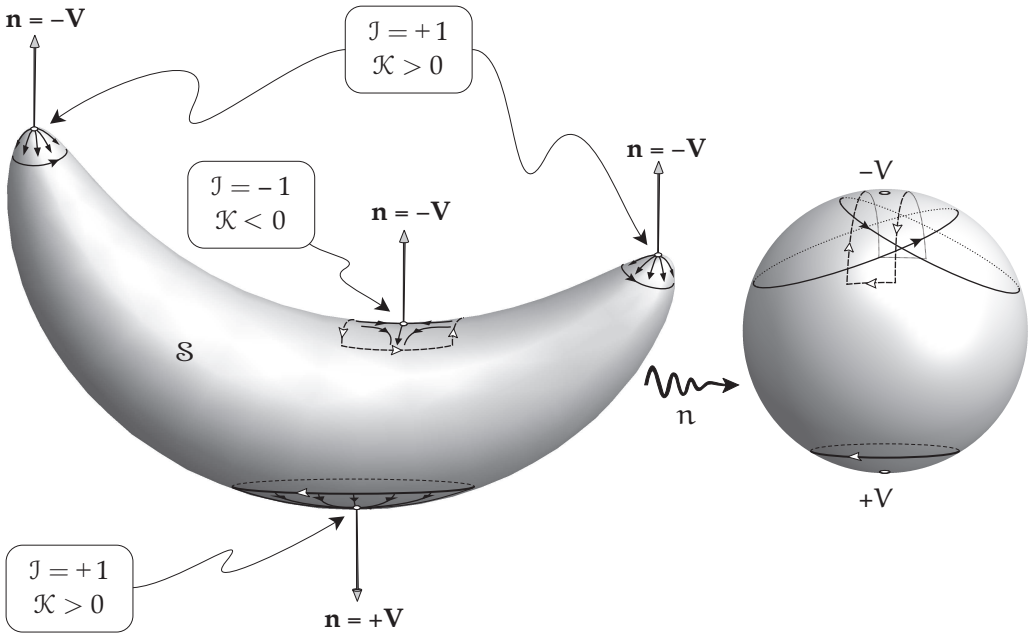


[19.16] *The* ***honey-flow*** $\mathbf{v}$ *on the surface* $\mathcal{S}$ *in the direction* $\mathbf{V}$ *is the orthogonal project of* $\mathbf{V}$ *onto the tangent plane* T_p *at* p.

To put this more mathematically, our "honey-flow" vector field is obtained by orthogonal projection of $\mathbf{V}$ onto the surface, or, more precisely, onto the tangent plane T_p of the surface. See [19.16]. Let us make this definition official:

The ***honey-flow*** vector field in the direction $\mathbf{V}$ is $\mathbf{v}(p) \equiv \text{proj}_{T_p}\mathbf{V}$. (19.13)

[8] As we have discussed, the exceptions are points of vanishing curvature, which contribute nothing to the total curvature.



[19.17] *The singular points of the honey-flow* $\mathbf{v}$ *on* $\mathcal{S}$ *in the direction* $\mathbf{V}$ *are mapped by the spherical map* $\mathfrak{n}$ *to either* $+V$ *or* $-V$ *on* $\mathbb{S}^2$*, and* ***the orientation of the covering on*** $\mathbb{S}^2$ ***is determined by the sign of the index*** *of* $\mathbf{v}$ *on* $\mathcal{S}$*. Here* $+V$ *is covered once, positively, by the image of the sink at the bottom of the banana—where* $\mathcal{I} = +1$ *and* $\mathcal{K} > 0$*. On the other hand,* $-V$ *is covered three times: twice positively by the images of the two sources—where* $\mathcal{I} = +1$ *and* $\mathcal{K} > 0$*—and once negatively by the image of the saddle point—where* $\mathcal{I} = -1$ *and* $\mathcal{K} < 0$*.*

The singular points $\mathbf{v}(p_i) = 0$ occur when $\mathbf{n}(p_i) = \pm\mathbf{V}$. In other words,

> *If honey flows over* $\mathcal{S}$*, pulled in the direction* $\mathbf{V}$ *in space, the singular points of the honey-flow* $\mathbf{v}$ *consist of the complete set of points that are sent by the spherical map to either* V *or to the antipodal point* $-V$ *on* $\mathbb{S}^2$*.*

Out of all possible vector fields on a surface, why have we lavished so much attention on the *honey-flow*? The answer is that there exists a crucial link between the *geometry* of the surface, and the *topology* of the honey-flow. This in turn will yield our third explanation of GGB, as a consequence of the Poincaré–Hopf Theorem.

Reconsider the singular points p_i of the honey-flow vector field $\mathbf{v}$ illustrated in [19.5]. These are shown again in [19.17], but now focusing attention on how the spherical map $\mathfrak{n}$ sends them to $\mathbb{S}^2$.

The decisive observation is the distinction between the singular points of $\mathbf{v}$ that have positive curvature, and those that have negative curvature.

Recall (12.9) and [12.7], on page 136. If $\mathcal{K}(p_i) > 0$, the surface is locally a dome facing either straight up (in which case it is a source with $\mathcal{I}(p_i) = +1$) or straight down (in which case it is a sink, again with $\mathcal{I}(p_i) = +1$). But if $\mathcal{K}(p_i) < 0$, then the surface is locally a saddle, producing a corresponding saddle point in the flow, in which case $\mathcal{I}(p_i) = -1$.

This immediately implies that the spherical map preserves orientation if $\mathcal{I}(p_i) = +1$, and reverses it if $\mathcal{I}(p_i) = -1$:

If $\mathfrak{p}$ *is a singular point of the honey-flow* $\mathbf{v}$*, then its image* $(\pm V)$ *under the spherical map* $\mathfrak{n}$ *is covered positively if* $\mathcal{I}(\mathfrak{p}) = +1$*, and negatively if* $\mathcal{I}(\mathfrak{p}) = -1$*.*

Now let us combine this result with the Poincaré–Hopf Theorem. If $\mathfrak{p}_i$ is the set of singular points of $\mathbf{v}$ on $\mathcal{S}$, i.e., those whose spherical image is either $+V$ or $-V$, then

$$\chi(\mathcal{S}) = \sum_i \mathcal{I}_v(\mathfrak{p}_i) = \{\mathcal{P}(+V) - \mathcal{N}(+V)\} + \{\mathcal{P}(-V) - \mathcal{N}(-V)\}. \tag{19.14}$$

In reality, both bracketed terms on the right are equal to each other, being the topological degree of the spherical map. See (19.12). For example, in [19.17] we see that

$$\mathcal{P}(+V) - \mathcal{N}(+V) = 1 - 0 = 1$$

and

$$\mathcal{P}(-V) - \mathcal{N}(-V) = 2 - 1 = 1.$$

However, (19.14) enables to prove GGB *without* assuming this fact. To do so, note that the singular points of the single flow $\mathbf{v}$, corresponding to honey flowing in the direction $\mathbf{V}$, actually include the singular points of *both* $\mathbf{V}$ *and* $-\mathbf{V}$. For reversing the direction of the honey flow does not alter the locations of the singular points, nor their indices.

Now if V roams over just the northern hemisphere of S^2 (the area of which is 2π) then $-V$ roams over the southern hemisphere, and therefore $\pm V$ covers the *entire* S^2. Thus, by virtue of (19.10), as V roams over just the northern hemisphere, the singular points of $\mathbf{v}$ cover the *entire* surface $\mathcal{S}$. And since we have just seen that the indices of the singular points correctly count the covers of their images under the spherical map, (19.11) and (19.14) yield our third proof of GGB, in the form

$$\mathcal{K}(\mathcal{S}) = 2\pi\chi(\mathcal{S}).$$

Let us return to the topic of the topological degree, and thereby obtain a bonus result. If we *assume* that the topological degree exists, then $\deg(V) = \deg(W)$, for all V and W on S^2. In particular, it follows that $\deg(+V) = \deg(-V)$. In that case, (19.14) becomes,

$$\begin{aligned} \chi(\mathcal{S}) &= \{\mathcal{P}(+V) - \mathcal{N}(+V)\} + \{\mathcal{P}(-V) - \mathcal{N}(-V)\} \\ &= \deg(+V) + \deg(-V) \\ &= 2\deg(V) \\ &= 2\deg, \end{aligned}$$

in which the final equality reflects the fact that the degree is independent of the choice of V. Thus, if we assume the existence of the topological degree, we recover Dyck's result, (19.12):

$$\deg = \tfrac{1}{2}\chi(\mathcal{S}).$$

19.8 The Road Ahead

All three of the proofs of GGB thus far presented have relied upon the interpretation of $\mathcal{K}$ as the local area expansion factor of the spherical map. This is an *extrinsic* conception of curvature, depending as it does upon the normal $\mathbf{n}$, which is invisible and unknowable to the inhabitants of $\mathcal{S}$. While the three proofs have afforded us wonderful new insights in many different directions, something is lacking, for we know that $\mathcal{K}$ is also, in fact, an *intrinsic* property of the surface, knowable to and measurable by its inhabitants. In principle, these inhabitants could measure $\mathcal{K}$ throughout the surface, and from these purely local *geometric* measurement they could determine the *topology* of their world!

We are about to embark on Act IV, which introduces a brand new, and extremely powerful, *intrinsic* way of understanding and measuring curvature. This will allow us to finally explain some of the fundamental results we have been forced to assume up till now, such as the Local Gauss–Bonnet Theorem—and with it the *Theorema Egregium*—as well as the remarkable ("*Star Trek* phaser") formula (4.10), page 38, for the curvature in terms of the metric. Furthermore, it will allow us to remedy the noted deficiency in our three existing proofs of GGB. Indeed, using an idea of Heinz Hopf, we shall finally be able to provide a proof of GGB that is entirely *intrinsic*.

Chapter 20

Exercises for Act III

Curvature of Plane Curves

1. **Computational Proof of the Curvature Formula.** Figure [8.7] provided a geometric proof of (8.7). Prove this instead by calculation. (*Hint*: If the curve $[x(t), y(t)]$ is traced at unit speed, then $\dot{x} = \cos\varphi$, and $\dot{y} = \sin\varphi$.)

Curves in 3-Space

2. **Binormal Cannot Tip in the Direction of Motion.** Prove by calculation that as the Frenet frame $(\mathbf{T}, \mathbf{N}, \mathbf{B})$ moves along its curve, the binormal $\mathbf{B}$ only spins around $\mathbf{T}$; it cannot tip in the direction of $\mathbf{T}$. (*Hint*: Symbolically, what must be proved is that $\mathbf{B}'$ has no component in the $\mathbf{T}$ direction: $\mathbf{B}' \bullet \mathbf{T} = 0$.)

3. **The Darboux Vector.** As the Frenet frame moves along its curve, it rotates. Verify by calculation that its instantaneous *angular velocity vector* is

$$\mathbf{A} \equiv \tau\,\mathbf{T} + \kappa\,\mathbf{B}.$$

   This vector $\mathbf{A}$ is sometimes referred to as the ***Darboux vector*** (see Stoker 1969, p. 62) in honour of its discoverer, Jean-Gaston Darboux (1842–1917), a pioneer of Differential Geometry, and a teacher of Cartan. *Hint*: Use (9.3) to prove that

$$\begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}' = \mathbf{A} \times \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}.$$

4. **Variable Speed Frenet–Serret Equations.** The Frenet–Serret equations (9.3) assume the derivative is with respect to arc length, which is only the same as the time derivative if the curve is traced at unit speed. Suppose instead that the particle has variable speed $v = \dot{s}$.

   (i) Show that $[\Omega] \mapsto v[\Omega]$:

$$\begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}^{\bullet} = v \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}.$$

   (ii) Prove that the acceleration is $\dot{\mathbf{v}} = [v\mathbf{T}]^{\bullet} = \dot{v}\,\mathbf{T} + \kappa v^2 \mathbf{N}$, and interpret and explain both terms geometrically.

   (iii) Show that $\mathbf{B} = \dfrac{\mathbf{v} \times \dot{\mathbf{v}}}{|\mathbf{v} \times \dot{\mathbf{v}}|}$.

   (iv) Show that $\kappa = \left| \dfrac{\mathbf{v} \times \dot{\mathbf{v}}}{v^3} \right|$.

   (v) Show that $\tau = \dfrac{(\mathbf{v} \times \dot{\mathbf{v}}) \bullet \ddot{\mathbf{v}}}{|\mathbf{v} \times \dot{\mathbf{v}}|^2}$.

5. **The Helix.** Consider the path of a particle whose position at time t is $(R\cos\omega t, R\sin\omega t, qt)$.
   (i) Explain why this is a helix, and state the geometrical/physical interpretations of R, ω, and q.
   (ii) Prove that $v=\sqrt{(R\omega)^2+q^2}$, and explain this geometrically/physically.
   (iii) Use the previous question to calculate κ and τ.
   (iv) Explain *geometrically* why $\lim_{\omega\to\infty}\kappa=(1/R)$, and $\lim_{q\to\infty}\kappa=0$.
   (v) Use (iii) to confirm the predictions of (iv).

6. **The Frenet–Serret Approximation to a Curve.** Let $\mathbf{x}(t)$ be the position at time t of a particle that traces a curve C in space at unit speed. Let $(\mathbf{T}_0,\mathbf{N}_0,\mathbf{B}_0)$ be the Frenet–Serret frame at time $t=0$, and let κ_0 and τ_0 be the curvature and torsion at this time.
   (i) Using Taylor's Theorem and the Frenet–Serret equations (9.3), prove that the motion along C is initially given by the ***Frenet Approximation***:
   $$\mathbf{x}(t)\asymp\mathbf{x}(0)+t\,\mathbf{T}_0+\kappa_0\frac{t^2}{2}\,\mathbf{N}_0+\kappa_0\tau_0\frac{t^3}{6}\,\mathbf{B}_0.$$
   (ii) Explain the first three terms geometrically.
   (iii) What does the fourth term describe, geometrically?
   (iv) In light of your answer to (iii), why *must* this term vanish when $\tau_0=0$?
   (v) Why does it make sense that this fourth term should also vanish if $\kappa_0=0$?

The Principal Curvatures of a Surface

7. **Series Expansion of Surfaces.** For each of the following equations, use a computer to draw the surface. Use series expansions to calculate the quadratic approximation near the origin, then use (10.6), page 111, to deduce the principal curvatures and $\mathcal{K}$ there. Visually confirm that your calculations are at least qualitatively correct.
   (i) $z=\exp(x^2+4y^2)-1$.
   (ii) $z=\ln\cos y-\ln\cos 2x$.

8. **Variable-Speed Formulas for the Curvature of a Surface of Revolution**. In the text, we pictured a particle moving at constant unit speed along the generating curve of a surface of revolution, in which case the principal curvatures are given by (10.11) and (10.12). If the speed $v(t)$ is *not* held constant, recall that (10.11) becomes (8.6):
   $$\kappa_1=\frac{\dot{x}\,\ddot{y}-\dot{y}\,\ddot{x}}{v^3}.$$
   (i) By modifying [10.6], page 114, deduce that the second principal curvature is given by
   $$\kappa_2=-\frac{\dot{x}}{yv}.$$
   (ii) Deduce that
   $$\boxed{\mathcal{K}=-\frac{\dot{x}\left[\dot{x}\,\ddot{y}-\dot{y}\,\ddot{x}\right]}{y\left[\dot{x}^2+\dot{y}^2\right]^2}.}$$

(iii) If the particle only moves to the right, never retracing any x-value, then we are free to adjust the speed along the curve so that x represents *time*: $x(t) = t$, in which case $y(t) = y(x)$, and the time derivative becomes the x-derivative: $\dot{y} = y'$. Show that the curvature formula in (ii) then reduces to

$$\boxed{\mathcal{K} = -\frac{y''}{y\left[1 + (y')^2\right]^2}.} \tag{20.1}$$

9. Polar Formula for the Curvature of a Surface of Revolution. In the previous exercise, the surface of revolution was defined using the distance of the generating curve from the rotation axis, which was taken to be the x-axis. Let us now, instead, describe such a surface by giving its height above a plane perpendicular to its axis of symmetry, which we now take to be the z-axis. If r is the usual polar coordinate in the (x, y)-plane, consider the surface of revolution $z = f(r)$.

(i) By simply changing names appropriately, and using the chain rule, show that (20.1) becomes

$$\boxed{\mathcal{K}(r) = \frac{f'(r)\, f''(r)}{r\left\{1 + [f'(r)]^2\right\}^2}.}$$

(ii) Find $f(r)$ for a sphere or radius R centred at the origin. Check that the formula in (i) gives the correct value of $\mathcal{K}$.

(iii) Manually sketch the surface $z = \exp[-r^2/2]$, calculate $\mathcal{K}$, and find the regions of positive, negative, and zero curvature. (A computer-generated graph of the surface should visually confirm your answers.)

10. In the text we noted that it is visually evident that the surface $z = r^4$ has a planar point (with $\mathcal{K} = 0$) at the origin, and that the surface has positive curvature everywhere else. Prove this is correct by using the formula in the previous question to calculate $\mathcal{K}(r)$.

Gauss's *Theorema Egregium*

11. Why Paper Folds into a Straight Line. We take it for granted that when we fold a piece of paper, the fold automatically forms itself into a *straight line*, but *why* does this happen? Show (without any calculation) that this is a direct consequence of the *Theorema Egregium*! (I owe this delightful insight to my colleague Dr. Robert Wolf, a former student of Chern.)

The Shape Operator

12. Visual Linear Algebra. Do all of the following by reasoning directly and *geometrically* about the linear transformations themselves, *not* by applying the "Devil's machine" (see Prologue) to the matrices that represent them! We hope that these examples (together with those in the text) will inspire you to bring this geometrical perspective to bear whenever you next encounter Linear Algebra.

(i) Verify that the geometric interpretation [15.4] of the SVD generalizes to $\mathbb{R}^3$, and therefore the interpretation [15.5] of M^T does too.
NOTE: Although visualization becomes harder, the same is true in $\mathbb{R}^n$.

(ii) Recall that an ***orthogonal*** linear transformation is one that preserves lengths, and therefore angles (in general), and orthogonality (in particular)—examples include rotations and reflections. In $\mathbb{R}^3$, use (i) to explain why any orthogonal transformation R has the property that $R^{-1} = R^T$.

(iii) Recall that a matrix $[M]$ is called ***skew symmetric*** if $[M]^T = -[M]$. In $\mathbb{R}^2$, use (15.12) to give a *geometric* characterization (in terms of the twist, τ) of the underlying skew-symmetric linear transformation M.

(iv) A symmetric linear transformation P is called ***positive definite*** if $\mathbf{x} \bullet (P\mathbf{x}) > 0$, for all $\mathbf{x}$. Use (15.13) to show that P is positive definite if (and only if) all of its eigenvalues are positive.

(v) Arguing directly, or, alternatively, building on (iv), show that a symmetric, positive-definite linear transformation must have positive determinant (and hence be invertible). Show that the converse is *false*, by finding a simple counterexample in $\mathbb{R}^3$. (*Hint*: Geometrically, the determinant is the *(signed) volume-expansion factor* of the linear transformation, the sign being postive (+) or negative (−) according to whether orientation is preserved or reversed, respectively.)

(vi) Given (iv) above, it follows from (15.15) that M^TM and MM^T are always symmetric and positive definite. Conversely, show that *any* symmetric, positive-definite linear transformation P can be factorized as $P = M^TM$, in infinitely many ways.

(vii) Building on (vi), show that just as any positive number has a real square root, so any positive-definite linear transformation P has a square root Q, such that $P = Q^2$. In greater detail, show that $Q = R^{-1}DR$, where R is orthogonal, and $[D]$ is the diagonal matrix whose entries are the square roots of the eigenvalues of P.

(viii) A symmetric linear transformation S is called ***positive semidefinite*** if $\mathbf{x} \bullet (S\mathbf{x}) \geqslant 0$, for all $\mathbf{x}$. If M is any linear transformation of $\mathbb{R}^3$, show that $M = RS$, where R is orthogonal, and S is symmetric and positive semidefinite. This is called the ***Polar Decomposition*** of M.

13. Vanishing Shape Operator $\Longleftrightarrow$ Flat. Show, both geometrically and by calculation, that if the Shape Operator vanishes identically, then the surface is a portion of a plane.

14. Gaussian Curvature in terms of the Shape Operator. If $\mathbf{u}$ and $\mathbf{v}$ are short tangent vectors at a point p on a surface with Shape Operator S, show geometrically that $S(\mathbf{u}) \times S(\mathbf{v}) = \mathcal{K}(p)\ [\mathbf{u} \times \mathbf{v}]$. (NOTE: The equation itself is valid even if the tangent vectors are *not* short, but the geometric *explanation* relies on thinking in terms of ultimate equalities.)

15. Cartesian Formulas for the Shape Operator, Curvature, and Mean Curvature.
Let T_p be the tangent plane of a general surface $\mathcal{S}$ at a point p. Introduce Cartesian coordinates (x, y) into T_p, and take $p = (0, 0)$, so that the normal $\mathbf{n}$ there points along the z-axis. Then, locally, $\mathcal{S}$ is given by

$$z = f(x, y) \quad \textit{where} \qquad f(0,0) = 0 \quad \text{and} \quad \partial_x f = 0 = \partial_y f \quad \text{at } (0,0).$$

(i) Note that if we define $F(x, y, z) \equiv f(x, y) - z$, then F is constant on $\mathcal{S}$. Deduce that ∇F is normal to $\mathcal{S}$, and therefore,

$$\mathbf{n} = \frac{1}{\sqrt{1 + (\partial_x f)^2 + (\partial_y f)^2}} \begin{pmatrix} \partial_x f \\ \partial_y f \\ -1 \end{pmatrix}.$$

(ii) Deduce that the matrix of the Shape Operator at $(0,0)$ is

$$\boxed{[S] = \begin{bmatrix} \partial_x^2 f & \partial_y \partial_x f \\ \partial_x \partial_y f & \partial_y^2 f \end{bmatrix},} \tag{20.2}$$

from which it follows immediately that

$$\boxed{\mathcal{K} = \kappa_1 \kappa_2 = \det [S] = (\partial_x^2 f)\,(\partial_y^2 f) - (\partial_x \partial_y f)^2.} \tag{20.3}$$

Likewise, $\overline{\kappa} \equiv$ ***mean curvature*** $= \left[\frac{\kappa_1 + \kappa_2}{2}\right] = \frac{1}{2}\mathrm{Tr}\,[S] = \frac{1}{2}\left[\partial_x^2 f + \partial_y^2 f\right]$, which may be written

$$\boxed{\overline{\kappa} = \tfrac{1}{2}\nabla^2 f,}$$

where $\nabla^2 = \partial_x^2 + \partial_y^2$ is the ***Laplacian operator***, first encountered on page 41.

16. Nonprincipal Coordinates. Let us reanalyze the saddle surface, this time *without* first aligning the coordinate axes with the principal directions, as we did in the text.

(i) Sketch the surface $z = xy$ near the origin. (*Hint*: Recalling that $\sin 2\theta = 2\sin\theta\cos\theta$, find the height of the surface above and below the points of the origin-centred circle of radius r, for which $x = r\cos\theta$ and $y = r\sin\theta$.)

