Newton's Second Law of Motion now yields *exactly* the same equation of motion as Jacobi's Equation, (28.2), but with *the surface's curvature replaced by the spring's stiffness*:

$$\boxed{\ddot{\boldsymbol{\xi}} = -k\,\boldsymbol{\xi}, \qquad \text{with solution} \qquad \boldsymbol{\xi} = \boldsymbol{\xi}_0 \sin(\sqrt{k}\,t),}$$

where $\boldsymbol{\xi}_0$ denotes the maximum extension, occurring at the moment that the spring halts the downward motion and starts pulling the weight back up.

The appropriateness of the name *harmonic oscillator* is clear from this experiment and from the mathematical solution above: the weight *oscillates* up and down, sinusoidally, achieving its maximum speed (equal to its launch speed) each time it passes through the original equilibrium position, $\boldsymbol{\xi} = \mathbf{0}$.

Now, with this *mathematically perfect* physical analogy in mind, let us return to [28.3]. First, notice that immediately after we launch the two particles, they appear to diverge at the same (acceleration-free) rate they would in the plane, as illustrated in [28.2b]. This is clear geometrically, but can also be confirmed by noting that $\sin\epsilon \asymp \epsilon$, and therefore, as r goes to zero,

$$\xi = R\,\delta\theta\,\sin\left(\frac{r}{R}\right) \asymp R\,\delta\theta\left(\frac{r}{R}\right) = r\,\delta\theta.$$

Despite the fact that the particles are moving freely over the surface, as their separation grows the curvature effectively exerts an *attractive force* that is *exactly* like a spring, and this mysterious curvature "force" starts to slow their separation.

As the separation grows, so does the force of attraction, in direct proportion, and hence the speed of separation is reduced more rapidly, until, finally, the relative speed drops to zero, just as the particles reach maximum separation at the equator. (This corresponds to the weight being at its lowest point, with the spring at maximum extension.) The attraction is now at maximum strength, and it starts pulling the geodesics back together again, until they are finally focussed to a *second* intersection point at the south pole, S—this is called the ***conjugate point*** of N.

As the particles converge and intersect at S (where $\boldsymbol{\xi} = \mathbf{0}$) they achieve maximum relative speed, just as our oscillating weight achieves maximum speed as it passes through the equilibrium point, $\boldsymbol{\xi} = \mathbf{0}$.

This journey from N to S is only one-*half* of a complete $\boldsymbol{\xi}$-oscillation, the remaining half being completed on the back side of the sphere, with the direction of $\boldsymbol{\xi}$ now reversed, as illustrated, switching over from pointing to the left of p's trajectory to its right.

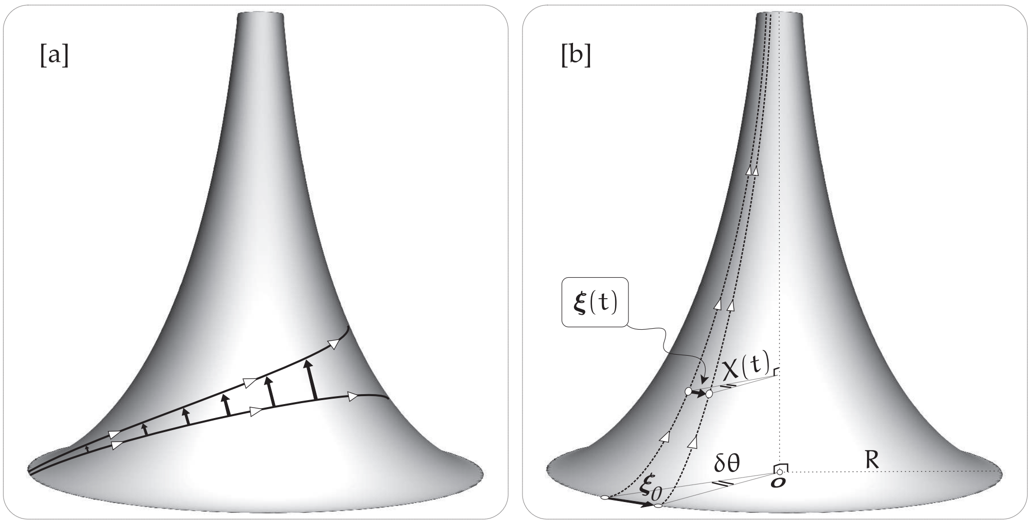
Once the particles have returned to N, the $\boldsymbol{\xi}$-oscillation begins anew, and repeats forever.

28.1.3 Negative Curvature: The Pseudosphere

Just as geodesics are attracted to each other on the positively curved sphere, so they are repelled by each other on the negatively curved pseudosphere. To make this especially clear, [28.4a] shows two neighbouring geodesics being launched from the rim in the "*same*" direction (in the sense that $\dot{\boldsymbol{\xi}}(0) = \mathbf{0}$). Soon, however, the repulsive negative curvature makes itself felt, and we see the geodesics flying apart.

But in order to *calculate* the relative acceleration with ease, we shall instead consider [28.4b], which shows two neighbouring particles on the rim being launched straight up the pseudosphere at unit speed, so that they are *approaching* each other as they travel along their respective tractrix generators, the arc length being equal to the time: $\sigma = t$.

The initial velocities of the particles are horizontal, both pointing at the centre o of the rim, distance R away. Since they are travelling at unit speed, in the absence of force they should therefore collide at time $t = R$. But clearly they don't collide, ... ever! Yes, the two tractrix generators



[28.4] **[a]** *Two neighbouring geodesics start off in the same direction, but the negative curvature forces them apart.* **[b]** *The connecting vector* $\boldsymbol{\xi}$ *between neighbouring tractrix generators is subject to* $\dot{\boldsymbol{\xi}} = -\left(\frac{1}{R}\right)\boldsymbol{\xi}$*, and so* $\ddot{\boldsymbol{\xi}} = +\left(\frac{1}{R^2}\right)\boldsymbol{\xi} = -\mathcal{K}\,\boldsymbol{\xi}$.

approach each other, but their approach speed *reduces* as they travel up the pseudosphere, thanks to the repulsive effective force arising from the negative curvature.

If $\delta\theta$ is the illustrated angle through which one of these generators must be rotated about the symmetry axis to obtain the other, then their initial small separation will be $|\boldsymbol{\xi}_0| \asymp R\,\