(ii) Use (10.4), p. 111, to identity the principal directions.

(iii) Use (20.2) to deduce that the Shape Operator has the matrix

$$[S] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

and deduce that $\mathcal{K} = -1$ and $\overline{\kappa} = 0$.

(iv) What is the geometric transformation S represented by $[S]$?

(v) Use (iv) to determine the eigenvectors and eigenvalues of S, geometrically (*not* by calculation). (NOTE: The eigenvectors should point in the same directions you found in (ii).)

(vi) Guided by either (ii) or (v), rotate the (x,y)-axes to obtain new (X,Y)-axes aligned with the principal directions, and show that the original equation now becomes, $z = \frac{1}{2}(X^2 - Y^2)$.

(vii) By comparing the new equation in (v) with (10.6), page 111, confirm the values of $\mathcal{K}$ and $\overline{\kappa}$ you found in (iii).

(viii) Use (20.2) to write down $[S]$ in the new (X,Y)-coordinate system of (vi), and confirm that $\mathcal{K}$ and $\overline{\kappa}$ have not changed.

(ix) Verify that the two different matrices $[S]$ in (iii) and (viii) (representing the same linear transformation S) conform to the general matrix formula (15.20), page 161.

17. Use (20.2) and (20.3) to verify your conclusions in Exercise 7.

18. Curvature Formula in Polar Coordinates. If $z = f(r,\theta)$, show that the curvature is given by this disappointingly complicated formula:

$$\mathcal{K} = \frac{r^2\,\partial_r^2 f\,\left(\partial_\theta^2 f + r\,\partial_r f\right) - \left[\,\partial_\theta f - r\,\partial_r\partial_\theta f\,\right]^2}{\left\{r^2\,\left[1 + (\partial_r f)^2\right] + (\partial_\theta f)^2\,g\right\}^2} \tag{20.4}$$

19. Nonisometric Surfaces with Equal Curvatures. As we discussed in Act I, Minding proved that if two surfaces have the same *constant* $\mathcal{K}$, then they are locally isometric to each other. Thus, according as this constant curvature $\mathcal{K} > 0$, $\mathcal{K} = 0$, $\mathcal{K} < 0$, the surface is locally isometric to a sphere, plane, or pseudosphere, respectively. We now provide an example to show that the converse is false.

(i) Using the same notation as in the previous exercise, find the shapes of these two surfaces: $\mathcal{S}_1$ with $f_1(r,\theta) = \ln r$, and $\mathcal{S}_2$ with $f_2(r,\theta) = \theta$. (*Hint*: $\mathcal{S}_2$ is called the ***helicoid***.) Check your answers by using a computer to draw them.

(ii) Let us say that a point in $\mathcal{S}_1$ and a point in $\mathcal{S}_2$ "correspond" if they have the same (r,θ)-coordinates. By finding the formulas for the metrics of both surfaces, show that this correspondence is *not* an isometry.

(iii) Nevertheless, use (20.4) to deduce that corresponding points of the two surfaces have the *same* curvature!

$$\mathcal{K}_1(r,\theta) = \frac{-1}{(r^2+1)^2} = \mathcal{K}_2(r,\theta).$$

20. Curvature of the *n*-Saddle. Given that the height of the n-saddle is $z = f(r,\theta) = r^n \cos n\theta$, use (20.4) to prove that its (negative) curvature is constant on each circle $r = const.$, and is given by the following formulas.

(i) The common 2-saddle (surrounding a generic hyperbolic point) has

$$\mathcal{K} = -\frac{4}{(1+4r^2)^2}.$$

(ii) The monkey 3-saddle has

$$\mathcal{K} = -\frac{36r^2}{(1+9r^4)^2}.$$

(iii) The n-saddle has

$$\mathcal{K} = -\frac{(n-1)^2 n^2 r^{2n}}{(r^2+n^2r^{2n})^2}.$$

Introduction to the Global Gauss–Bonnet Theorem

21. Genus-Dependent Predictions of GGB.

(i) Use GGB to deduce that any topological sphere *must* have regions of positive curvature.

(ii) Use GGB to deduce that any topological doughnut *must* have regions of positive curvature.

(iii) In fact *every* smooth closed surface $\mathcal{S}_g$ with $g \geqslant 2$ must also have a point of positive curvature, though GGB alone no longer guarantees this. Let $S(r)$ be a sphere of radius r centred at an arbitrary point inside $\mathcal{S}_g$. Imagine we start with a sufficiently large value of

r that $S(r)$ contains $\mathcal{S}_g$. Now shrink the sphere until it makes first contact with $\mathcal{S}_g$, when $r = R$, say. Argue that at the point of contact with $S(R)$, $\mathcal{K} > (1/R^2) > 0$.
Hint: Consider the principal directions at the point of contact.

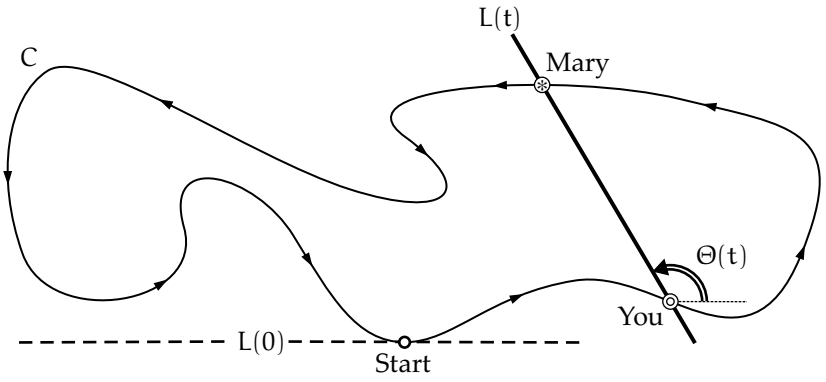
22. Holy Broken Gauss–Bonnet, Batman! Recall our definition of a ***wormhole*** in [16.7], and also recall that it has total curvature, $\mathcal{K} = -4\pi$.

(i) If n small bagels are placed at the vertices of a large regular n-gon, and are joined together with wormhole connectors along the edges of the n-gon into a surface $\mathcal{S}$, explain why $\mathcal{K}(\mathcal{S}) = -4\pi n$. Why is this *not* a violation of GGB, as it appears to be?

(ii) Suppose $n = 4$, so that the bagels are at the corners of a square, joined along the four edges. Now add a *fifth* wormhole connector along one of the diagonals, connecting two previously disconnected bagels. What is the total curvature now, and why is this *not* a violation of GGB, as it appears to be?

First (Heuristic) Proof of GGB

23. Hopf's Umlaufsatz. Let us attempt to reduce Hopf's (1935) ingenious argument to its essence. Imagine that the (smooth and "simple") closed curve C of the theorem is a convoluted hiking/running path snaking through sand dunes. See map below. The entrance is at the southern-most point (downward on the page); the rest of C lies to the north (upwards). As illustrated, let $\Theta(t)$ be the angle of the line $L(t)$ connecting you (⊙) and your friend, Mary (⊛), at time t. Initially, Mary is standing next to you, so $L(0)$ (dashed line) is horizontal, with $\Theta(0) = 0$.



(i) Mary decides to do a warm-up lap, but you stay at the start and just watch her run. As she runs the length of C, counterclockwise, you turn your head to follow. Explain why your line of sight $L(t)$ must execute a *net* rotation of π by the time Mary returns to you: $\delta\Theta = \pi$. Next, Mary stays put and rests while *you* run a lap, so L rotates another π. Thus, after you and Mary have each run one lap, L has executed one full revolution: $\delta\Theta = 2\pi$.

(ii) Next, Mary runs another lap, but this time you don't wait till she finishes before you start to move. You start walking *very slowly* along the track behind her. You only cover a small distance before Mary has returned to the start line. Once she's there, you jog along the remainder of the path back to her. Explain why the net rotation of L must *still* be $\delta\Theta = 2\pi$.

(iii) Next, you decide to race Mary, despite the fact that she *always* wins. Sure enough, you never even get close. Explain why the net rotation of L after you and Mary have each

completed your lap is *still* $\delta\Theta = 2\pi$. (*Hint*: Imagine *gradually* increasing your speed in (ii), and picture the *continuous* evolution of the graph of $\Theta(t)$: (i) $\rightsquigarrow$ (ii) $\rightsquigarrow$ (iii).)

(iv) You decide to have a final race, and this time you give it all you've got, and you succeed in staying right on Mary's heels the whole time. As you stare at her back, eyes front along the path, your line of sight L is now the *tangent line* to C. And since the net rotation of L must *still* be $\delta\Theta = 2\pi$, Hopf's proof is complete!

Second (Angular Excess) Proof of GGB

24. There Are Only Five Regular Polyhedra. If each face of a regular polyhedron is a regular n-gon, and m such faces meet at each vertex, use the following steps to show that the *only* possibilities are the five Platonic solids illustrated in [18.2], namely, $(n, m) = (3, 3)$ or $(4, 3)$ or $(3, 4)$ or $(5, 3)$ or $(3, 5)$.

(i) Explain why $m \geqslant 3$.

(ii) If θ_n is the internal angle of the regular n-gon face, explain why $m\theta_n < 2\pi$. (*Hint*: Imagine cutting out the m faces that meet at a vertex v, now cut along one of the edges containing v, and finally flatten out the surface onto the plane.)

(iii) Write down θ_n for $n = 3, 4, 5, 6$ and exhaust the possibilities.

25. There Are Only Five *Topologically* Regular Polyhedra. Let us use the same (n, m) notation as in the previous exercise, only now the faces of our topological polyhedron can be curved, irregular, and/or with wavy edges. As usual, let V, E, and F denote the number of vertices, edges, and faces.

(i) Explain why $V = Fn/2$.

(ii) Explain why $E = Fn/m$.

(iii) By substituting the previous two equations into Euler's Polyhedral Formula, deduce that

$$F = \frac{4m}{2(m+n) - mn}.$$

(iv) Deduce that $\Omega(n, m) \equiv 2(m+n) - mn > 0$.

(v) Think of (n, m) as grid points in $\mathbb{R}^2$. In the relevant region ($n \geqslant 3$ and $m \geqslant 3$), mark each grid point with the value $\Omega(m, n)$, noting and using the symmetry about the line $n = m$.

(vi) Deduce that the only solutions (n, m) are the same ones we found in the previous exercise: the five Platonic solids, but now topologically deformed.

This result is due to Simon Antoine Jean L'Huilier (1811). See Robin Wilson in James 1999, page 516.

26. Legendre's Projection of Convex Polyhedra. As noted in the text, Legendre's proof of Euler's Polyhedral Formula assumed that the polyhedron was *convex*. The proof we presented did *not* require this assumption, but Legendre's assumption of convexity gave him an elegant advantage: it allowed him to instantly jump from the original polyhedron $\mathcal{P}$ to a topologically equivalent *geodesic* polygonal partition of the sphere (see [18.7]), as follows. Imagine that the convex polyhedron $\mathcal{P}$ is a wire frame, and a light bulb B is placed somewhere inside it. If $\mathcal{S}$ is a sphere centred at B and enclosing $\mathcal{P}$, prove that the *shadow* of the frame on the sphere is the desired *geodesic* partition.

27. Hopf's Proof That $\chi(\mathcal{S}_g) = 2 - 2g$. In the text, we proved $\chi(\mathcal{S}_g) = 2 - 2g$ by first showing that $\chi(S^2) = 2$, and then showing that gluing on a "Toblerone® handle" (see [18.9]) reduces χ by 2 (see [18.10]). Here is an elegant alternative argument, due to Hopf (1956, pp. 8–10).

(i) Glue together two Toblerone® handles to create a topological torus, $\mathcal{S}_1$, and deduce that $\chi(\mathcal{S}_1) = 0$.

(ii) Remove one triangle from this torus, and remove one triangle from a general surface $\mathcal{S}_g$ of genus g. Now glue the edges of the two triangular holes to each other, thereby creating a new closed surface $\mathcal{S}_{g+1}$ of one higher genus. Deduce that $\chi(\mathcal{S}_{g+1}) = \chi(\mathcal{S}_g) - 2$.

(iii) Use (i) and (ii) to deduce that $\chi(S^2) = 2$, and that $\chi(\mathcal{S}_g) = 2 - 2g$.

28. Dual Polyhedra. Place a vertex in the centre of each face of a cube, and join them together to create an octahedron; this is the ***dual*** of the cube, each face having become a vertex, and each vertex having become a face. What if we keep going, by taking the dual of the octahedron, i.e., the dual of the dual of the cube? With the assistance of [18.2], repeat this exercise for the three remaining Platonic solids.

Third (Vector Field) Proof of GGB

29. Dipole Streamlines Are Circular. Prove that the dipole streamlines in [19.3] are *circles*,

(i) by calculation

(ii) geometrically.

30. Honey-Flow on $\mathcal{S}_g$. Take $\mathcal{S}_g$ in [19.11] and rotate it by a right angle about an axis perpendicular to the page, so that the holes are aligned horizontally, instead of vertically. Sketch the new honey-flow, and deduce that there are a total of $6g - 2$ singular points. Confirm that the sum of their indices is indeed $\chi = 2 - 2g$, in accord with the Poincaré–Hopf Theorem.

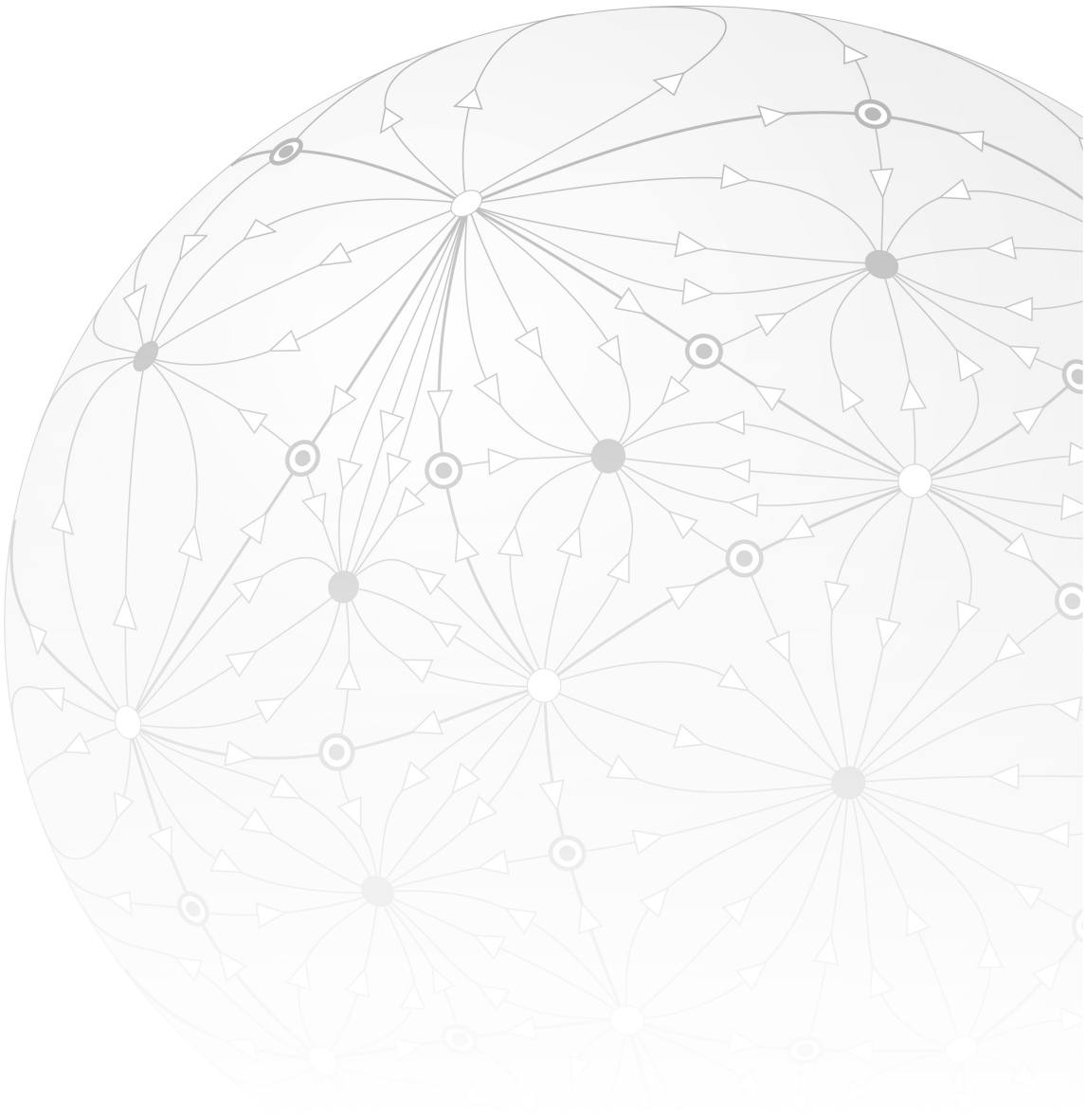
31. Existence of the Stiefel Vector Field. Explain, as explicitly as possible, *why* the construction in [19.12] is guaranteed to create a consistent vector field on $\mathcal{S}_g$, regardless of the specific partition into polygons.

32. A Vector Field on $\mathcal{S}_g$ with *One* Singular Point? The dipole field on the sphere [19.8b] has only one singular point, and its index is 2 ($= \chi$), in accord with the Poincaré–Hopf Theorem. On a torus ($g = 1$) we have seen we need not have any singular points in the flow. But if $g \geqslant 2$, there must be singular points, for the sum of their indices is $\chi \neq 0$. In this general case, is it always possible to generalize the dipole field on the sphere, to find a flow with only *one* singular point of index χ? (*Hint*: A new field can be created by coalescing two or more of its singular points into one. For example, we can imagine creating the dipole field in [19.8b] by coalescing the two vortices in [19.8a].)

33. Vector Fields on $\mathcal{S}_g$ with $-\chi$ Saddle Points. Show (by drawing it!) that it is possible to construct a vector field on $\mathcal{S}_g$ (with $g \geqslant 2$) such that there are precisely $(2g - 2) = -\chi$ singular points, each with $\mathcal{I} = -1$. (*Hints*: (A) First tackle $g = 2$; the generalization to higher genus turns out to be obvious. (B) Picture $\mathcal{S}_2$ as two bagels still joined together. Next, imagine this surface crushed almost flat under a great weight, so that it resembles two joined annuluses. Finally, confirm (by drawing it!) that it is possible to place one saddle point on the "top" surface, and one on the "bottom," *and* to have the top and bottom fields agree where they meet, at the edges.)

ACT IV

Parallel Transport



Chapter 21

An Historical Puzzle

Parallel transport[1] is now understood to be one of the most fundamental and powerful concepts of Differential Geometry, but it was remarkably late to the game.

By 1915 most of the fundamental concepts and computational tools of modern Differential Geometry in n-dimensional space were already in place, and not a moment too soon.

Einstein had spent the previous *decade* struggling to reconcile his 1905 Special Theory of Relativity with gravity. He had known since 1905 that no physical effect could travel faster than light, but Newton's Inverse-Square Law—unquestioned since 1687—insisted that if a giant solar flare erupted from the surface of the Sun, its tiny gravitational tug on the Earth would be felt *instantly*, despite the fact that the light from the flare would take eight minutes to reach Earth—a naked contradiction!

Only very gradually did Einstein's physical intuition drive him towards Differential Geometry, where he was miraculously blessed to find that in 1901 Gregorio Ricci (1853–1925) and Tullio Levi-Civita (1873–1941) had jointly published *"the marvelously prepared and almost predestined instrument for the exposition of his theory"*[2]—that instrument was ***Tensor Calculus***.[3]

On the 25th of November, 1915, Einstein finally ended his terrible, decade-long struggle, harnessing Tensor Calculus to write down his famous ***Field Equation*** that reconciled gravity with Special Relativity; he christened this union the ***General Theory of Relativity***. Einstein had succeeded in understanding the true nature of gravity: it is the ***Riemannian curvature***[4] impressed upon 4-dimensional spacetime by matter and energy; free-falling particles then respond to the gravitational field by travelling through curved spacetime along *geodesics*.[5]

As of the time of this writing, in 2020, every single testable prediction of the theory has been confirmed, including the astonishing, Nobel Prize–winning[6] confirmation—on the 14th of September, 2015—that *gravitational waves exist*, which Einstein had predicted almost exactly a century earlier, in 1916!

Little appreciated is the fact that in several instances these experimental confirmations have attained a breathtaking degree of accuracy, rivalling or exceeding those of the former gold standard: Quantum Electrodynamics. Indeed, our daily use of Global Positioning System (GPS) technology[7] would simply be impossible without taking into account the *precise* time-warping effect of gravity predicted by General Relativity! Einstein's 1915 discovery therefore stands as

---

[1] Sometimes instead referred to as "parallel displacement" or "parallel propagation."

[2] Levi-Civita (1931). For the history of the slow acceptance of Tensor Calculus before 1915, and the dramatic change post-1915, see Bottazzini (1999).

[3] This purely mathematical discovery was originally called the ***Absolute Differential Calculus***, but later became known as ***Ricci Calculus***, and finally as ***Tensor Calculus***. Élie Cartan's more powerful and elegant version of Tensor Calculus constitutes the dénouement of our drama (Act V) and is called the ***Exterior Calculus of Differential Forms***.

[4] This intrinsic measure of curvature will be described at the conclusion of Act IV: it generalizes Gaussian curvature to n-dimensional spaces.

[5] Though geodesics now *maximize* the "distance" measured using the metric of curved spacetime, which generalizes the Minkowski spacetime interval $\beth$; see (6.15), page 75.

[6] The 2017 Nobel Prize for Physics was awarded to Rainer Weiss, Barry C. Barish, and Kip S. Thorne for their joint construction of the Laser Interferometer Gravitational-Wave Observatory (LIGO) detector, which made the discovery possible.

[7] See Taylor and Wheeler (2000, A-1).

not only a supremely beautiful scientific triumph of the human intellect, but also as one of the *best-tested* physical theories we have.

But there is a serious puzzle[8] here, one that is not widely known or recognized. Einstein's success was all the more remarkable, and remains all the more *puzzling*, because he achieved it *before* Levi-Civita—pictured in [21.1]—discovered[9] the concept of ***parallel transport***, which did not occur until *1917!* Without this concept, I personally do not know of any way to make complete geometrical sense of Einstein's 1915 discovery.

So what *is* parallel transport? Like the concept of the geodesic, it is a queer amphibian, equally at home in the world of extrinsic geometry and in the world of intrinsic geometry, and we shall now describe what it looks like in each of these worlds, and attempt to understand how it is able to straddle the two.

[21.1] *Tullio Levi-Civita (1873–1941)*

[8]For readers familiar with General Relativity, there is in fact a second, even more spectacular puzzle: in November 1915, *Einstein did not know the Differential Bianchi Identity*—stated later as (29.17)—and therefore he did not realize that energy–momentum conservation followed automatically from his law! See Pais (1982, p. 256).

[9]Levi-Civita (1917).

Chapter 22

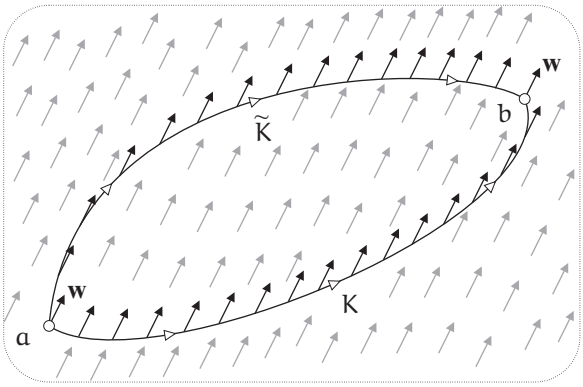
Extrinsic Constructions

22.1 Project into the Surface as You Go!

Here is the heuristic idea of what Levi-Civita sought to achieve in his fundamental paper of 1917. To ***parallel transport*** a tangent vector $\mathbf{w}$ to a surface $\mathcal{S}$ along a curve K connecting two points a to b within $\mathcal{S}$, we want to always keep $\mathbf{w}$ pointing in the same direction (and having the same length) *while remaining tangent to* $\mathcal{S}$ as it moves along K, the direction of $\mathbf{w}$ at each moment always being parallel to its direction a moment before.

But this is seemingly *impossible!* If we move even a small distance ϵ from p to q along K, then the vector $\mathbf{w}$ in the tangent plane T_p, when rigidly moved parallel to itself (in $\mathbb{R}^3$) to q, will generally *not* lie in T_q, but will instead stick out of $\mathcal{S}$.

In the Euclidean plane no such problem arises, and there is a simple *global* sense of parallelism, resulting from the fact that through any point not lying on a line L (in the direction $\mathbf{w}$), there is a unique line that is parallel to L, in the same direction $\mathbf{w}$. This global parallelism can be pictured as a uniform flow across the plane, with velocity $\mathbf{w}$; see [22.1]. The parallel-transported vector along K is simply the restriction to K of this constant velocity field. It is therefore trivially clear that we always obtain the *same* parallel-transported vector $\mathbf{w}$ at b, regardless of whether we get there by travelling along K, or along some other route $\widetilde{K}$.



[22.1] *In the Euclidean plane there is a* global *concept of parallelism, manifested by the flow with constant velocity vector* $\mathbf{w}$. *To parallel transport* $\mathbf{w}$ *along* K, *we need only restrict this vector field to* K. *Thus we always obtain the* same *vector* $\mathbf{w}$ *at* b, *whether we travel along* K *or* $\widetilde{K}$. *This path-independence holds* only *in this case of vanishing curvature.*

Levi-Civita's key observation was that this path-independence of the parallel-transported vector *only* occurs if the space is flat; if the space is curved then parallel transport (as it will be defined momentarily) along two different paths will yield *different* final vectors. This crucial phenomenon is called ***holonomy***, and is the subject of an upcoming chapter by that name. As we shall see, this holonomy can be used to *measure* curvature, and not only the Gaussian curvature of surfaces, but also the Riemannian curvature of higher-dimensional curved spaces, like Einstein's curved, 4-dimensional spacetime.

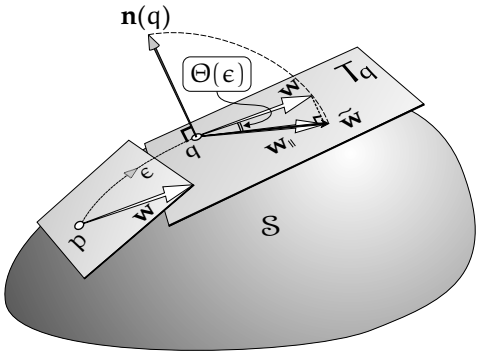
If we try to repeat the Euclidean construction in the hyperbolic plane, $\mathbb{H}^2$, then we immediately run into trouble, for if we are given a point and a line L though it, we know that there are now *infinitely many* "parallel" lines running through a neighbouring point; which one should we choose?!

Later we shall return to this topic of how to carry out parallel transport in $\mathbb{H}^2$, or equivalently on the pseudosphere, but let us now immediately jump in at the deep end and see how to do parallel transport on a *general* curved surface.

Our first extrinsic method of parallel transport is illustrated in [22.2], using the notation that was introduced in the opening paragraph, above. If we move the tangent vector $\mathbf{w}$ at p parallel to itself (as a vector in $\mathbb{R}^3$) a small (ultimately vanishing) distance ϵ along the curve K, then $\mathbf{w}$ is no longer tangent to S at q: it sticks out of the tangent plane T_q, but only slightly, at angle $\Theta(\epsilon)$. The best approximation to $\mathbf{w}$ that *is* tangent to S at q is $\mathbf{w}_{||}$, obtained by projecting $\mathbf{w}$ orthogonally down onto T_q. If $\mathcal{P}$ denotes this projection, and $\mathbf{n}(q)$ is the unit normal at q, then, as explained by [22.2],

$$\mathbf{w}_{||} = \mathcal{P}[\mathbf{w}] = \mathbf{w} - (\mathbf{w} \bullet \mathbf{n})\mathbf{n}. \tag{22.1}$$

To parallel transport $\mathbf{w}$ along K, we must imagine breaking K into an enormous number oftiny steps of length ϵ, repeating this process of translating in $\mathbb{R}^3$ and projecting back into the surface, over and over, one ϵ-step at a time, until we finally arrive at the end of K. Even then, this is merely an *approximation* of parallel transport—perfect parallel transport is only achieved when we take the *limit* that ϵ vanishes, in which case we *continually project into the surface as we go*.



[22.2] *To* **parallel transport** *the tangent vector* $\mathbf{w}$ *from* p *to the neighbouring point* q *a distance* ϵ *away, we move it parallel to itself in* $\mathbb{R}^3$*, then either* (i) project *it down onto* T_q *to get* $\mathbf{w}_{||}$ *or* (ii) rotate *it down to get* $\widetilde{\mathbf{w}}$*. The two constructions are equivalent, because we ultimately take the limit that* ϵ *vanishes.*

Note that the projection $\mathbf{w}_{||}$ in [22.2] is slightly shorter than the original vector $\mathbf{w}$. If we instead simply *rotate* $\mathbf{w}$ straight down onto T_q then we obtain $\widetilde{\mathbf{w}}$, pointing in the same direction as $\mathbf{w}_{||}$ but with $|\widetilde{\mathbf{w}}| = |\mathbf{w}|$. (The length $|\mathbf{w}|$ is irrelevant to the construction, but here we have chosen it have the same unit length as $\mathbf{n}$, for ease of visualization.) But, since $\lim_{\epsilon \to 0} \Theta(\epsilon) = 0$, we see that

$$\mathbf{w}_{||} \asymp \widetilde{\mathbf{w}}.$$

Finally, because perfect parallel transport takes this limit of vanishing ϵ, we deduce that *parallel transport preserves length*.

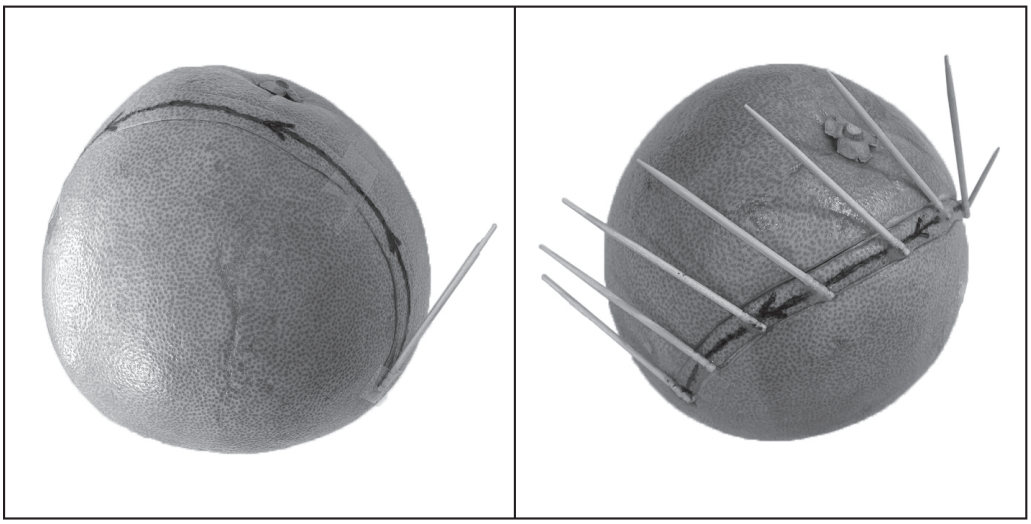
We have thus arrived at two *equivalent* extrinsic methods of parallel transport:

To ***parallel transport*** *the tangent vector* $\mathbf{w}$ *to* S *along a curve* K*, move* $\mathbf{w}$ *parallel to itself in* $\mathbb{R}^3$ *while continually either* (i) projecting *into* S *as you go or* (ii) rotating *down to* S *as you go. The length of the vector remains constant as it is parallel transported.* (22.2)

Method (ii) is much easier in practice, and we strongly encourage you to try it out yourself on any curved object you have at hand—fruits and vegetables are convenient, and they have a crucial advantage that will soon become apparent: they can be peeled!

Draw any curve K you please on your surface, connecting point a to point b, press one end of a toothpick down against the surface at a, in any tangent direction you like. Now move it parallel to itself (in space) a small distance along K, then press that same end straight down against the surface (so that it becomes tangent again), ... repeat, ... repeat, ... repeat, ... until you arrive at b!

Figure [22.3] illustrates this on a pomelo. Although we may choose any direction at a that is tangent to S, here we have chosen $\mathbf{w}$ to be the initial velocity along K, in order to make this point:



[22.3] *Parallel transport on a pomelo: continually press the toothpick straight down onto the surface while moving the toothpick parallel to itself along the curve.*

in general, $\mathbf{w}_{||}$ *does not remain tangent* to K, though it certainly *does* remain tangent to the surface, by construction.

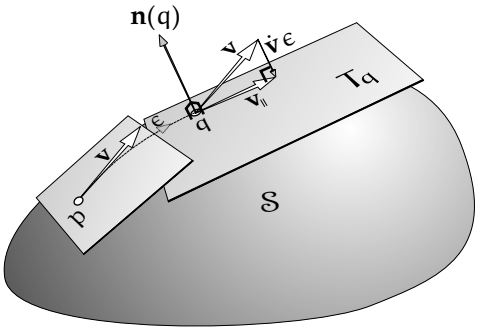
Here we have chosen the path K so as to be close to a vertical section of the surface of the pomelo, and in the photograph on the right we have deliberately taken the picture from almost directly overhead, so that it *looks* remarkably straight. It is therefore perhaps very surprising and confusing how fast the toothpick rotates from initially being tangent to K, to being almost perpendicular to K by the end. This because our photograph does not reveal just how *curved* K is respect to the *intrinsic* geometry. The intrinsic, geodesic curvature of this particular path K will be revealed in Section 22.3, where a different method of parallel transport will be revealed.

22.2 Geodesics and Parallel Transport

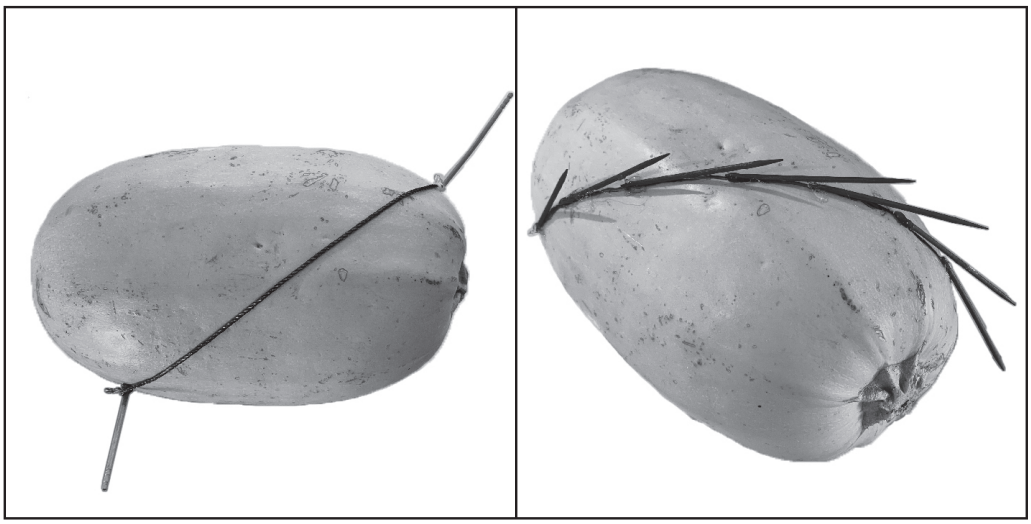
Under what circumstances *would* parallel transport of the initial tangent vector along a curve G *maintain* tangency to G? The answer is that this happens precisely when G is a *geodesic*!

To see why, recall that the local definition of a geodesic G is as a unit-speed curve for which the acceleration is always directed along the surface normal, $\mathbf{n}$. That means that if we look at the velocity at neighbouring points p and q, distance/time ϵ apart, then the difference of the two velocities is ultimately $\dot{\mathbf{v}}\epsilon$ and is ultimately directed along $\mathbf{n}$. Figure [22.4] depicts this new, special case of [22.2], now with $\mathbf{v}$ as the velocity, pointing along the segment ϵ. We see that this means that the new velocity at q is ultimately the original velocity $\mathbf{v}$ at p, *parallel-transported along itself.*

Conversely, suppose we are given a point p and that we launch a particle from there, out



[22.4] Constructing a Geodesic via Parallel Transport. *Launch a particle from* p *with velocity* $\mathbf{v}$*, and define* G *to be the curve obtained by parallel-transporting* $\mathbf{v}$ along itself*, so that the new velocity at the neighbouring point* q *(distance/time* ϵ *away) is* $\mathbf{v}_{||}$*. Then the acceleration* $\dot{\mathbf{v}}$ *is given by* $\dot{\mathbf{v}}\epsilon \asymp \mathbf{v}_{||} - \mathbf{v}$*. But then* $\dot{\mathbf{v}} \propto \mathbf{n}$*, and therefore* G *is geodesic.*



[22.5] [a] *A geodesic segment* G *on a yellow squash is first constructed by stretching a string across its surface.* **[b]** *The string is removed, but a toothpick is left at the initial point, tangent to* G*. As predicted, parallel transporting the initial toothpick along itself yields toothpicks that are everywhere tangent to the* same geodesic *curve* G*.*

across the surface, in an arbitrary direction $\mathbf{v}$ within $\mathcal{S}$. The same geometry as before tells us that we may *construct* the geodesic motion that ensues by performing parallel transport: carry $\mathbf{v}$ a short distance ϵ within $\mathcal{S}$ along itself, then press it down onto the surface to obtain the new velocity there . . . repeat!

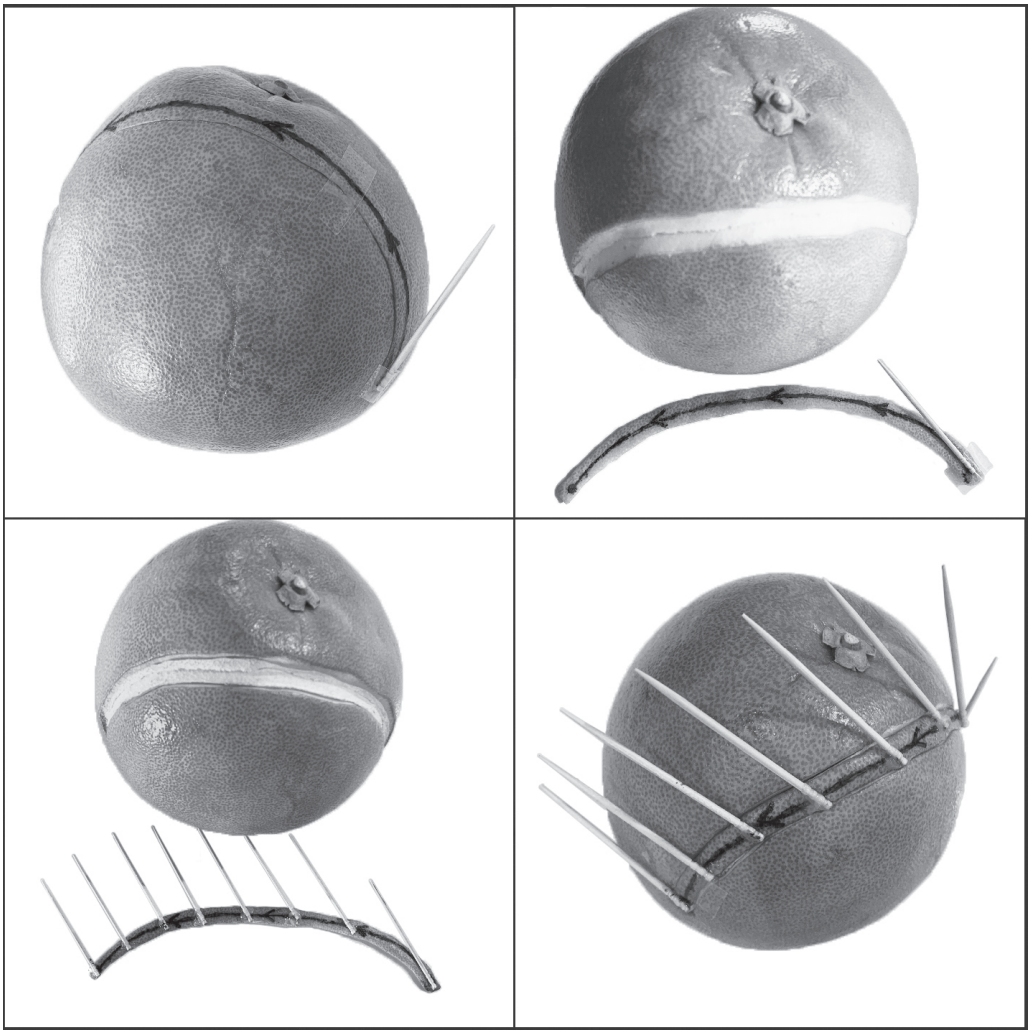
Figure [22.5] illustrates this new method of constructing geodesics. In order to check the correctness of our ultimate solution, we first construct a geodesic segment G on our surface (a squash) using our original trick of stretching a string across it, here tied to two (normal) toothpicks at the ends. If we remove the string, but leave behind an initial toothpick tangent to the start of the string, we can then take this initial-velocity toothpick and parallel transport it along itself. As you see, and as we urge you to try for yourself, this construction yields the *same* geodesic segment G as the string did.

22.3 Potato-Peeler Transport

We will now describe a third extrinsic method of parallel transport that will quickly prove its mettle,[1] yielding important *properties* of parallel transport that are not immediately evident from our first two constructions seen in (22.2).

In [22.2] imagine that we use a potato peeler to peel off a narrow strip of $\mathcal{S}$ surrounding the movement ϵ along K. Now, instead of rotating $\mathbf{v}$ down to the surface, imagine that we take this short strip of peel and bend it up, pressing it flat onto T_p, bringing $\mathbf{v}_{||}$ to $\mathbf{v}$, instead of the other way around. If we keep peeling a narrow strip all along K, we arrive at our new construction:

[1] This new construction is not readily found in modern textbooks, and I recall being elated when I first hit upon it, more than 30 years ago. Other authors likewise believed they had found something new: e.g., Koenderink (1990), Casey (1996), and Henderson (1998). But when the time came to write this book, I turned to original sources, and there I found that what we had *re*discovered was *first* discovered, more than a century ago, by Levi-Civita himself! To read Levi-Civita's own account of his idea, translated into English, see Levi-Civita (1926, p. 102). For more on the history of the discovery, see Goodstein (2018, Ch. 12).

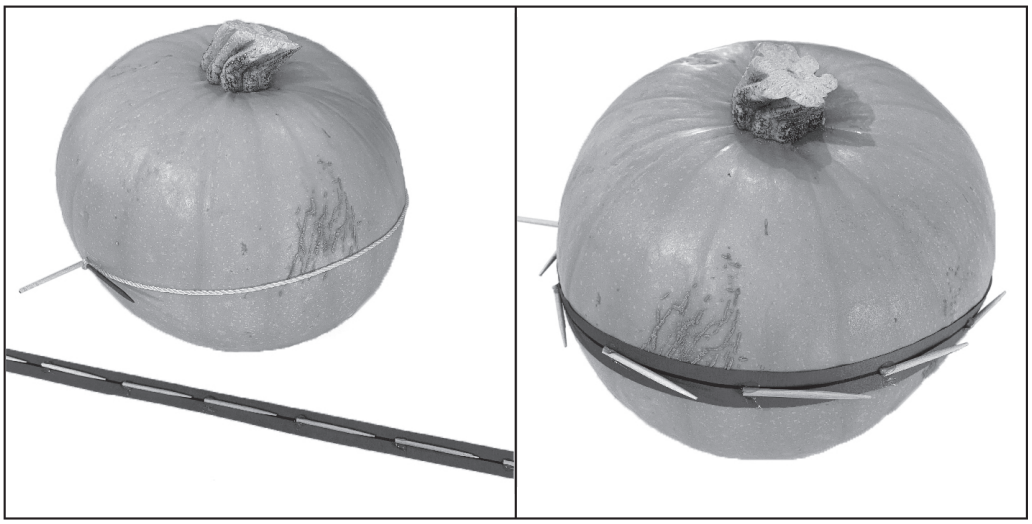


[22.6] "Potato-Peeler" Parallel Transport *on a pomelo: remove a narrow strip of peel along the curve, flatten it out on the table, do Euclidean parallel transport within the flat tabletop, then finally reattach the peel and vectors to the surface.*

> **Potato-Peeler Parallel Transport.** *To* **parallel transport** *the tangent vector* $\mathbf{v}$ *to* $\mathbb{S}$ *along a curve* K*, peel a narrow (ultimately vanishingly thin) strip of* $\mathbb{S}$ *containing* K *as its centre line. Lay this strip flat in the plane, then do ordinary Euclidean parallel transport of* $\mathbf{v}$ *along the flattened* K*. Finally, reattach* K *(and the constructed vectors) to* $\mathbb{S}$ *in its original location.* (22.3)

Figure [22.6] reveals how we cheated and actually used *this* method (rather than those in (22.2)) to carry out the parallel transport previously shown in [22.3].

While a potato peeler conjures up the right mental image, it is *not* in fact the best tool for the job. To obtain a tidy, *narrow* strip, take a small sharp knife and make a very shallow incision along a curve just to one side of K, cutting beneath K; now do the same from the other side, thereby cutting out the desired narrow strip of peel, with a shallow V-shaped cross-section.



[22.7] [a] *A geodesic segment* G *on a pumpkin is first constructed by stretching a string across its surface. Meanwhile, on the tabletop, a toothpick pointing along a straight strip of tape is parallel transported along that strip, automatically remaining tangent to the centre line.* **[b]** *The string is removed, and the tape (with attached toothpicks) is unrolled onto the surface, starting at the same point and heading off in the same direction. As predicted, the tape automagically rolls down onto the surface along the* same geodesic *curve* G*, and the toothpicks are parallel-transported along it.*

Two important properties of parallel transport become immediately clear from the new construction.

First, if two vectors are parallel transported along a curve in the Euclidean plane, then clearly the angle between them remains constant. The new construction immediately implies that the same must therefore be true on a curved surface:

> *If two vectors are parallel transported along a curve on a curved surface, then the angle between them remains constant.* (22.4)

Conversely, if only one of the vectors is known to be parallel-transported, and the angle with the second is kept fixed, then that second vector is necessarily parallel-transported, also.

Second, recall from (1.6), page 13 and from (1.7), page 14, that if a *geodesic* is peeled from the surface and laid flat on the table, it becomes a *straight line*; conversely, if a straight and narrow strip of sticky tape is gradually rolled onto the surface, it will automatically generate a geodesic. In the Euclidean plane, the direction vector of a line continues to point along the line as it is parallel-transported along the line. We thus have a much more intuitive, immediate, *visual* proof the previously established fact that,

> *If* G *is a geodesic traced at unit speed, launched with initial velocity* **v***, then the velocity at any future time is obtained by parallel-transporting the initial velocity along* G*. Conversely, given an arbitrary initial point, and an arbitrary initial velocity* **v** *tangent to* $\mathbb{S}$*, parallel-transport of* **v** *along* **v** *yields the unique geodesic arising from these initial conditions.* (22.5)

This is illustrated in [22.7]. Again, we urge you to try this for yourself.[2]

Figure [22.8] simply illustrates the same idea again. The strip form [22.7] has been removed from the pumpkin and rolled down onto the surface of the yellow squash shown in [22.5], along the same geodesic as before. But whereas in that figure we were forced to manually carry out the parallel transport and associated generation of the illustrated geodesic, *this* time both things happen automatically.

[22.8] *The same strip shown in [22.7] is here rolled down onto the surface shown in [22.5]. Unlike the manual construction in [22.5], here both the generation of the geodesic and the parallel transport of its velocity vector along it are* automatic.

[2]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. As for the toothpicks, we used a hot glue gun to attach them, but just at their bases, so that they were free to become tangent to the surface once the strip was attached

Chapter 23

Intrinsic Constructions

23.1 Parallel Transport via Geodesics

In the Euclidean plane, here is an easy, obvious way to parallel-transport a vector $\mathbf{w}$ along a straight line L (with direction $\mathbf{v}$): *keep the angle between* $\mathbf{w}_{||}$ *and* $\mathbf{v}$ *constant as it moves along* L. See [23.1a].

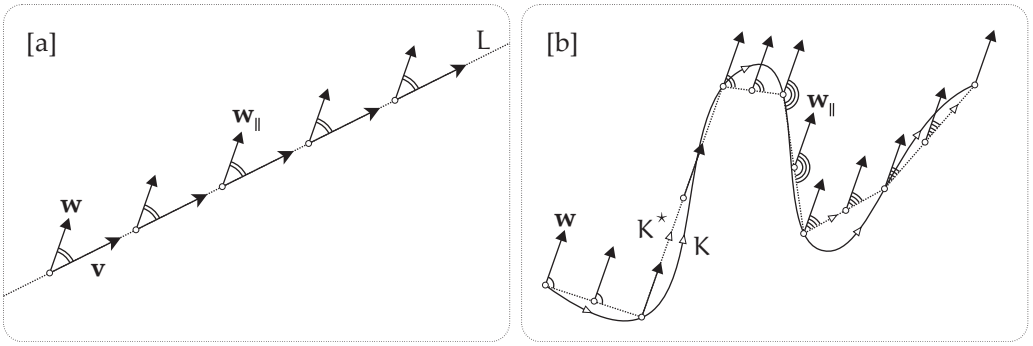
Well, on a curved surface the analogue of a line L is a geodesic G, so it is natural to guess that the Euclidean construction generalizes as follows:

> *To parallel transport a tangent vector* $\mathbf{w}$ *to a surface* $\mathcal{S}$ *along a geodesic* G *with velocity* $\mathbf{v}$*, simply keep the angle between* $\mathbf{w}_{||}$ *and* $\mathbf{v}$ constant *as it moves along* G. (23.1)

This *intrinsically* defined method does indeed produce the *same* parallel-transported vector as the three extrinsic methods above, as follows immediately [exercise] by combining (22.4) and (22.5). This is illustrated in [23.2].

Let us repeat our mantra: please try this construction for yourself! Here we have used a stretched string to emphasize the intrinsic nature of this construction, but in practice it is often easier to construct the geodesic with narrow-sticky-tape construction, for the tape will continue to roll down as a geodesic even in parts of the surface that bend towards you, across which a string cannot be stretched, at least not from the outside of the surface.

To (intrinsically) parallel transport $\mathbf{w}$ along an arbitrary curve K in the Euclidean plane, the first step is to approximate K by a sequence $K^\star$ of short, straight line segments, $\{L_i\}$, all of which have length less than ϵ, say. See [23.1b]. Keeping the angle constant along each successive L_i, we have parallel transported $\mathbf{w}$ along $K^\star$. Finally, taking the limit that ϵ vanishes, $K^\star$ becomes K, and we have achieved intrinsic parallel transport along K.



[23.1] [a] *In the Euclidean plane, to parallel transport* $\mathbf{w}$ *along the line* L *in the direction* $\mathbf{v}$*, keep the angle between* $\mathbf{w}_{||}$ *and* $\mathbf{v}$ constant. **[b]** *To parallel transport along a general curve* K*, approximate it with a sequence of line segments,* $K^\star$*, and keep the angle constant with each one, in succession. Finally, let the length of each segment go to zero, so that* $K^\star$ *becomes* K.

Of course this is all somewhat theatrical in the Euclidean plane, where we have a global concept of parellelism, but the point is that it is now clear what we must do on a curved surface:

> *To parallel transport a tangent vector* $\mathbf{w}$ *to a surface* $\mathcal{S}$ *along a general curve* K, *approximate* K *by a sequence* $K^\star$ *of geodesic segments* $\{G_i\}$, *each of length less than* ϵ, *then carry* $\mathbf{w}$ *along each* G_i, *maintaining a* constant angle *between the direction of* G_i *and* $\mathbf{w}_{||}$. *Finally, take the limit that* ϵ *vanishes, so that* $K^\star$ *becomes* K.

23.2 The Intrinsic (aka, "Covariant") Derivative

When we parallel-transport a vector along a curve K in a surface $\mathcal{S}$, it appears to be *unchanging* to an inhabitant of $\mathcal{S}$ who walks along K watching it. This idea of constancy opens the door to an *intrinsic* measure of how fast a vector *changes* as it moves along a curve.

Suppose a particle moves at unit speed along a curve K on a surface $\mathcal{S}$, its position at time t being $p(t)$, and its unit velocity being $\mathbf{v}(t)$. Furthermore, suppose that emanating from $p(t)$ we have a vector $\mathbf{w}(t)$ that is tangent to $\mathcal{S}$, so that $\mathbf{w}$ is defined everywhere along K (though not necessarily anywhere else). What is the *intrinsic* rate of change of $\mathbf{w}(t)$?

[23.2] *A geodesic segment* G *on a pumpkin is first constructed by stretching a string across its surface. Any initial tangent vector to the surface can then be parallel transported along* G *simply by keeping the angle between it and* G constant.

Well, if p moves for a short time ϵ to $q = p(t+\epsilon)$, then the *extrinsically* defined rate of change $\nabla_{\mathbf{v}}\mathbf{w}$ is given by

$$\epsilon\nabla_{\mathbf{v}}\mathbf{w} = \epsilon\mathbf{w}'(t) \asymp \mathbf{w}(q) - \mathbf{w}(p),$$

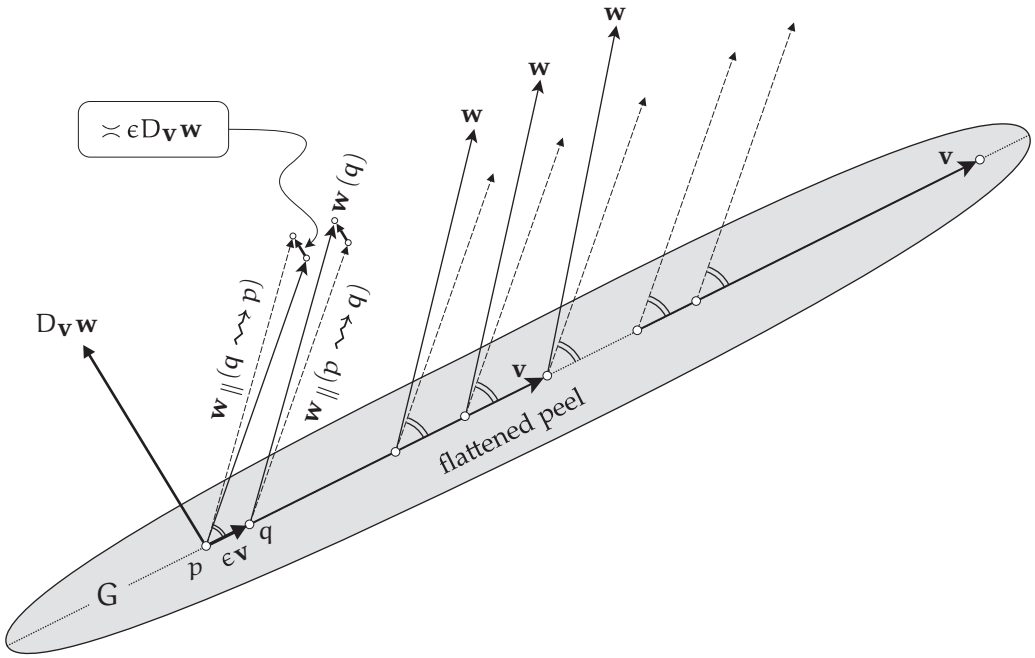
where the difference of the two vectors on the right makes perfectly good sense in $\mathbb{R}^3$. But $\mathbf{w}(q)$ lives in T_q, while $\mathbf{w}(p)$ lives in T_p, so this difference lives in neither, and so is certainly *not* intrinsic to $\mathcal{S}$.

It is precisely the parallel-transport depicted in [22.2] that allows us to bring $\mathbf{w}(p)$ to q as $\mathbf{w}_{||}(p \rightsquigarrow q)$, keeping it "constant." We can now compare this old value $\mathbf{w}_{||}(p \rightsquigarrow q)$ to the new value $\mathbf{w}(q)$, to see how much it has changed. Let $D_{\mathbf{v}}$ denote this *intrinsic* rate of change (operator) along K, which we shall call the ***intrinsic derivative***, then

$$\epsilon D_{\mathbf{v}}\mathbf{w} \asymp \mathbf{w}(q) - \mathbf{w}_{||}(p \rightsquigarrow q).$$

Since both these vectors lie in T_q, so does their difference. Since $D_{\mathbf{v}}\mathbf{w}$ is tangent to $\mathcal{S}$, it may be viewed as intrinsic to $\mathcal{S}$.

WARNING: The standard name for $D_{\mathbf{v}}$ is the ***covariant derivative***. Historically, the word "covariant" had to do with how the coordinate expression of $D_{\mathbf{v}}$ transforms under a change of coordinates, but since "covariant" means little or nothing to the ears of modern students, we would argue (by example!) that the time has come for a new, *self-explanatory* name. That said, we



[23.3] *The* **Intrinsic (aka "Covariant") Derivative**, $D_{\mathbf{V}}\mathbf{w}$, *which measures the intrinsic rate of change of* $\mathbf{w}$ *along* $\mathbf{v}$, *is especially easy to visualize if* $\mathbf{v}$ *is the velocity of a* geodesic G, *for then the strip of peel surrounding* G *flattens into a straight line, and parallel transport of* $\mathbf{w}_{||}$ *within the surface becomes ordinary Euclidean parallel transport within the plane.*

would be remiss if we did not warn the reader that (as of this writing in 2019) *essentially every other Differential Geometry (and physics) book instead refers to* $D_{\mathbf{V}}$ *as the* **covariant derivative**. *Finally, we note that this is also called the* **Levi-Civita Connection**.

In order to give a vivid picture of this intrinsic derivative, suppose for simplicity's sake that the curve is a *geodesic*, G. If we peel the strip surrounding G and press it onto the table, it becomes the straight line illustrated in [23.3], and both the tangent vectors $\mathbf{v}(t)$ to G, and the tangent vector field $\mathbf{w}(t)$ to $\mathcal{S}$ (defined along G) all get pressed down onto this table top, $\mathbb{R}^2$, just as in [22.6] and [22.7]. Parallel transport of $\mathbf{w}$ along G (as in [23.2]) now amounts to ordinary Euclidean parallel transport along this line in $\mathbb{R}^2$, keeping the angle between $\mathbf{w}_{||}$ and $\mathbf{v}$ constant, as in [23.1a].

As is evident in the figure, here $\mathbf{w}$ is growing in length and rotating counterclockwise as it moves along G. The job of $D_{\mathbf{V}}\mathbf{w}$ is to *quantify* the rate at which this is happening, i.e., the rate at which $\mathbf{w}(t)$ departs from its initial value (as represented by $\mathbf{w}_{||}$) just as we *begin* to move away from p along G.

Alternatively, and this is in fact the more standard definition, we may parallel transport $\mathbf{w}(q)$ *back* to p to obtain $\mathbf{w}_{||}(q \rightsquigarrow p)$, and then compare it to the original vector $\mathbf{w}(p)$, as illustrated in [23.3]. Thus the ***intrinsic derivative*** $D_{\mathbf{V}}\mathbf{w}$ at p may be pictured as emanating from p, lying with T_p, given by

$$D_{\mathbf{V}}\mathbf{w} \asymp \frac{\mathbf{w}_{||}(q \rightsquigarrow p) - \mathbf{w}(p)}{\epsilon}. \tag{23.2}$$

Figure [23.3] illustrates this geometric construction with $\epsilon = 0.1$: we draw the change in $\mathbf{w}$ resulting from moving one tenth of $\mathbf{v}$, then stretch this change vector by a factor of ten. Of course the *exact* value of $D_{\mathbf{V}}\mathbf{w}$ is obtained only when we take the limit of this figure as ϵ vanishes.

In the general case, where K is not geodesic, everything we have just visualized remains the same, provided that we only peel off the part of the surface surrounding the very short segment of K connecting p and q.

Note that the intrinsic constancy of $\mathbf{w}_{||}$ along K can now be neatly expressed by saying that its intrinsic derivative vanishes:

$$\boxed{\mathrm{D}_{\mathbf{v}}\mathbf{w}_{||}=0 \quad \Longleftrightarrow \quad \mathbf{w}_{||} \text{ is parallel-transported along } \mathbf{v}.}$$

Here is an *extrinsic* way of looking at the intrinsic derivative. Instead of thinking of $\mathrm{D}_{\mathbf{v}}$ as measuring the rate of change of the projection onto the tangent plane, we may instead take the projection of the rate of change $\nabla_{\mathbf{v}}$ itself. Once again letting $\mathcal{P}$ denote orthogonal projection onto the tangent plane, this generalization of (22.1) takes the form

$$\mathrm{D}_{\mathbf{v}}\mathbf{w}=\mathcal{P}\left[\nabla_{\mathbf{v}}\mathbf{w}\right]=\nabla_{\mathbf{v}}\mathbf{w}-\left(\mathbf{n}\bullet\nabla_{\mathbf{v}}\mathbf{w}\right)\mathbf{n}. \tag{23.3}$$

Note that the extrinsic term may also be expressed in terms of the Shape Operator:

$$\boxed{\mathrm{D}_{\mathbf{v}}\mathbf{w}=\nabla_{\mathbf{v}}\mathbf{w}+\left[\mathbf{w}\bullet S(\mathbf{v})\right]\mathbf{n}.}$$

In other words, to obtain $\mathrm{D}_{\mathbf{v}}\mathbf{w}$ we take the full rate of change $\nabla_{\mathbf{v}}\mathbf{w}$ in $\mathbb{R}^3$, then subtract out the part that is not tangent to the surface, thereby leaving behind the part that *is* intrinsic to the surface.

The formula (23.3) may be used to prove [exercise] that the intrinsic derivative $\mathrm{D}_{\mathbf{v}}$ obeys all the same linearity and Leibniz (product) rules as the full vector derivative $\nabla_{\mathbf{v}}$, so that if a and b are constants, and f and g are scalar functions defined along K, and $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ are tangent to $\mathcal{S}$ along K, then

$$\begin{aligned}
\mathrm{D}_{\mathbf{v}}[a\mathbf{x}+b\mathbf{y}] &= a\mathrm{D}_{\mathbf{v}}\mathbf{x}+b\mathrm{D}_{\mathbf{v}}\mathbf{y},\\
\mathrm{D}_{[f\mathbf{x}+g\mathbf{y}]}\,\mathbf{z} &= f\mathrm{D}_{\mathbf{x}}\mathbf{z}+g\mathrm{D}_{\mathbf{y}}\mathbf{z},\\
\mathrm{D}_{\mathbf{v}}[f\mathbf{x}] &= f\,\mathrm{D}_{\mathbf{v}}\mathbf{x}+[\mathrm{D}_{\mathbf{v}}f]\,\mathbf{x}=f\,\mathrm{D}_{\mathbf{v}}\mathbf{x}+f'\mathbf{x},\\
\mathrm{D}_{\mathbf{v}}[\mathbf{x}\bullet\mathbf{y}] &= [\mathrm{D}_{\mathbf{v}}\mathbf{x}]\bullet\mathbf{y}+\mathbf{x}\bullet[\mathrm{D}_{\mathbf{v}}\mathbf{y}]\,.
\end{aligned}$$

(Here the prime denotes differentiation with respect to either time or distance along K.) However, instead of establishing these facts by calculation, it is much simpler to think of flattening onto the tabletop the strip surrounding K, together with the vector fields $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, for then $\mathrm{D}_{\mathbf{v}}$ simply *is* $\nabla_{\mathbf{v}}$.

The intrinisic derivative sheds new light on our earlier discussion of *geodesic curvature*. Recall from [11.2], page 116, that the full acceleration $\boldsymbol{\kappa}\equiv\nabla_{\mathbf{v}}\mathbf{v}$ of a particle travelling at unit speed over the surface can be broken down into two components:

$$\nabla_{\mathbf{v}}\mathbf{v}=\boldsymbol{\kappa}=\boldsymbol{\kappa}_g+\boldsymbol{\kappa}_n.$$

The first component $\boldsymbol{\kappa}_g$ is the ***geodesic curvature vector***: it is the component of the acceleration tangent to $\mathcal{S}$. It is automatically perpendicular to the trajectory, and it points towards the centre of curvature *as perceived by inhabitants of* $\mathcal{S}$, the magnitude $|\boldsymbol{\kappa}_g|$ being the curvature of this circle; in other words, it measures the part of the curvature of the trajectory that is intrinsic to $\mathcal{S}$. In contrast to this, the ***normal curvature vector*** $\boldsymbol{\kappa}_n$ is directed along $\mathbf{n}$ and is invisible to these inhabitants.

Intuitively, $\boldsymbol{\kappa}_g$ should be the intrinsic rate of rotation of $\mathbf{v}$ within the surface:

$$\boldsymbol{\kappa}_g = D_{\mathbf{v}}\mathbf{v}.$$

That this is indeed true follows immediately from (23.3), since $\boldsymbol{\kappa}_n = (\mathbf{n} \bullet \nabla_{\mathbf{v}}\mathbf{v})\,\mathbf{n}$.

If $\theta_{||}$ denotes the angle between $\mathbf{v}$ and any vector $\mathbf{w}_{||}$ that it parallel transported along the trajectory, then (see Ex. 4) *the geodesic curvature is simply the rate of turning of the velocity vector within* $\mathcal{S}$*, relative to the "constant" vector* $\mathbf{w}_{||}$:

$$|\boldsymbol{\kappa}_g| = |D_{\mathbf{v}}\theta_{||}| = |\theta'_{||}|. \tag{23.4}$$

A geodesic is a curve that seems *straight* to the inhabitants of $\mathcal{S}$: in other words, $\boldsymbol{\kappa}_g = 0$. Thus the ***geodesic equation*** takes the form,

$$\boldsymbol{\kappa}_g = 0 \quad \Longleftrightarrow \quad D_{\mathbf{v}}\mathbf{v} = \mathbf{0}, \tag{23.5}$$

which is simply another way of looking at the fact that $\mathbf{v}$ is parallel transported along itself, and appears *constant* to the inhabitants of $\mathcal{S}$.

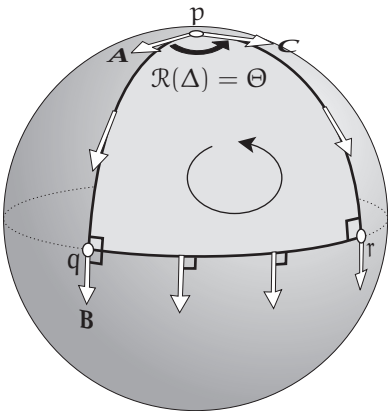
On the other hand, if a curve within $\mathcal{S}$ is *not* geodesic, then (see Ex. 4) when a strip surrounding it is peeled off and pressed flat onto the plane, it will appear *curved*, and its curvature within the plane is none other than κ_g!

Chapter 24

Holonomy

24.1 Example: The Sphere

Consider [24.1], which depicts a geodesic triangle Δ on the sphere of radius R, the angle between the meridians being Θ, and the third side being a segment of the equator.



[24.1] *When* **A** *is parallel transported round the geodesic triangle* Δ*, it returns to* p *rotated by the* **holonomy** *of* Δ*, namely,* $\mathcal{R}(\Delta) = \Theta$.

Suppose we start with the south-pointing vector **B** at q on the equator, and then use (23.1) to parallel transport it to p, along two different geodesic routes. If we carry it due north along the meridian geodesic qp, we obtain **A**. But if we instead parallel transport it due east along the geodesic equator segment qr (maintaining the right angle with the geodesic), and then carry it due north along the meridian geodesic rp, we obtain the quite different vector **C**. This discrepancy between the result of parallel transport along different routes is the ***holonomy*** we alluded to earlier, which Levi-Civita discovered in 1917.

It turns out to be fruitful to look at this holonomy in a slightly different way. Instead of carrying **B** to p along two different routes, suppose we start with **A** at p and then parallel transport it counterclockwise around the closed loop, $p \rightsquigarrow q \rightsquigarrow r \rightsquigarrow p$. The figure shows that it returns to p having undergone a counterclockwise rotation of $\mathcal{R}(\Delta) = \Theta$. This is the holonomy of Δ, and we can now introduce the general definition:

> *The* **holonomy** $\mathcal{R}(L)$ *of a simple closed loop* L *on a surface* $\mathcal{S}$ *is the net rotation of a tangent vector to* $\mathcal{S}$ *that is parallel transported around* L. (24.1)

Note that this definition does not specify *which* tangent vector is to be parallel-transported. The reason is that (22.4) implies that *all* tangents vectors must rotate rigidly together, through the *same* angle $\mathcal{R}(L)$. Thus,

> *We may think of the holonomy as the rotation of the entire tangent plane as it is parallel transported around the loop.*

The definition of holonomy also does not specify *where* on L we should begin. To see why this, too, does not matter, suppose that instead of starting with **A** at p, we start with **B** at q, then parallel transport it $q \rightsquigarrow r \rightsquigarrow p \rightsquigarrow q$. Using (23.1), you may visually confirm that upon return to q, the vector has undergone the *same* rotation $\mathcal{R}(\Delta) = \Theta$ as before. More generally [exercise], convince

yourself that *the holonomy is independent of the starting point of the loop* (as well as the initial tangent vector).

Next, note that in [24.1] the counterclockwise *sense* of the holonomy on the sphere *matches* the sense in which we traverse Δ. This will turn out to be because the sphere has positive curvature.

If we had instead transported the vector on a surface of *negative* curvature, then the rotation would have been *opposite* to the direction of transport. At this point we strongly encourage you to verify this empirically, using the sticky-strip construction of geodesics to create a geodesic triangle on a negatively curved patch of a suitable fruit or vegetable. You may then easily parallel-transport a toothpick around the triangle by maintaining a constant angle with each successive edge.

Not only is the sign of $\mathcal{R}(L)$ determined by the curvature within Δ, its *magnitude* is too! The constant curvature of the sphere is $\mathcal{K} = (1/R^2)$, so the total curvature residing within Δ is

$$\mathcal{K}(\Delta) = \iint_\Delta \mathcal{K}\, dA = \frac{1}{R^2} \iint_\Delta dA = \frac{1}{R^2}[R^2\,\Theta] = \Theta,$$

and therefore, for $L = \Delta$,

$$\boxed{\mathcal{R}(L) = \mathcal{K}(L).} \tag{24.2}$$

As we shall see, this is no accident—it is true for *any* simple loop L on *any* surface $\mathcal{S}$! Establishing this result (in the next chapter) will furnish us with a seemingly universal key, capable of unlocking some of the deepest mysteries we have encountered. It will unlock the *Theorema Egregium*. It will unlock the ***intrinsic*** nature of the Global Gauss–Bonnet Theorem. It will unlock the metric curvature formula, (4.10), the "*Star Trek* phaser" delivered from our future. And its generalization to higher dimensions will unlock the Riemannian curvature that lies at the heart of Einstein's curved-spacetime theory of gravity.

In fact the list goes on, extending far beyond the confines of this book. It includes Sir Michael Berry's remarkable 1983 discovery (see Shapere and Wilczek (1989) and Berry (1990)) of what is now called the ***Berry phase*** in quantum mechanics, as well as other "geometric phases" in physics. For a lovely selection of applications of holonomy to physics, see Berry (1991) (but be warned that we call ***holonomy***, physicists often call ***anholonomy***).

24.2 Holonomy of a General Geodesic Triangle

Figure [24.2] depicts a general geodesic triangle Δ on a general surface, the interior angles being θ_i and the exterior angles being φ_i, so

$$\theta_i + \varphi_i = \pi. \tag{24.3}$$

We know that the holonomy $\mathcal{R}(\Delta)$ is independent of the vector that is parallel transported around it, and we now use this freedom to make a choice that will make the answer vividly clear: we choose the tangent vector $\mathbf{v}$ to the first edge of Δ.

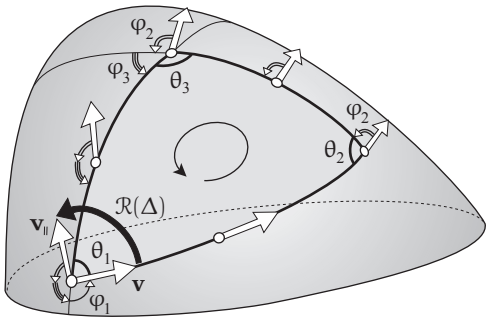
Parallel transport keeps $\mathbf{v}$ tangent to the first edge, so when it reaches the end, it makes angle φ_2 with the second edge. Since this second edge is also geodesic, the angle φ_2 is maintained as $\mathbf{v}$ is parallel transported along it. Thus, when it arrives at the end of the second edge, it makes angle $(\varphi_2 + \varphi_3)$ with the third and final edge, and this angle is maintained as it moves along that edge,

finally returning to its starting point as $\mathbf{v}_{||}$, making angle $(\varphi_1+\varphi_2+\varphi_3)$ with the initial edge. Thus we can *see* that the holonomy is

$$\mathcal{R}(\Delta)=2\pi-(\varphi_1+\varphi_2+\varphi_3). \tag{24.4}$$

In our discussion of Hopf's *Umlaufstatz*, we began by noting (see [17.1], p. 176) that if a particle travels round a Euclidean triangle Δ then the rotation of the velocity vector is $(\varphi_1+\varphi_2+\varphi_3)=2\pi$. Thus the holonomy formula (24.4) measures how much this total rotation $(\varphi_1+\varphi_2+\varphi_3)$ of the velocity *differs* from the Euclidean prediction of 2π.

Throughout this book we have used a different way of measuring the degree to which a geodesic triangle on a curved surface departs from a Euclidean one, namely, its ***angular excess***, $\mathcal{E}(\Delta)$. But in fact these two conceptually different measures of the curvature within Δ are *equal*! To see this, we combine (24.3) with (24.4):



[24.2] *The tangent vector* $\mathbf{v}$ *to the first edge of a general geodesic triangle* Δ *on a general surface is parallel transported around* Δ*, returning to its starting point as* $\mathbf{v}_{||}$*, rotated by the holonomy* $\mathcal{R}(\Delta)$.

$$\mathcal{R}(\Delta)=2\pi-[(\pi-\theta_1)+(\pi-\theta_2)+(\pi-\theta_3)]=\theta_1+\theta_2+\theta_3-\pi,$$

so,

$$\mathcal{R}(\Delta)=\mathcal{E}(\Delta). \tag{24.5}$$

All of the above generalizes easily from a geodesic 3-gon to a geodesic m-gon, P_m. First, [24.2] clearly generalizes to yield

$$\mathcal{R}(P_m)=2\pi-\sum_{i=1}^{m}\varphi_i. \tag{24.6}$$

On the other hand, the angular excess of P_m is given by (18.4), page 189:

$$\mathcal{E}(P_m)=\sum_{i=1}^{m}\theta_i-(m-2)\pi.$$

Once again using (24.3), we find that the two seemingly different measures of the curvature within P_m are actually equal:

$$\mathcal{R}(P_m)=\mathcal{E}(P_m). \tag{24.7}$$

24.3 Holonomy Is Additive

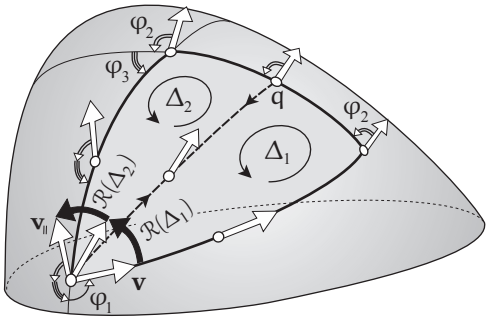
Recall from [2.8], page 23, that if we split Δ into two geodesic triangles Δ_1 and Δ_2, then the angular excess $\mathcal{E}$ is *additive*:

$$\mathcal{E}(\Delta) = \mathcal{E}(\Delta_1) + \mathcal{E}(\Delta_2).$$

It follows from (24.5) that the holonomy $\mathcal{R}$ is additive, too. However, we consider $\mathcal{R}$ to be the more fundamental of the two concepts, so rather than view its additivity as being *inherited* from $\mathcal{E}$, we should try to understand this *directly*.

Figure [24.3] is such a direct, geometric proof that $\mathcal{R}$ is additive:

$$\boxed{\mathcal{R}(\Delta) = \mathcal{R}(\Delta_1) + \mathcal{R}(\Delta_2).} \tag{24.8}$$



[24.3] Holonomy Is Additive. *The geodesic triangle Δ is split into Δ_1 and Δ_2 by the insertion of the dashed geodesic. The tangent vector $\mathbf{v}$ to the first edge of Δ_1 is parallel transported around Δ_1 and then around Δ_2. We see that the parallel transport back and forth along the dashed geodesic "cancels," and therefore $\mathcal{R}(\Delta) = \mathcal{R}(\Delta_1) + \mathcal{R}(\Delta_2)$.*

Here we have inserted the dashed geodesic into our original geodesic triangle Δ, thereby splitting it into the two geodesic triangles Δ_1 and Δ_2. The tangent vector $\mathbf{v}$ to the first edge of Δ_1 is parallel-transported around Δ_1, returning home rotated by $\mathcal{R}(\Delta_1)$. It is then parallel-transported around Δ_2, returning home rotated by $\mathcal{R}(\Delta_2)$. So the total rotation after parallel translation around both is $[\mathcal{R}(\Delta_1) + \mathcal{R}(\Delta_2)]$, as illustrated.

But because the final edge of Δ_1 (starting at q) is also the first edge of Δ_2 (ending at q), we parallel-trans-port the vector along the dashed geodesic twice, in succession, in opposite directions, so the vector returns to q *unchanged*. Successively parallel-transporting a vector around Δ_1 and then Δ_2 is therefore equivalent to parallel-transporting it around Δ, as was to be shown.

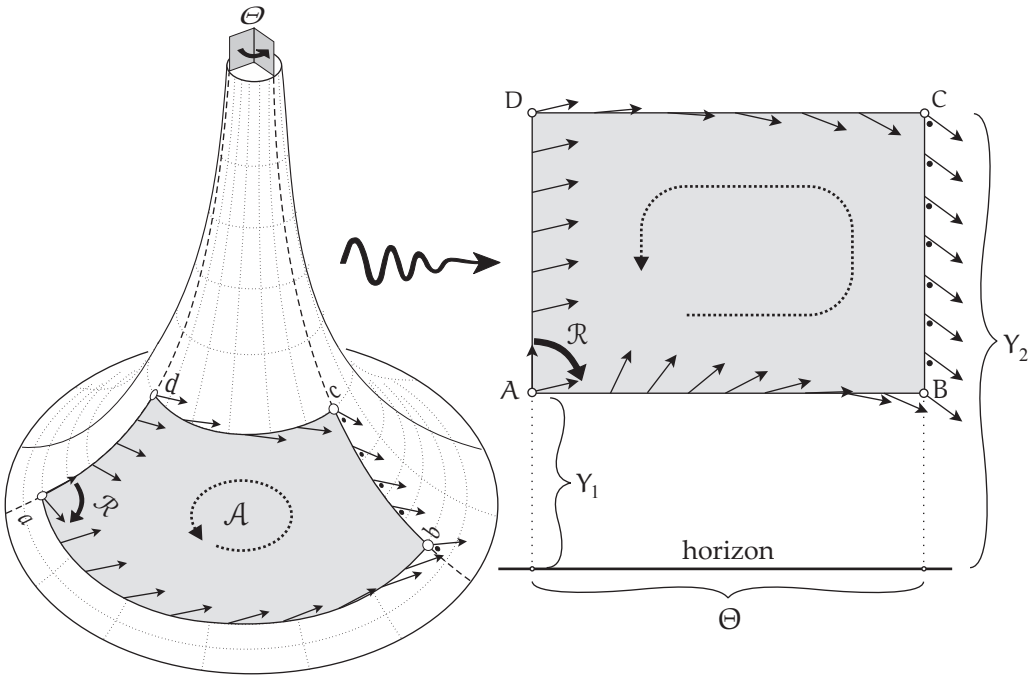
In this sense, parallel-transportation back and forth along the same curve "cancels," even if that curve is not a geodesic.

24.4 Example: The Hyperbolic Plane

We end this chapter by applying the concept of parallel transport to the pseudosphere (via the Beltrami–Poincaré half-plane model), to obtain a new, simple *intrinsic* geometric demonstration[1] of the constant negative curvature of the hyperbolic plane. In order to do this we shall *assume*, for now, the fundamental fact (24.2)—which will be proved in the next chapter—that the holonomy of a loop measures the total curvature within it.

Before we begin the demonstration, we make the important observation: just as we originally used the angular excess to give an intrinsic definition of the curvature $\mathcal{K}(\mathrm{p})$ *at* a point [see (2.1), p. 18], so (24.2) can now be used in the same way to find the curvature *at* a point p.

[1] The following argument previously appeared in Needham (2014).



[24.4] *The "rectangle"* abcd *(with area $\mathcal{A}$) on the pseudosphere (left) is conformally mapped to* ABCD *in the Beltrami–Poincaré upper-half-plane (right). When the illustrated vector at* a *is parallel-transported counterclockwise around* abcd*, it returns rotated clockwise by $\mathcal{R}$. The conformality of the map ensures that the parallel-transported vector in the map undergoes the* same *rotation $\mathcal{R}$.*

If L_p is a small loop around p, then we can apply (24.2) to L_p as it shrinks down to p to find the curvature *at* p:

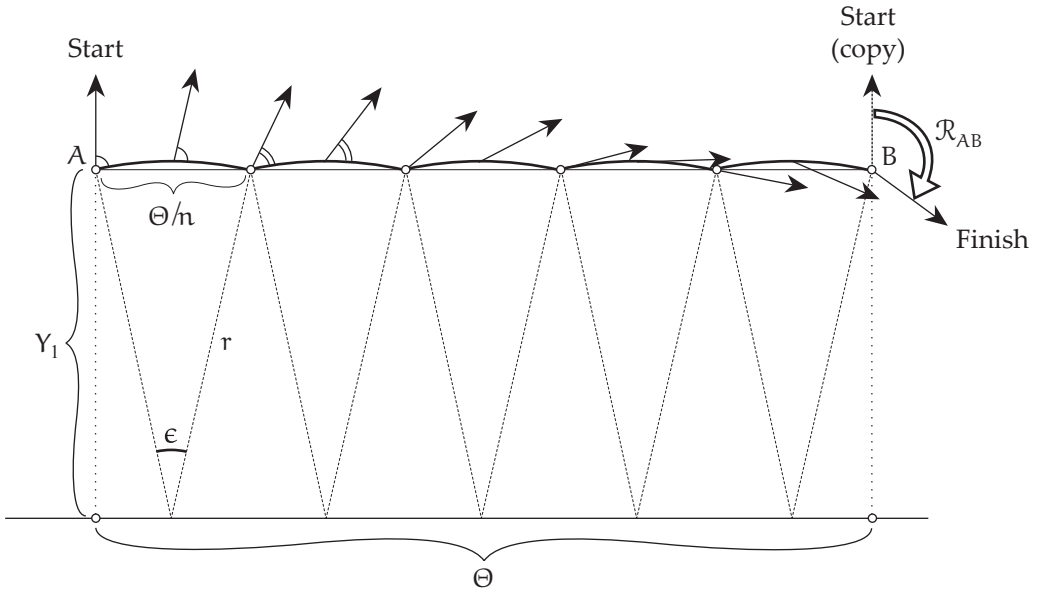
$$\mathcal{K}(p) = \lim_{L_p \to p} \frac{\mathcal{R}(L_p)}{\mathcal{A}(L_p)} = \textit{holonomy per unit area at } p. \tag{24.9}$$

We can now return to the problem at hand. On the pseudosphere of radius R, consider the "rectangle" abcd (traced counterclockwise) bounded by the vertical segments ad and bc of geodesic tractrix generators (Θ being the angle from the first to the second) together with the *non*geodesic horizontal circular arcs ab, cd. See the left side of [24.4]. As illustrated, let us parallel-transport a vector round abcd to discover the total curvature within it.

The right side of [24.4] depicts the conformal image in the Beltrami–Poincaré model: abcd is mapped to the rectangle with vertices $A = (x, Y_1)$, $B = (x + \Theta, Y_1)$, $C = (x + \Theta, Y_2)$, $D = (x, Y_2)$. Thus, using (5.6), page 55, the area $\mathcal{A}$ of the rectangle abcd on the pseudosphere is

$$\mathcal{A} = \int_{x=0}^{x=\Theta} \int_{Y_1}^{Y_2} \frac{R^2 \, dx \, dy}{y^2} = R^2 \Theta \left[\frac{1}{Y_1} - \frac{1}{Y_2} \right]. \tag{24.10}$$

At a we have chosen an initial vector pointing up the pseudosphere, along ad. As we parallel-transport it along ab, it rotates clockwise relative to the direction of motion; along bc it maintains the constant illustrated angle • with the direction of motion (because it is a geodesic); along cd it rotates counterclockwise relative to the direction of motion, *but not as much as it did on* ab; finally,



[24.5] *The initially vertical vector at* A *is parallel transported along the horizontal Euclidean line segment* AB *in the Beltrami–Poincaré half-plane. To do so, we approximate* AB *with* n *geodesic segments* (arcs of circles centred on the horizon, $y = 0$)*, then maintain constant angle with each geodesic segment, in succession. Finally, we let* n *go to infinity.*

it maintains constant angle with the geodesic da, returning to a having undergone a negative net rotation of $\mathcal{R}$.

Because the Beltrami–Poincaré map is conformal, when the vector is transported around ABCD it undergoes the *same* net rotation $\mathcal{R}$. But, as we now explain, the crucial advantage of the map is that it enables us to *see*[2] what this rotation actually is.

Divide the nongeodesic horizontal segment AB of Euclidean length Θ into n small segments of length (Θ/n). Next, as illustrated in [24.5], approximate these segments with geodesic segments: recall that these are arcs of circles centred on the horizon. Let ϵ be the angle that each such arc subtends on the horizon, as illustrated.

As the initially-vertical *Start* vector is parallel-transported along the first geodesic segment, its angle with that segment remains constant, and it therefore rotates through angle $-\epsilon$. Likewise for each successive segment, so that after all n segments have been traversed the total rotation from *Start* to *Finish* is $-n\epsilon$. But since

$$r\epsilon \asymp \frac{\Theta}{n} \qquad \text{and} \qquad r \asymp Y_1,$$

we deduce that the total angle through which the vector is rotated in the map is

$$\mathcal{R}_{AB} \asymp -n\epsilon \asymp -\frac{\Theta}{r} \asymp -\frac{\Theta}{Y_1}.$$

The same reasoning yields $\mathcal{R}_{CD} = (\Theta/Y_2)$. And since the vector does not rotate along either of the geodesics BC or DA, we deduce that the net rotation upon returning to A is

$$\mathcal{R} = \mathcal{R}_{AB} + \mathcal{R}_{CD} = -\frac{\Theta}{Y_1} + \frac{\Theta}{Y_2} = \left[-\frac{1}{R^2}\right] \mathcal{A},$$

by virtue of (24.10).

---

[2] A small amount of calculation is called for, but this is greatly simplified by our ability to manipulate "ultimate equalities" (involving $\asymp$) exactly as though they were ordinary equalities (involving $=$).

Thus,

$$\textit{rotation per unit area} = -\frac{1}{R^2}\,.$$

The fact that this answer is independent of the size, shape, and location of the rectangle proves, via (24.9), that the hyperbolic plane does indeed have *constant negative intrinsic curvature* $-1/R^2$, as was to be shown.

If you would like to try your own hand at performing parallel transport on specific surfaces (the cone and the sphere), try Exercise 5.

Chapter 25

An Intuitive Geometric Proof of the *Theorema Egregium*

25.1 Introduction

Several proofs of the *Theorema Egregium* have been found since Gauss first shocked the world with it in 1827. By far the most common proof, to be found in essentially all textbooks, is a variant of Gauss's original proof: perform a lengthy computation, sleep through the journey, then awake to find yourself staring at a formula for the *extrinsic* curvature $\mathcal{K}_{\text{ext}}$ that *simply turns out* to depend only on the metric, thereby establishing that $\mathcal{K}_{\text{ext}} = \kappa_1\kappa_2$ is in fact *intrinsic* to the surface.

There is no question that this is a watertight *proof*, just as it was when Gauss first wrote it down in 1827, but we are left clueless as to *why* the result is true!

In his authoritative (and excellent) book, *A Panoramic View of Riemannian Geometry* (Berger 2003, p. 106), Professor Marcel Berger (1927–2016)—one of the foremost geometric authorities of the late twentieth century—provides summaries of the various known proofs, and concludes, *"To our knowledge there is no simple geometric proof of the Theorema Egregium today."*[1] The purpose of this chapter is to rectify this situation by providing such a simple geometric proof.[2]

Rather than trying to understand the *Theorema Egregium* directly, we shall instead seek to understand Gauss's initial discovery, his so-called "Beautiful Theorem," (13.1). For convenience's sake, we restate the result here, again quoting Gauss's own words, exactly as he recorded the discovery in his private notebook of 1816:

> "**Beautiful Theorem**. *If a curved surface on which a figure is fixed takes different shapes in space, then the surface area of the spherical image of the figure is always the same.*"

In Section 13.3 we explained that the local *Theorema Egregium* follows easily and intuitively from the Beautiful Theorem, simply by letting the figure shrink down towards a point. So, to understand the Beautiful Theorem is to understand the *Theorema Egregium*.

But, as we noted in that same section, Dombrowski (1979) found that even in his private notebooks Gauss did not leave any trace of how he discovered the result, nor did he give any indication of a proof (assuming that he had one).

As should be clear from the present context, *we* shall prove the Beautiful Theorem using parallel transport, a concept whose discovery by Levi-Civita lay more than a century in the future in

[1] As soon as I discovered the simple geometric proof that I am about to present, I sent it to Banchoff, Berger, and Penrose, none of whom had ever seen it before. In particular, Berger, whose first language was French, emailed back, "I am glad to be able to congratulate you for this amazing proof, bypassed for above more than a century, for not only [is it] one of the most beautiful result[s] in geometry, but also this result started the whole Chern, etc. business."

[2] As this book neared completion, I learned that the same insight was previously published by Professor David W. Henderson (1998, Problem 6.3), of Cornell University. I was about to write to him when I learned that he had been struck by a car and killed, just days earlier, on December 20th, 2018. So, in the interest of keeping the record straight, the credit for this discovery should go to the late Professor Henderson (unless there is a precursor that I am not aware of). Regardless of its provenance, I hope that my *visualization* here (which Henderson does not provide) will help to make this simple and intuitive proof much more widely known. (And, on an extremely personal note, priority isn't everything: the startling, unanticipated flash of clarity—in the Sierra mountains, surrounded by pristine snow—was one of the happiest moments of my life.)

1816. Thus, as intuitive and simple as we hope you will find our parallel-transport proof, it could *not* have been Gauss's original approach.

We stress that Gauss discovered the Beautiful Theorem fully 11 years prior to developing his 1827 calculational apparatus for analyzing general surfaces. Thus the enigma remains, and, with it, the tantalizing possibility that a quite different, even simpler approach may exist—but this may have to await the second coming of Gauss.

25.2 Some Notation and Reminders of Definitions

Let $\mathcal{S}$ be the surface with unit normal vector field $\mathbf{n}$, $\mathbb{S}^2$ be the unit sphere, and $\mathfrak{n}:\mathcal{S}\mapsto\mathbb{S}^2$ be the spherical map.

Let $\mathcal{K}_{\text{ext}}$ be the *extrinsic* curvature, defined as the local (signed) area expansion factor of the spherical map, as introduced in (12.3), p. 132.

Let a tilde $\widetilde{\ }$ denote a quantity defined on $\mathbb{S}^2$. So, for example, $\mathcal{A}$ denotes area on $\mathcal{S}$, while $\widetilde{\mathcal{A}}$ denotes area on $\mathbb{S}^2$.

Let $\mathcal{E}(\Delta)$ be the angular excess of a geodesic triangle Δ on $\mathcal{S}$, or more generally let $\mathcal{E}(P_m)$ denote the angular excess of a geodesic m-gon, P_m, given by (18.4), p. 189.

Let L be a simple loop bounding a region Ω of $\mathcal{S}$, and let $\mathcal{R}(L)$ denote the net rotation (holonomy angle) of a tangent vector $\mathbf{w}$ to $\mathcal{S}$ that is parallel transported counterclockwise round L, the sense of rotation being determined by having $\mathbf{n}$ pointing at our eye. For ease of visualization, let us assume that $\mathcal{K}_{\text{ext}}$ has a *single sign* (either always positive, or else always negative) throughout Ω, so that $\mathfrak{n}(\Omega)$ does not contain any *folds*; see the discussion in Section 17.5.

Likewise, let $\widetilde{\mathcal{R}}(\widetilde{L})$ denote the holonomy of a tangent vector to $\mathbb{S}^2$ that is parallel transported around the image $\widetilde{L}\equiv\mathfrak{n}(L)$ on $\mathbb{S}^2$ of L.

Finally, let $\mathcal{K}_{\text{ext}}(\Omega)$ denote the total amount of *extrinsic* curvature within Ω,

$$\mathcal{K}_{\text{ext}}(\Omega)=\iint_{\Omega}\mathcal{K}_{\text{ext}}\,\mathrm{d}\mathcal{A},$$

and let $\mathcal{K}(\Omega)$ denote the total amount of *intrinsic* curvature within Ω,

$$\mathcal{K}(\Omega)=\iint_{\Omega}\mathcal{K}\,\mathrm{d}\mathcal{A}=\mathcal{R}(L).$$

25.3 The Story So Far

Let us briefly recap what we actually *know* so far. This is important, because throughout this book we have felt at liberty to quote (and use) future results long before we were in a position to prove or explain them. That said, we have at all times scrupulously avoided so much as a whiff of circular reasoning.

Nevertheless, it is quite possible that in the course of our time-travelling exposition the line between what is already established and what has yet to be proved may have become blurred. Here, then, are the few facts that we shall need in the following explanation of the Beautiful Theorem, all of which *have* been properly established:

- Since $\mathcal{K}_{\text{ext}}$ is defined to be the local (signed) area expansion factor of the spherical map, *the total amount of extrinsic curvature within Ω on $\mathcal{S}$ is the area of its image $\widetilde{\Omega}$ on $\mathbb{S}^2$*:

$$\mathcal{K}_{\text{ext}}(\Omega)=\iint_{\Omega}\mathcal{K}_{\text{ext}}\,\mathrm{d}\mathcal{A}=\iint_{\widetilde{\Omega}}\mathrm{d}\widetilde{\mathcal{A}}=\widetilde{\mathcal{A}}(\widetilde{\Omega}).$$

- As we proved in (12.8), page 134, the extrinsic curvature can be expressed as the product of the principal curvatures:

$$\mathcal{K}_{\text{ext}} = \kappa_1 \kappa_2.$$

Recall that we initially proved this by examining the effect of n on a small rectangle aligned with the principal directions. However, later ((15.8), p. 153) we established that *all* shrinking shapes ultimately undergo the *same* expansion, $\kappa_1\kappa_2$.

Thus the area on S^2 of the image $\widetilde{\Omega} = n(\Omega)$ can also be expressed as,

$$\widetilde{\mathcal{A}}(\widetilde{\Omega}) = \mathcal{K}_{\text{ext}}(\Omega) = \iint_{\Omega} \kappa_1 \kappa_2 \, d\mathcal{A}.$$

- Harriot's 1603 result ((1.3), p. 8) on a sphere of radius R implies (with $R = 1$) that on S^2 we have $\widetilde{\mathcal{E}}(\widetilde{\Delta}) = \widetilde{\mathcal{A}}(\widetilde{\Delta})$, and this easily generalizes to a geodesic m-gon, $\widetilde{P}_m$:

$$\widetilde{\mathcal{E}}(\widetilde{P}_m) = \widetilde{\mathcal{A}}(\text{interior of } \widetilde{P}_m).$$

- But on S^2 we also have proved in (24.5) that $\widetilde{\mathcal{R}}(\widetilde{\Delta}) = \widetilde{\mathcal{E}}(\widetilde{\Delta})$, and in (24.7) we generalized this to geodesic m-gons. Thus if $\widetilde{P}_m$ is a geodesic m-gon on S^2 then the *net rotation [holonomy] of a vector that is parallel-transported around* $\widetilde{P}_m$ *is the same as the area it encloses*:

$$\widetilde{\mathcal{R}}(\widetilde{P}_m) = \widetilde{\mathcal{A}}(\text{interior of } \widetilde{P}_m).$$

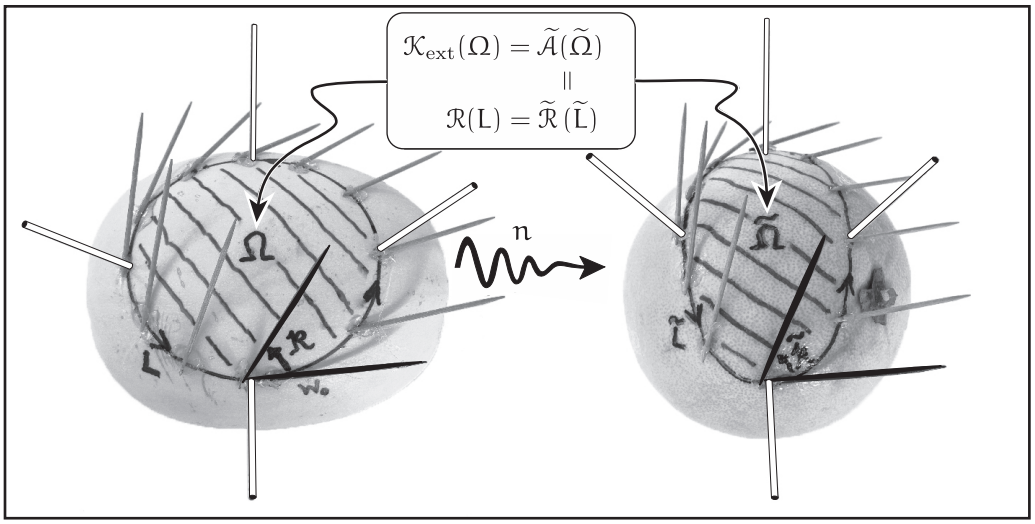
- If $\widetilde{L}$ is the (generally nongeodesic) image on S^2 of the simple loop L on S, we may approximate $\widetilde{L}$ with a geodesic m-gon and then let $m \to \infty$. Thus, combining the above results, we may summarize the relevant known facts as follows:

> ***The Story So Far***. *If the simple loop L on S (enclosing the region Ω of single-signed extrinsic curvature) is mapped by n to $\widetilde{L}$, then the total amount of extrinsic curvature within Ω on S is the (signed) area of $\widetilde{\Omega}$ on S^2, bounded by $\widetilde{L}$, and this is in turn is equal to the* holonomy *of a tangent vector to S^2 that it parallel-transported around $\widetilde{L}$:*
>
> $$\iint_{\Omega} \kappa_1 \kappa_2 \, d\mathcal{A} = \mathcal{K}_{ext}(\Omega) = \widetilde{\mathcal{A}}(\widetilde{\Omega}) = \widetilde{R}(\widetilde{L}). \tag{25.1}$$

25.4 The Spherical Map Preserves Parallel Transport

Consider (25.1), which shows the simple loop L on S (a yellow squash, on the left) and its image on S^2 (a pomelo, on the right) under the spherical map, $\widetilde{L} = n(L)$. The tangent planes at p and $n(p)$ are parallel, so any tangent vector field $\mathbf{w}$ defined along L on S is—after being transplanted from p to $n(p)$—automatically also a tangent vector field to S^2 along $\widetilde{L}$.

While the shape of L is immaterial to the following argument, here we have chosen to construct an intrinsic circle, using a stretched piece of string (not shown) to trace the locus of points at constant distance from a fixed point. To help visualize the effect of the spherical map on L, we have erected normals at four equally spaced points along L; their image points on S^2 can be identified



[25.1] Geometric Proof of the Beautiful Theorem. *The total extrinsic curvature* $\mathcal{K}_{ext}(\Omega)$ *within* Ω *on* $\mathcal{S}$ (left) *equals the area* $\widetilde{\mathcal{A}}\,\widetilde{\Omega})$ *of its spherical image* $\widetilde{\Omega}$ *on* $\mathbb{S}^2$ (right). *Furthermore,* $\widetilde{\mathcal{A}}(\widetilde{\Omega}) = \widetilde{\mathcal{R}}(\widetilde{L})$. *But* **the spherical map preserves parallel transport**, *so* $\widetilde{\mathcal{R}}\,(\widetilde{L}) = \mathcal{R}(L)$. *Thus the total* **extrinsic** *curvature* $\mathcal{K}_{ext}(\Omega)$ *is equal to the* **intrinsically** *defined holonomy* $\mathcal{R}(L)$ *on* $\mathcal{S}$.

by the fact that they have the *same* normals, as shown. Due to the disparity in the principal curvatures, note that the spherical image $\widetilde{L}$ of this circle is stretched out rather dramatically into an oval on $\mathbb{S}^2$.

Now suppose that $\mathbf{w}$ is not an arbitrary tangent field to $\mathcal{S}$ along L, but, as illustrated, is instead obtained by *parallel-transporting* an initial tangent vector $\mathbf{w}_0$ along L to generate $\mathbf{w}_{||}$. The choice of $\mathbf{w}_0$ is immaterial to the argument, but in [25.1] we have chosen it to be tangent to L.

We now come to the surprisingly simple and elegant crux of the matter. By the definition of parallel transport, the rate of change of $\mathbf{w}_{||}$ is always perpendicular to $\mathcal{S}$ as we travel along L, i.e., the rate of change is along $\mathbf{n}$. But this means that the rate of change of $\mathbf{w}_{||}$ is *also* perpendicular to $\mathbb{S}^2$ at $n(p)$, for the normal $\mathbf{n}$ to $\mathcal{S}$ at p is the *same* as the normal to $\mathbb{S}^2$ at $n(p)$. In other words, the *same* vector $\mathbf{w}_{||}$—transplanted from p to $n(p)$—is automatically parallel-transported along $\widetilde{L}$ on the sphere!

Let us make the same point differently, in the hope that (one way or the other) this critically important result will become crystal clear in the process. Our recipe [22.2] for parallel transport calls for us to move $\mathbf{w}_{||}$ along L parallel to itself in $\mathbb{R}^3$, while continually projecting back into the surface (or, technically, into the tangent plane T_p) as we go, thereby obtaining $\mathbf{w}_{||}$. Now transplant (by simple translation in $\mathbb{R}^3$) both $\mathbf{w}_{||}$ and T_p from p on $\mathcal{S}$ to $n(p)$ on $\mathbb{S}^2$. The transplanted tangent plane simply *is* the tangent plane to $\mathbb{S}^2$ at $n(p)$, so projecting into it yields parallel transport on $\mathbb{S}^2$, too.

In summary, and as illustrated,

> **The Spherical Map Preserves Parallel Transport.** *As* $\mathbf{w}_{||}$ *at* p *is parallel-transported with respect to the surface* $\mathcal{S}$ *along* L, *the* exact same *vector* $\mathbf{w}_{||}$ *at* $n(p)$ *is automatically* also *parallel-transported with respect to* $\mathbb{S}^2$ *along* $\widetilde{L} = n(L)$. (25.2)

25.5 The Beautiful Theorem and *Theorema Egregium* Explained

Upon executing a full circuit of L, *we can now literally* see *that the net rotation (holonomy) on the surface is equal to the net rotation on the sphere:*

$$\boxed{\mathcal{R}(L) = \widetilde{\mathcal{R}}\,(\widetilde{L}).}$$

But, according to "The Story So Far," (25.1), the net rotation on the sphere is simply the area enclosed on the sphere, which in turn is the total amount of *extrinsic* curvature enclosed by L back on the surface $\mathcal{S}$:

$$\boxed{\mathcal{R}(L) = \mathcal{R}(\widetilde{L}) = \widetilde{\mathcal{A}}(\widetilde{\Omega}) = \mathcal{K}_{\text{ext}}(\Omega) = \iint_{\Omega} \kappa_1\kappa_2 \, d\mathcal{A}.}$$

Since $\mathcal{R}(L)$ is defined intrinsically on $\mathcal{S}$, it is invariant under isometries of $\mathcal{S}$, and we have therefore proved Gauss's original Beautiful Theorem of 1816.

If we take L_p to be a small loop on $\mathcal{S}$ surrounding p, and let Ω_p be the region it encloses, and then shrink L_p down to p, we recover a more standard local statement of the *Theorema Egregium*:

$$\boxed{\kappa_1\kappa_2 = \lim_{L_p \to p} \frac{\mathcal{R}(L_p)}{\mathcal{A}(\Omega_p)}.}$$

Alternatively, suppose we take L_p to be a small geodesic triangle Δ_p containing p, and let $\mathcal{A}(\Delta_p)$ denote the area of the interior of Δ_p. Then, by virtue of (24.5), we have also recovered our original form of the *Theorema Egregium*, in terms of our original intrinsic definition ((2.1), p. 18) of curvature:

$$\boxed{\kappa_1\kappa_2 = \lim_{\Delta_p \to p} \frac{\mathcal{E}(\Delta_p)}{\mathcal{A}(\Delta_p)} = \mathcal{K}(p).}$$

Since the quantity on the right of either of the previous two equations is intrinsic to $\mathcal{S}$, it follows that the *extrinsically* defined curvature $\kappa_1\kappa_2$ on the left must *also* be—with thrilling unexpectedness!—*invariant under isometries*.

Thus we have arrived (at last!) at a satisfying geometric *explanation* of the empirical phenomena that we first observed in [13.1], page 139, and in [13.2], page 140.

Chapter 26

Fourth (Holonomy) Proof of the Global Gauss–Bonnet Theorem

26.1 Introduction

Recall that as the curtain fell on Act III, no fewer than three distinct explanations of GGB had played out, but all of them had been forced to rely on *extrinsic* geometry.

Only now, armed with parallel transport, are we finally in a position to fulfill the promise of an *intrinsic* proof of GGB. The elegant argument that follows is entirely due to Hopf (1956, pp. 112–113), but we shall explain some important details that Hopf's original presentation took for granted, and, more significantly, we shall explicitly *visualize* the proof in a way that Hopf's exposition did not.

A significant bonus of Hopf's intrinsic proof of GGB is that it will simultaneously provide us with a brand new proof of the Poincaré–Hopf Theorem, (19.6), page 206.

26.2 Holonomy Along an *Open* Curve?

Holonomy, as we have defined it in (24.1), is a concept that *only* makes sense for a *closed loop*, L. We parallel transport an initial vector $\mathbf{w}_0$ around L, generating $\mathbf{w}_{||}$ along the way, and *when* $\mathbf{w}_{||}$ *returns to its starting point*, we can compare it to the initial vector $\mathbf{w}_0$ to see the angle $\mathcal{R}(L)$ through which it has been rotated.

The first step in Hopf's argument is to generalize the holonomy concept to an *open* curve K. Clearly we should define this to be the rotation within the surface of $\mathbf{w}_{||}$ as it is parallel-transported along K. But rotation relative to *what*?

Recall that we previously faced a similar problem in trying to define the index $\mathcal{I}_F(s)$ of a singular point s of a vector field $\mathbf{F}$ on a surface $\mathcal{S}$. As we illustrated in [19.7], page 205, our answer was to introduce a *fiducial vector field* $\mathbf{U}$, the only requirement being that $\mathbf{U}$ not have any singular points on or inside the loop L surrounding s. This allowed us to define the index as the *number of revolutions of* $\mathbf{F}$ *relative to* $\mathbf{U}$ as we travel round s along the loop L, namely, (19.5), page 205:

$$2\pi\mathcal{I}_F(s) = \delta_L\,(\angle UF).$$

As we proved, this definition of the index is indeed well-defined, which is to say that *the choice of the vector field* $\mathbf{U}$ *does not matter*, despite the fact that the precise variation in the angle $\angle UF$ along L certainly *does* depend on the specific choice of the $\mathbf{U}$ field.

Let us now apply this same idea to the concept of holonomy. Reconsider [24.2], but now thinking of the geodesic triangle Δ as made up of its successively traversed (geodesic) edges, K_j:

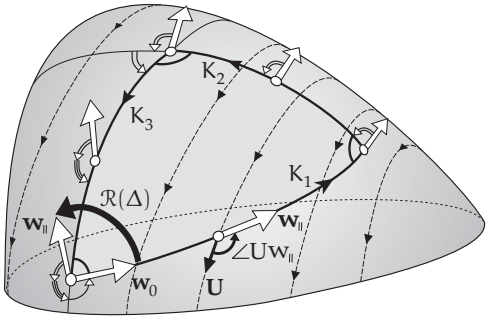
$$\Delta = K_1 + K_2 + K_3,$$

as illustrated in [26.1]. (NOTE: We are reusing this example because having geodesic edges makes parallel transport easy to visualize, but the following reasoning applies equally well to *any* (nongeodesic) curve.)

Figure [26.1] reproduces [24.2], but it introduces a fiducial vector field **U**, enabling us to generalize the concept of holonomy. We define $\mathcal{R}(K)$ for the *open* curve K to be the net change in the angle $\angle U w_{||}$—from **U** to the parallel-transported vector $\mathbf{w}_{||}$—as we travel along K:

$$\mathcal{R}_U(K) \equiv \delta_K\,(\angle U w_{||}). \tag{26.1}$$

NOTES: The subscript U is essential here, for $\mathcal{R}_U(K)$ *does indeed depend on our arbitrary choice of* **U**; thus $\mathcal{R}_U(K)$ has no true mathematical meaning; it is merely a stepping stone in the argument that is to follow. That said, if two different vectors are parallel transported along K, the angle between them remains fixed (by virtue of (22.4)) and therefore they both rotate the same amount relative to **U**. Thus, as the notation indicates, $\mathcal{R}_U(K)$ *is independent of the choice of* $\mathbf{w}_{||}$.



[26.1] *The holonomy $\mathcal{R}(\Delta)$ can be found by covering Δ with a nonsingular fiducial vector field* **U**, *and then summing over the edges K_j the change $\mathcal{R}_U(K_j)$ in the illustrated angle $\angle U w_{||}$.*

The holonomy $\mathcal{R}(\Delta)$ of the closed triangular loop Δ can then be expressed as the sum of the "holonomies" of its edges:

$$\mathcal{R}(\Delta) = \mathcal{R}_U(K_1) + \mathcal{R}_U(K_2) + \mathcal{R}_U(K_3).$$

While each of the three individual terms on the right *does* depend on the arbitrary choice of **U**, their sum $\mathcal{R}(\Delta) = \mathcal{K}(\Delta)$ is *independent of* **U**.

We note that this approach provides a better definition of holonomy than we had before, solving a potential problem that we glossed over in our initial discussion: what if $\mathcal{R}(\Delta) > 2\pi$? Then if we naively compare $\mathbf{w}_{||}$ to $\mathbf{w}_0$ we will only (and incorrectly) perceive the rotation $\mathcal{R}(\Delta)$ as being the *excess* over 2π. But the use of a fiducial vector field **U** allows us to keep continuous track of the rotation of $\mathbf{w}_{||}$, and the above formula will yield the true value of $\mathcal{R}(\Delta)$.

26.3 Hopf's Intrinsic Proof of the Global Gauss–Bonnet Theorem

Now suppose that $\mathcal{S}_g$ is a closed surface of genus g, and suppose that **F** is a vector field on $\mathcal{S}$ with a finite number of singular points, s_i. Our ultimate aim will be to prove that

$$\mathcal{K}(\mathcal{S}_g) \equiv \iint_{\mathcal{S}_g} \mathcal{K}\, dA = 2\pi \sum_i \mathcal{I}_F(s_i). \tag{26.2}$$

But, before we do so, let us explain how the Poincaré–Hopf Theorem and GGB both follow immediately from this.

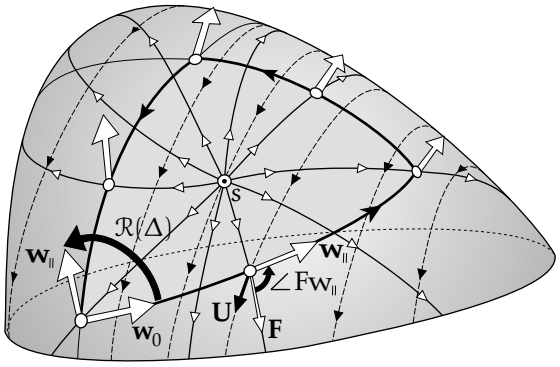
First, since the left-hand side of (26.2) is independent of **F**, it follows that the sum of the indices on the right-hand side must have the same value for *all* vector fields. But, by examining any of our previous examples of vector fields on $\mathcal{S}_g$, such as the honey-flow shown in [19.11],

page 209, this universal index sum must equal $\chi(\mathcal{S}_g) = 2 - 2g$, thereby proving the Poincaré–Hopf Theorem. Thus, if we can prove (26.2), we will *also* have our proof of GGB:

$$\mathcal{K}(\mathcal{S}_g) = 2\pi\chi(\mathcal{S}_g).$$

To begin to understand (26.2), consider [26.2] which shows the same geodesic triangle Δ as before, and the same vector $\mathbf{w}_{||}$ parallel-transported around its boundary. But now the figure also imagines a vector field $\mathbf{F}$ on the surface, with a singular point s within Δ. (Here we have specifically pictured a source, but the argument applies to any singular point.)

From this figure we deduce that,



[26.2] *The difference $\mathcal{R}(\Delta) - 2\pi\mathcal{I}_F(s)$ can be found by summing over the edges K_j the change $\Phi(K_j)$ in the illustrated angle $\angle \mathrm{Fw}_{||}$, i.e., the rotation of $\mathbf{w}_{||}$ relative to $\mathbf{F}$.*

$$\begin{aligned}
\mathcal{K}(\Delta) - 2\pi\mathcal{I}_F(s) &= \mathcal{R}(\Delta) - \delta_\Delta(\angle \mathrm{UF}) \\
&= \sum_j \Big[\mathcal{R}_U(K_j) - \delta_{K_j}(\angle \mathrm{UF})\Big] \\
&= \sum_j \Big[\delta_{K_j}(\angle \mathrm{Uw}_{||}) - \delta_{K_j}(\angle \mathrm{UF})\Big] \\
&= \sum_j \delta_{K_j}(\angle \mathrm{Fw}_{||}).
\end{aligned}$$

The final expression *measures the rotation of* $\mathbf{w}_{||}$ *relative to* $\mathbf{F}$; it is manifestly *independent* of the arbitrary fiducial vector field $\mathbf{U}$.

To simplify matters, let us adopt Hopf's notation, and *define* $\Phi(K_j)$ *to be the net rotation along* K_j *of* $\mathbf{w}_{||}$ *relative to* $\mathbf{F}$, *i.e., the net change in the angle between the vector field* $\mathbf{F}$ *and the parallel-transported vector* $\mathbf{w}_{||}$ *as we traverse* K_j:

$$\Phi(K_j) \equiv \delta_{K_j}(\angle \mathrm{Fw}_{||}).$$

NOTE: $\Phi(K_j)$ *is independent of the choice of* $\mathbf{w}_{||}$, for the same reason that $\mathcal{R}_U(K_j)$ was.

Then the previous result may be written

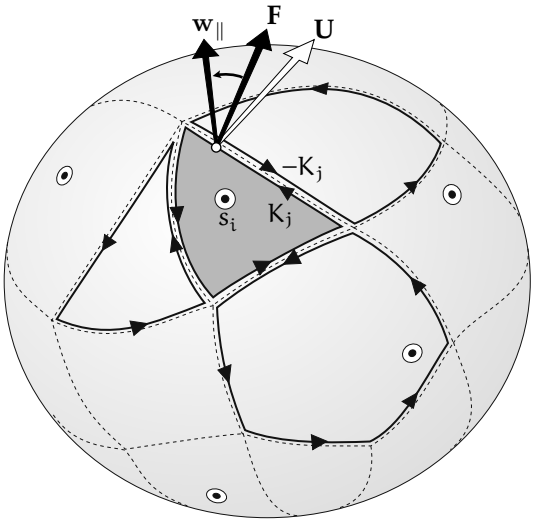
$$\mathcal{K}(\Delta) - 2\pi\mathcal{I}_F(s) = \textstyle\sum_j \Phi(K_j). \tag{26.3}$$

Clearly, this result holds equally well if Δ is replaced with a polygon, and we also remind the reader that this conclusion does *not* require that the edges be geodesics.

As usual, let $(-K_j)$ denote K_j traversed in the opposite direction. Since parallel transport does not depend on the direction in which a curve is traversed, we deduce that

$$\Phi(-K_j) = -\Phi(K_j).$$

The remainder of the proof of (26.2) now follows the exact same path as our original proof of the Poincaré–Hopf Theorem. Consider [26.3], which is simply a relabelled copy of [19.10], page 208. We partition $\mathcal{S}_g$ into polygons P_l with edges K_j (which, again, need *not* be geodesic), such that each polygon contains at most one singular point s_i of $\mathbf{F}$.



[26.3] *The surface $\mathcal{S}_g$ is partitioned into polygons P_l such that at most one singular point s_i of $\mathbf{F}$ lies in each one. The difference $\mathcal{K}(\mathcal{S}_g) - 2\pi \sum_i \mathcal{I}_{\mathbf{F}}(s_i)$ can be found by summing over all the edges K_j (of all the polygons P_l) the change $\Phi(K_j)$ in the illustrated angle $\angle \mathrm{F}w_{\|}$. But each edge of K_j abuts an oppositely directed edge of a neighbouring polygon, so the sum of these rotations over all polygons* vanishes. *Therefore, $\mathcal{K}(\mathcal{S}_g) = 2\pi \sum_i \mathcal{I}_{\mathbf{F}}(s_i)$.*

If we now sum (26.3) over all polygons, we find

$$\begin{aligned} &\mathcal{K}(\mathcal{S}_g) - 2\pi \sum_i \mathcal{I}_{\mathbf{F}}(s_i) \\ &= \sum_l \mathcal{K}(P_l) - 2\pi \sum_i \mathcal{I}_{\mathbf{F}}(s_i) \\ &= \sum_{\text{all polygons}} \Phi(K_j) \\ &= 0, \end{aligned}$$

because each edge K_j belongs to two adjacent polygons, and is therefore traversed twice in opposite directions, contributing $\Phi(K_j) + \Phi(-K_j) = 0$.

This completes Hopf's intrinsic proof of (26.2) and, with it, both the Poincaré–Hopf Theorem and the Global Gauss–Bonnet Theorem.

Chapter 27

Geometric Proof of the Metric Curvature Formula

27.1 Introduction

The sole purpose of this chapter is to use parallel transport to provide (at long last!) a geometric proof of the "*Star Trek* phaser" formula, (4.10), page 38, for the curvature in terms of the metric, which we repeat here:

$$\boxed{\mathcal{K} = -\frac{1}{AB}\left(\partial_v\left[\frac{\partial_v A}{B}\right] + \partial_u\left[\frac{\partial_u B}{A}\right]\right).} \tag{27.1}$$

We are now living more than 20 chapters in the future of the point at which this formula was first announced, so we begin by reminding the reader of the meaning of the (u, v) coordinates, and of the metric components A and B.

The construction of a general coordinate system (as illustrated in [4.3], page 35) can always be specialized to an *orthogonal* coordinate system (u, v) on a surface $\mathcal{S}$, as illustrated in [27.1]. We first draw a family of nonintersecting curves covering the patch of surface we are analyzing, so that one (and only one) curve from our family passes though each point $\widehat{p}$ on the patch. We now number these curves arbitrarily (but differentiably) to create the u-coordinate, and call these curves, with their assigned u-values, the u-*curves*.

To complete the orthogonal coordinate system, we now draw the ***orthogonal trajectories*** of the u-curves, and label them differentiably with the v-coordinate, and call these v-*curves*.

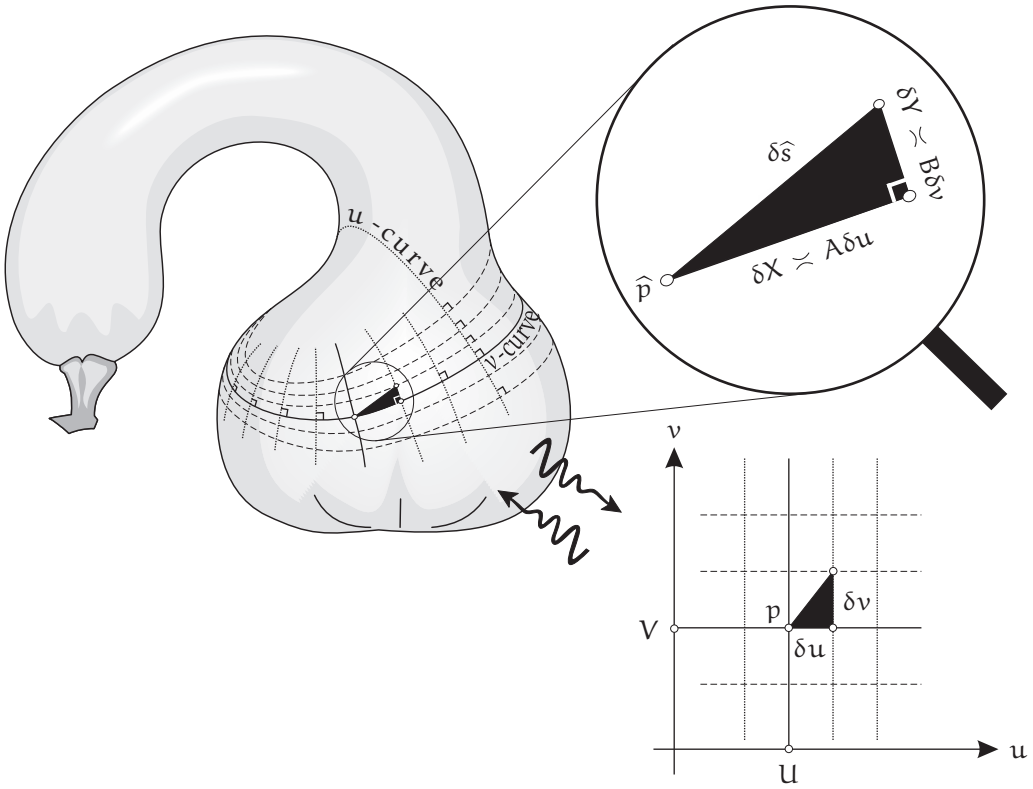
Thus, as illustrated, the point $\widehat{p}$ on $\mathcal{S}$ can be labelled by the unique u-curve (say $u = U$) and v-curve (say $v = V$) that intersect there. So, in the map, $\widehat{p}$ can now be represented as $p = (U, V)$. Thus, as illustrated, u-curves are represented in the map by vertical lines, while v-curves are represented by horizontal lines.

NOTE: Here we are reverting to our earlier notation, distinguishing objects in the map from corresponding objects in the surface by attaching a "hat" ($\widehat{\ }$) to the latter. So, for example, an element of area in the map will be denoted $\delta\mathcal{A}$, while the corresponding area on the surface will be denoted $\delta\widehat{\mathcal{A}}$.

Let X measure distance along v-curves on $\mathcal{S}$, which correspond to horizontal lines in the map. Likewise, let Y measure distance along u-curves, which correspond to vertical lines in the map. We can now use X and Y to explain the meanings of A and B. If we move a small (ultimately vanishing) distance δu along a horizontal line in the map, then the corresponding point on the surface moves an ultimately *proportional* distance δX along the corresponding v-curve:

$$\delta X \asymp A\,\delta u \qquad \Longleftrightarrow \qquad A \text{ is the local horizontal expansion factor.}$$

Likewise, if we move a distance δv along a vertical line in the map, then the corresponding point on the surface moves a distance δY along the corresponding u-curve, and



[27.1] The Metric in Orthogonal Coordinates. *Having drawn a family of (nonintersecting) "*u*-curves" (*$u = \text{const.}$*), we construct the* v*-curves as their* orthogonal trajectories. *If* $\hat{p}$ *is the intersection of the curves* $u = U$ *and* $v = V$*, then it represented by the point* $p = (U, V)$ *in the map. A small horizontal movement* δu *in the map produces an ultimately proportional movement* $\delta X \asymp A\,\delta u$ *on the surface (along a* v*-curve), and a vertical movement* δv *in the map produces an ultimately proportional* orthogonal *movement* $\delta Y \asymp B\,\delta v$ *on the surface (along a* u*-curve).*

$$\delta Y \asymp B\,\delta v \qquad \Longleftrightarrow \qquad B \text{ is the local vertical expansion factor.}$$

The *metric* then tell us the true distance $d\hat{s}$ within the surface in terms of apparent distances within the map, and Pythagoras's Theorem implies that in our *orthogonal* (u, v)-coordinate system it takes the form (4.9):

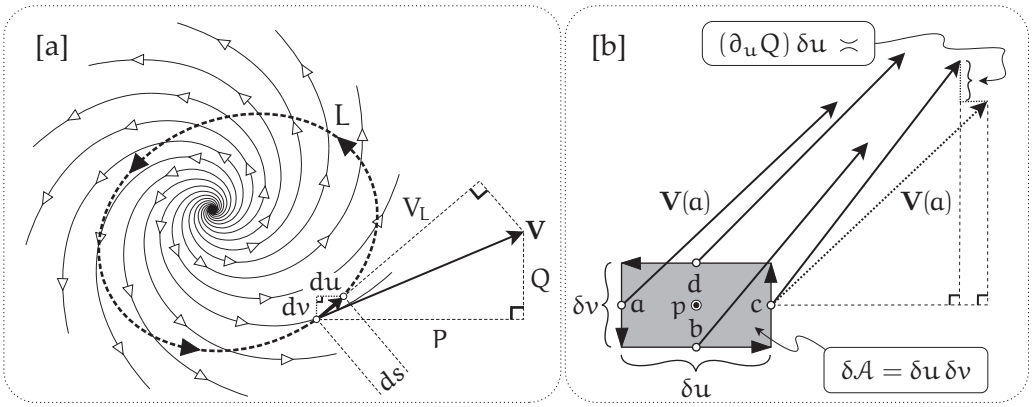
$$\boxed{d\hat{s}^2 = A^2\,du^2 + B^2\,dv^2.}$$

27.2 The Circulation of a Vector Field Around a Loop

To avoid disrupting the flow of the proof to follow, we first address a lemma, namely, how to calculate the circulation of a vector field $\mathbf{V}$ around a small (ultimately vanishing) loop.

The required concept—called the ***curl*** of $\mathbf{V}$—will be familiar to those who have studied undergraduate Vector Calculus, but a reminder can do no harm; furthermore, our exposition will be more geometrical than is customary.

Let $\mathbf{V} = \begin{bmatrix} P(u,v) \\ Q(u,v) \end{bmatrix}$ be a vector field in the (u,v)-plane, and let $\mathbf{r} = \begin{bmatrix} u \\ v \end{bmatrix}$ be the position vector of a point that traces (counterclockwise) a closed simple loop L. Then, as clarified in [27.2a],



[27.2] [a] *The* **circulation** $\mathcal{C}_L(\mathbf{V}) = \oint_L V_L\,ds = \oint_L [P\,du + Q\,dv]$. **[b]** *As the shaded rectangle shrinks, the circulation around its boundary is ultimately equal to* $\{\partial_u Q - \partial_v P\}\,\delta\mathcal{A}$.

we define the ***circulation*** $\mathcal{C}_L(\mathbf{V})$ ***of*** $\mathbf{V}$ ***around*** L to be the integral of the component of $\mathbf{V}$ in the direction of L:

$$\mathcal{C}_L(\mathbf{V}) \equiv \oint_L \mathbf{V} \bullet d\mathbf{r} = \oint_L V_L\,ds = \oint_L [P\,du + Q\,dv]\,, \tag{27.2}$$

where V_L is the (signed) projection of $\mathbf{V}$ onto the direction $d\mathbf{r}$ of L, and where $ds = |d\mathbf{r}|$. For the illustrated vortex, it is clear that V_L is always positive on L, so the circulation $\mathcal{C}_L(\mathbf{V})$ is positive.[1]

Now consider [27.2b], which illustrates a small (ultimately vanishing) coordinate rectangle R centred at p, with sides δu and δv, and hence with area $\delta\mathcal{A} = \delta u\,\delta v$. To calculate $\mathcal{C}_R(\mathbf{V})$, we perform a Riemann sum, and, as illustrated, we choose to use the *midpoints* of the sides. Of course in the limit that R vanishes, this specific choice does not matter, but we note that this choice of midpoints makes the approximation *much* more accurate than a random choice, the error for each side dying away as the *cube*[2] of the side.

The choice of midpoints also makes it especially easy to visualize and calculate the circulation. The contribution to the circulation from the upward vertical edge through c is ultimately equal to $Q(c)\,\delta v$. And the contribution from the downward edge through a is likewise $[-Q(a)]\,\delta v$. Thus the net contribution from the two vertical edges is governed by the illustrated *difference* in the vertical components of $\mathbf{V}$, namely, $\{Q(c) - Q(a)\}\,\delta v$. But, as illustrated, this change in Q is in turn governed by its partial derivative, which we may take to be evaluated at the centre p of R, as R shrinks down towards it. Thus the net contribution of the two vertical edges to the circulation is ultimately equal to $(\partial_u Q\,\delta u)\,\delta v$.

Exactly the same reasoning applies to the two horizontal edges, and therefore the total circulation around R is given by

$$\begin{aligned}
\mathcal{C}_R(\mathbf{V}) &= \oint_R [P\,du + Q\,dv] \\
&\asymp Q(a)\,(-\delta v) + P(b)\,(\delta u) + Q(c)\,(\delta v) + P(d)\,(-\delta u) \\
&= \{Q(c) - Q(a)\}\,\delta v - \{P(d) - P(b)\}\,\delta u
\end{aligned}$$

[1] In many important physical circumstances, the value of $\mathcal{C}_L(\mathbf{V})$ is *independent* of the precise shape of L; if [27.2a] were such an example, $\mathcal{C}_L(\mathbf{V})$ it would simply measure the total strength of the vortex. See VCA, Chapter 11.

[2] See VCA, p. 382

$$\begin{aligned} &\asymp \quad (\partial_u Q\, \delta u)\ \delta v - (\partial_v P\, \delta v)\ \delta u \\ &= \quad \{\partial_u Q - \partial_v P\}\ \delta\mathcal{A}. \end{aligned}$$

As this formula shows, we can now *define the* **curl** *of* $\mathbf{V}$ *to be the local circulation per unit area*:

$$\boxed{\text{curl}\begin{bmatrix} P \\ Q \end{bmatrix} \asymp \frac{\mathcal{C}_R(\mathbf{V})}{\delta\mathcal{A}} \asymp \partial_u Q - \partial_v P.} \tag{27.3}$$

27.3 Dry Run: Holonomy in the Flat Plane

In order to get our feet wet, let us see how to determine the holonomy per unit area ($\asymp \mathcal{K}$) for a small loop in the flat plane. Of course the answer had better turn out to be $\mathcal{K} = 0$!

In the previous chapter we generalized ***holonomy*** to an *open curve* $\widehat{K}$, defining it in (26.1) to be the rotation of any parallel-transported vector $\mathbf{w}_{||}$ along $\widehat{K}$, relative to an arbitrary (but singularity-free) fiducial vector field $\mathbf{U}$:

$$\mathcal{R}_U(\widehat{K}) \equiv \delta_{\widehat{K}}\,(\angle U w_{||}).$$

Recall that while this is independent of $\mathbf{w}_{||}$, it does not directly tell us *anything* about the geometry of the surface—it merely measures the rotation of $\mathbf{w}_{||}$ relative to the *arbitrarily* chosen $\mathbf{U}$ field. However, if $\widehat{K}$ becomes a *closed loop* $\widehat{L}$, then $\mathcal{R}_U(\widehat{L}) = \mathcal{R}(\widehat{L})$ becomes *independent* of $\mathbf{U}$, and it equals the total amount of curvature within $\widehat{L}$.

Let us take our surface $\mathcal{S}$ to be the flat plane, and let us apply the idea above, using ordinary polar coordinates: $u = r$, $v = \theta$, $A = 1$, and $B = r$:

$$d\hat{s}^2 = A^2\, du^2 + B^2\, dv^2 = dr^2 + r^2\, d\theta^2.$$

On the left of [27.3] we see a small (ultimately vanishing) rectangular loop $L = efghe$ in the polar-coordinate map plane, with sides δr and $\delta\theta$. On the right is its image $\widehat{L} = \widehat{ef}\widehat{gh}\widehat{e}$ on the (flat!) surface $\mathcal{S}$. Let us choose the fiducial vector field $\mathbf{U}$ to be radial, pointing outward along the rays $\theta = const.$, and let us find the holonomy along each of the four legs of the closed loop $\widehat{L}$. Starting at $\widehat{e}$, let us choose our initial vector $\mathbf{w}$ to be $\mathbf{U}$ at $\widehat{e}$, and now let us parallel transport it as $\mathbf{w}_{||}$ around the loop.

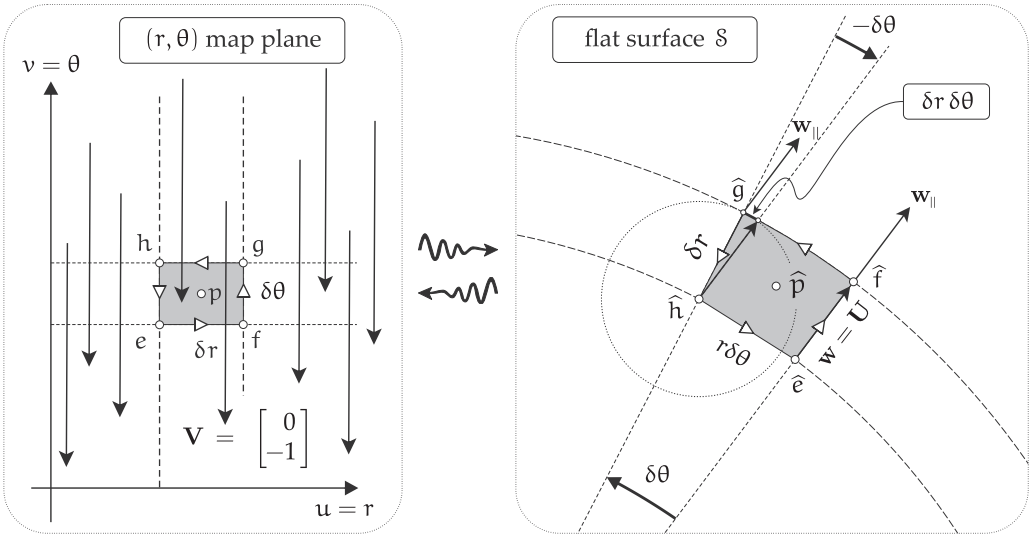
On the first leg of our journey, $\mathbf{w}_{||}$ does not rotate relative to $\mathbf{U}$, so $\delta\mathcal{R}_U(\widehat{ef}) = 0$. On the second leg, as we see clearly in [27.3], the rotation of $\mathbf{w}_{||}$ is $\delta\mathcal{R}_U(\widehat{fg}) = -\delta\theta$. On the third, radial leg $\widehat{gh}$ there is no rotation. Finally, as we return home along $\widehat{he}$, the rotation of $\mathbf{w}_{||}$ relative to $\mathbf{U}$ is $\delta\theta$.

Thus, the rotations of $\mathbf{w}_{||}$ relative to $\mathbf{U}$ cancel, and therefore the holonomy (which is independent of $\mathbf{U}$)—and with it the curvature—does indeed vanish, as it should:

$$\begin{aligned} \mathcal{R}(\widehat{L}) &= \delta\mathcal{R}_U(\widehat{ef}) + \delta\mathcal{R}_U(\widehat{fg}) + \delta\mathcal{R}_U(\widehat{gh}) + \delta\mathcal{R}_U(\widehat{he}) \\ &= 0 + (-\delta\theta) + 0 + \delta\theta \\ &= 0. \end{aligned}$$

But there is another way to determine the rotation of $\mathbf{w}_{||}$ relative to $\mathbf{U}$, and it is this method that we shall generalize to derive the general formula for $\mathcal{K}$: we shall use local measurements of *distances* within the coordinate grid—in other words, the *metric*.

The apparent rotation along $\widehat{ef}$ arises because the far side of the quadrilateral at radius $r + \delta r$ is longer than the near side, at radius r. How much longer? Well, the far side has length $(r + \delta r)\,\delta\theta$, while the near side has length $r\,\delta\theta$, so, as illustrated, the increase in length is $\delta r\,\delta\theta$.



[27.3] *The fact that* $\delta\mathcal{R}_U(\widehat{he}) = \delta\theta$ *can be deduced from the* metric *fact that this side is shorter than the opposite side by* $\delta r\,\delta\theta$. *The circulation of* $\mathbf{V}$ *around the loop in the map plane yields the (vanishing) holonomy on the (flat) surface on the right.*

If we think of this as ultimately equal to a small arc of the illustrated circle (of radius δr) centred at $\widehat{h}$ passing through $\widehat{g}$, then this increase in length subtends an angle at $\widehat{h}$ that is the rotation that we seek:

$$-\delta\mathcal{R}_U(\widehat{fg}) \asymp \frac{\text{arc}}{\text{radius}} \asymp \frac{\text{increase in side length}}{\text{orthogonal side}} = \frac{\delta r\,\delta\theta}{\delta r} = \delta\theta.$$

More generally, if we call the side length $\delta Y \asymp B\,\delta v$, then the increase $\delta^2 Y$ resulting from increasing u by δu will be

$$\delta^2 Y \asymp [\partial_u B\,\delta u]\,\delta v.$$

In the present case, $B = r$, $u = r$, and $v = \theta$, and therefore

$$\delta^2 Y \asymp [\partial_u B\,\delta u]\,\delta v = [\partial_r r\,\delta r]\,\delta\theta = \delta r\,\delta\theta,$$

as it should.

Before we turn to the general case of a curved surface, let us reinterpret what we have done, but now in the (r, θ) map plane on the left of [27.3]. On each leg of the loop L, the rotation on S is given by

$$\delta\mathcal{R}_U(\text{edge}) = 0\,\delta r + (-1)\,\delta\theta = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} \delta r \\ \delta\theta \end{bmatrix}.$$

The vector field $\mathbf{V} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ is illustrated in the map plane of [27.3]. We now recognize the fact the holonomy of the loop on S is equal to the circulation of $\mathbf{V}$ around the loop in the map plane. The flow of $\mathbf{V}$ is perpendicular to the top and bottom edges, so they contribute nothing to the circulation. But $\mathbf{V}$ does flow directly along the edge he, and it flows directly against the edge fg, and these two contributions are exactly equal and opposite to each other.

In summary,

$$\mathcal{R}(\widehat{L}) = \mathcal{C}_L(\mathbf{V}).$$

As we shall now see, in the general case, too, it is possible to visualize the holonomy of a loop on a curved surface $\mathcal{S}$ as the circulation of a vector field $\mathbf{V}$ around the corresponding loop in the (u, v) map plane, but in this case the circulation will generally *not* vanish, corresponding to the presence of curvature within the loop on $\mathcal{S}$

27.4 Holonomy as the Circulation of a Metric-Induced Vector Field in the Map

In this section we will generalize the above argument and derive a formula for the holonomy $\mathcal{R}(\widehat{L})$ of a simple loop $\widehat{L}$ on $\mathcal{S}$ in terms of the circulation $\mathcal{C}_L(\mathbf{V})$ of a special vector field $\mathbf{V}$—determined by the metric—around the corresponding loop L in the (u, v)-map.

Let R denote the small rectangle (with boundary $L = efgh$) shown in the map plane of [27.3] (with centre p) only now with general (u, v)-coordinates, and so with sides δu and δv. This maps to a curvilinear quadrilateral $\widehat{R}$ on the curved surface $\mathcal{S}$ with corresponding "centre" $\widehat{p}$.

Since the sides of $\widehat{R}$ are ultimately $\delta X \asymp A\,\delta u$ and $\delta Y \asymp B\,\delta v$, the area $\delta\widehat{\mathcal{A}}$ of $\widehat{R}$ is ultimately related to the area $\delta\mathcal{A} = \delta u\,\delta v$ of R by,

$$\boxed{\delta\widehat{\mathcal{A}} \asymp \delta X\,\delta Y \asymp (A\delta u)(B\delta v) = (AB)\,\delta\mathcal{A}.} \tag{27.4}$$

In other words, as we pass from the map to the surface, *the local area expansion factor is* AB.

As R shrinks down to the centre point p, $\widehat{R}$ shrinks down to the point $\widehat{p}$, so that to the naked eye it will appear to be a true, plane rectangle. However, in order to investigate curvature we must place $\widehat{R}$ under a powerful microscope of magnification $(1/\delta u)$—or $(1/\delta v)$—in which case it becomes clear that the lengths of opposite edges of this "rectangle" differ from each other, if only very slightly.

At this point, to follow along more vividly, we *strongly* encourage you to draw a small "rectangle" $\widehat{R}$ on an orange, apple, or grapefruit, perhaps using ordinary spherical polar coordinates to create your grid, by drawing circles of longitude and latitude. You may then parallel transport a toothpick along one of the edges of $\widehat{R}$, and watch how it appears to rotate relative to the coordinate grid you have chosen to draw on your fruit's surface.

Figure [27.4] shows the vertices of $\widehat{R}$ orthogonally projected onto the osculating plane of $\mathcal{S}$ at any interior point of $\widehat{R}$, say $\widehat{p}$. We have connected these with straight lines to form a quadrilateral. Note that while the (u, v)-coordinate curves on the surface are always orthogonal, the sides of this projected quadrilateral are not orthogonal. But as R shrinks, so too does $\widehat{R}$, and its form is ultimately a rectangle. Here we have deliberately drawn a rather extreme deviation from this limiting case, in order to make the geometrical reasoning easier to follow.

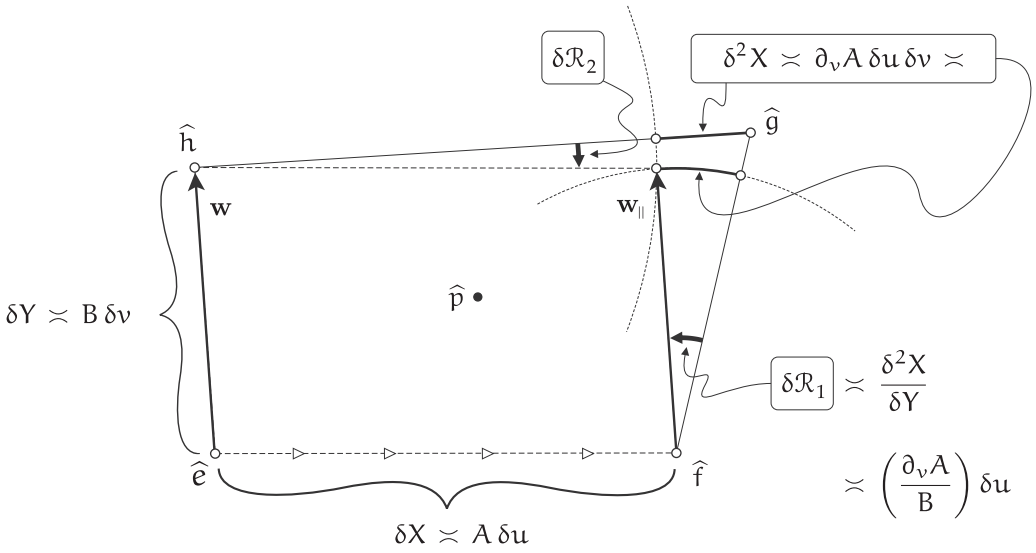
Using the same notation as in the previous subsection, let $\delta^2 X = \delta[\delta X]$ denote the increase in δX, as illustrated. This δ^2-notation serves to remind us that we are now dealing with a *second-order* infinitesimal, as is made manifest by this:

$$\delta^2 X \asymp [\partial_v A\,\delta v]\,\delta u.$$

As [27.4] demonstrates, $\delta^2 X$ is ultimately equal to the arc of the circle of radius $\delta Y \asymp B\,\delta v$, subtending the angle $\delta\mathcal{R}_1$ at $\widehat{e}$. Therefore the small holonomy resulting from the change δu in the map is ultimately given by

$$\delta\mathcal{R}_1 \asymp \frac{\text{arc}}{\text{radius}} \asymp \frac{\delta^2 X}{\delta Y} \asymp \frac{(\partial_v A\,\delta v)\,\delta u}{B\,\delta v} = \left(\frac{\partial_v A}{B}\right)\delta u,$$

in which both $\partial_v A$ and B are understood to be evaluated at the centre p of R, as R shrinks down to p.



[27.4] Geometric Proof of the Metric Curvature Formula. *Parallel transport along* $\widehat{e}\widehat{f}$ *induces a rotation* $\delta\mathcal{R}_1$*—relative to the* (u, v)*-curves—given by* $\delta\mathcal{R}_1 \asymp \left(\frac{\partial_v A}{B}\right)\delta u$.

Exactly the same reasoning applies to the holonomy resulting from the change δv in the map, simply by performing the simultaneous exchanges $u \leftrightarrow v$ and $A \leftrightarrow B$.

However, if $\partial_u B$ is *positive*—as we have drawn it to be in [27.4]—then the resulting rotation $\delta\mathcal{R}_2$ of the parallel transported vector relative to the coordinate grid is now *clockwise*, as illustrated, and therefore we must introduce a minus sign into the formula:

$$\delta\mathcal{R}_2 \asymp -\left(\frac{\partial_u B}{A}\right)\delta v.$$

The linearity of the intrinsic derivative implies that the net rotation resulting from the changes δu and δv is the sum of the separate rotations, and is therefore given by

$$\delta\mathcal{R} \asymp \left[\frac{\partial_v A}{B}\right]\delta u - \left[\frac{\partial_u B}{A}\right]\delta v = \begin{bmatrix} (\partial_v A)/B \\ -(\partial_u B)/A \end{bmatrix} \bullet \begin{bmatrix} \delta u \\ \delta v \end{bmatrix}. \tag{27.5}$$

To obtain the holonomy of a simple closed loop $\widehat{L}$ on $\mathcal{S}$, we must integrate this formula along its image L in the (u, v)-map. We have thus arrived at this important result:

Let $\mathcal{S}$ *be a surface with metric* $d\hat{s}^2 = A^2\,du^2 + B^2\,dv^2$*, and let* $\widehat{L}$ *be a simple loop on* $\mathcal{S}$*, represented by the simple loop* L *in the* (u, v)*-map plane. If in the map plane we define the vector field*

$$\mathbf{V} \equiv \begin{bmatrix} (\partial_v A)/B \\ -(\partial_u B)/A \end{bmatrix}, \tag{27.6}$$

then the holonomy of $\widehat{L}$ on $\mathcal{S}$ is equal to the circulation of $\mathbf{V}$ around L in the map*:*

$$\mathcal{R}(\widehat{L}) = \mathcal{C}_L(\mathbf{V}).$$

You may easily verify [exercise] that this general result is in accord with the special case illustrated in [27.3].

27.5 Geometric Proof of the Metric Curvature Formula

Now let us take the limit that the loop $\widehat{L}$—with area $\delta\widehat{\mathcal{A}}$—shrinks down towards $\widehat{p}$. The curvature $\mathcal{K}(\widehat{p})$ is ultimately equal to the holonomy per unit area, so by combining (27.4) and (27.6) we obtain,

$$\mathcal{K}(\widehat{p}) = \lim_{\widehat{L}\to\widehat{p}} \frac{\mathcal{R}(\widehat{L})}{\delta\widehat{\mathcal{A}}} \asymp \frac{1}{AB}\left[\frac{\mathcal{C}_L(\mathbf{V})}{\delta\mathcal{A}}\right].$$

But (27.3) now tells us that the bracketed term on the right—namely, the local circulation per unit area—is none other than the *curl* of $\mathbf{V}$.

This allows us to complete our geometric proof of the metric curvature formula, (27.1), one of the most elegant and *explicit* manifestations of Gauss's *Theorema Egregium* of 1827:

$$\begin{aligned}
\mathcal{K}(\widehat{p}) &= \frac{1}{AB}\left\{\text{curl of}\begin{bmatrix}(\partial_v A)/B\\ -(\partial_u B)/A\end{bmatrix}\right\}\\
&= \frac{1}{AB}\left\{\partial_u\left[-\frac{\partial_u B}{A}\right]-\partial_v\left[\frac{\partial_v A}{B}\right]\right\}\\
&= -\frac{1}{AB}\left(\partial_v\left[\frac{\partial_v A}{B}\right]+\partial_u\left[\frac{\partial_u B}{A}\right]\right).
\end{aligned}$$

Of course in proving this formula, we have also proved the important special case in which the map is *conformal*, so that $A = B = \Lambda$. In this case the formula reduces to the even more elegant form, (4.16):

$$\boxed{\mathcal{K} = -\frac{\nabla^2 \ln \Lambda}{\Lambda^2}.}$$

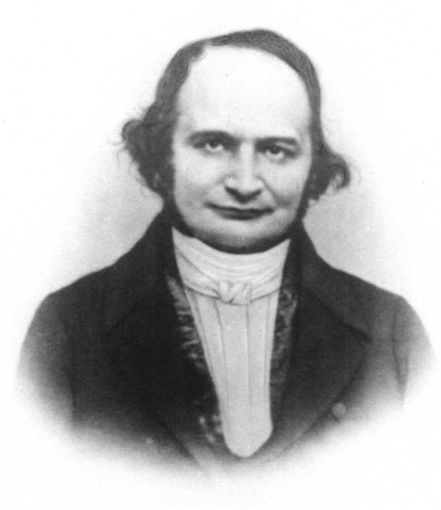
Chapter 28

Curvature as a Force between Neighbouring Geodesics

28.1 Introduction to the Jacobi Equation

This chapter introduces a brand new interpretation of curvature, one that is critically important in Einstein's curved-spacetime theory of gravity. The idea is to look at the separation of neighbouring unit-speed geodesics, and to examine their *relative acceleration*, which may be towards each other (attraction), or away from each other (repulsion).

We first examine this phenomenon in the three principal cases of *constant* curvature, and we then go on to analyse a general surface of variable curvature. In this general case, we will provide two different geometric proofs of the fundamental equation that governs this relative acceleration, called the ***Equation of Geodesic Deviation***, or the ***Jacobi Equation***, named after Carl Gustav Jacob Jacobi—pictured in [28.1]—who discovered it in 1837.



[28.1] *Carl Gustav Jacob Jacobi (1804–1851).*

As we shall see, the essential insight is that if two neighbouring geodesics pass through a region of *positive* curvature, they are *attracted* to each other. If these two geodesics are launched from a common point of origin o, in slightly different directions, this attractive force arising from the positive curvature may cause them to be focused back to a *second* intersection point, called a ***conjugate point*** of o.

On the other hand, neighbouring geodesics travelling through a region of *negative* curvature are *repelled* by each other, accelerating apart.

In both cases, the force of attraction or repulsion is directly *proportional to the separation* of the geodesics, and (locally) the proportionality "constant" *is equal to the curvature* of the surface at the location of the particle!

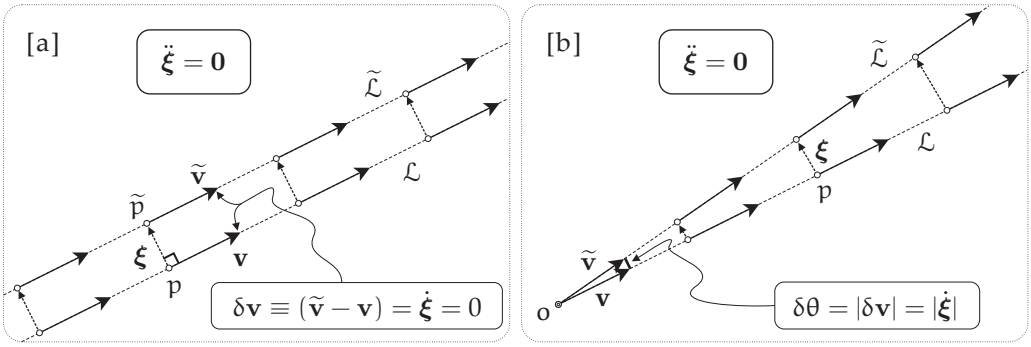
This is essence of Jacobi's discovery.

28.1.1 Zero Curvature: The Plane

Consider [28.2a], which shows a particle with position $\mathrm{p}(\mathrm{t})$ travelling at unit speed along a geodesic straight line $\mathcal{L}$ in the plane, with unit velocity $\mathbf{v} = \dot{\mathbf{p}}$. It also shows a neighbouring *parallel* geodesic $\widetilde{\mathcal{L}}$, traced by $\widetilde{\mathrm{p}}$, with velocity $\widetilde{\mathbf{v}}$.

As illustrated,

$$\text{Perpendicular connecting vector } \boldsymbol{\xi} \equiv \overrightarrow{\mathrm{p}\widetilde{\mathrm{p}}} \qquad \Longrightarrow \qquad \delta\mathbf{v} \equiv \widetilde{\mathbf{v}} - \mathbf{v} = \dot{\boldsymbol{\xi}}.$$



[28.2] **[a]** *In the flat plane, the separation* $\boldsymbol{\xi}$ *of neighbouring parallel lines is constant, so* $\dot{\boldsymbol{\xi}} = \mathbf{0}$*, and so (trivially)* $\ddot{\boldsymbol{\xi}} = \mathbf{0}$. **[b]** *If the lines instead diverge, separated by angle* $\delta\theta$*, then* $|\dot{\boldsymbol{\xi}}| = \delta\theta$*, and so* $\ddot{\boldsymbol{\xi}} = \mathbf{0}$*, again.*

Obviously in this case $\widetilde{\mathbf{v}} = \mathbf{v}$, so $\delta\mathbf{v} = \dot{\boldsymbol{\xi}} = \mathbf{0}$, but in the general case the equation simply says that the difference of the two velocities is the velocity with which $\widetilde{\mathcal{L}}$ departs from $\mathcal{L}$.

The focus of our attention in this chapter will not be on the relative velocity, but rather on the relative *acceleration*:

$$\ddot{\boldsymbol{\xi}} = \frac{\mathrm{d}(\delta\mathbf{v})}{\mathrm{dt}} = \textit{relative acceleration}. \tag{28.1}$$

In [28.2a] we find (trivially) that $\ddot{\boldsymbol{\xi}} = \mathbf{0}$.

Next, consider [28.2b], in which $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ are now rays emanating from their common point of origin o, separated by a small angle $\delta\theta$. These geodesics are spreading out, diverging from each other, but they are doing so at a *steady rate* $\delta\theta$. The diagram explains this by geometrically constructing $\delta\mathbf{v}$ as the connecting vector from the tip of $\mathbf{v}$ to the tip of $\widetilde{\mathbf{v}}$, which is ultimately equal to the arc of the circle of radius $|\mathbf{v}| = 1$, subtending the angle $\delta\theta$ at o.

We can also deduce this symbolically. If $\mathrm{r} = \mathrm{op}$, then the unit speed of p along $\mathcal{L}$ can be written $|\mathbf{v}| = \dot{\mathrm{r}} = 1$. Since $|\boldsymbol{\xi}| \asymp \mathrm{r}\,\delta\theta$, the speed of separation $|\dot{\boldsymbol{\xi}}| \asymp |(\mathrm{r}\,\delta\theta)\dot{}| = \delta\theta$.

Thus, even though these geodesics diverge from each other, they do so without any force *pushing* them apart—*their relative acceleration vanishes*:

$$\ddot{\boldsymbol{\xi}} = \mathbf{0}.$$

As we shall come to see, this vanishing of the relative acceleration of neighbouring geodesics is a new manifestation of the plane's vanishing curvature.

28.1.2 Positive Curvature: The Sphere

Consider [28.3]. We launch two particles from the north pole N of a sphere of radius R, the angle between them being $\delta\theta$. Recall that if the particles are stuck to the surface but are not subjected to any sideways forces *within* the surface, they will automatically follow geodesics of the surface, in this case, great circles.

After time t the particle with position $\mathrm{p}(\mathrm{t})$ on one of these geodesics will have travelled distance $\mathrm{r} = \mathrm{t}$ over the surface, subtending angle ϕ at the centre of the sphere. Thus $\mathrm{p}(\mathrm{t})$ lies on the

illustrated circle of latitude of intrinsic radius $r = t = R\phi$, but with the illustrated *extrinsic* radius,

$$\rho = R \sin\phi = R \sin\left(\frac{t}{R}\right).$$

Thus the length ξ of the illustrated connecting vector $\boldsymbol{\xi}$ is

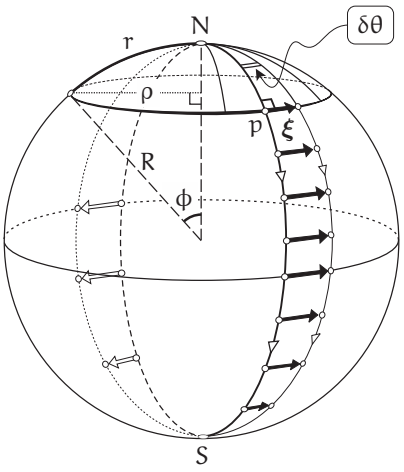
$$|\boldsymbol{\xi}| = \xi = \rho\,\delta\theta = R\,\delta\theta \sin\left(\frac{t}{R}\right).$$

Differentiating twice yields (first) the relative velocity, and (second) the relative acceleration as follows:

$$\frac{d\xi}{dt} = \left[\frac{1}{R}\right] R\,\delta\theta \cos\left(\frac{t}{R}\right)$$

$$\Longrightarrow \qquad \frac{d^2\xi}{dt^2} = -\left[\frac{1}{R^2}\right] R\,\delta\theta \sin\left(\frac{t}{R}\right).$$

But since the curvature of the sphere is $\mathcal{K} = +(1/R^2)$, this result can be written in the following remarkably compact and elegant form, called the ***Equation of Geodesic Deviation***, or the



[28.3] The Jacobi Equation. *Two particles are launched with unit speed from the north pole* N *of the sphere or radius* R*, a small angle* $\delta\theta$ *apart. After time* t*, their separation is* $\xi = R\,\delta\theta \sin(t/R)$*. Their relative acceleration is therefore given by the Jacobi Equation:* $\ddot{\boldsymbol{\xi}} = -\mathcal{K}\boldsymbol{\xi}$.

Jacobi Equation:

$$\boxed{\ddot{\boldsymbol{\xi}} = -\mathcal{K}\,\boldsymbol{\xi}.} \tag{28.2}$$

We have merely proved the special case of this equation in the case where $\mathcal{K}$ is constant (and positive), but soon we will be able to prove that it is true on a general surface of *variable* curvature. In the general case we must take $\mathcal{K}$ in the equation to be $\mathcal{K}(p)$, i.e., the curvature *at* the location of the particle as it travels across the surface along the geodesic.

There is much more to be said about this example before we move on. First, for some readers, (28.2) may already be ringing a loud, harmonic bell! For this Equation of Geodesic Deviation also goes by a quite different name—it is the equation of the ***harmonic oscillator***, which is ubiquitous in physics, in both the classical and quantum realms.

It is easy enough to create your own personal harmonic oscillator at home. Take a rubber band (or spring) and attach one end to the underside of a table. Now attach a small object (which we shall take to have unit mass) to the other end, and gently lower it until it hangs in equilibrium, the downward pull of gravity being balanced by the upward pull of the stretched rubber band or spring.

Now launch the weight straight down from this equilibrium position, thereby stretching the rubber band or spring. In 1676[1] Robert Hooke (a contemporary and a rival of Newton) discovered that the force pulling the weight back up is *proportional* to the displacement $\boldsymbol{\xi}$ from the equilibrium position—this is called ***Hooke's Law***, and the proportionality constant k is called the ***stiffness*** of the rubber band or spring.

---

[1] In fact Hooke merely *laid claim* to his discovery in 1676, without revealing *what* he had discovered, publishing it as an incomprehensible Latin anagram: *"ceiiinosssttuv"*! Only in 1678 did he reveal the *solution* to his anagram: *ut tensio, sic vis* ("as the extension, so the force").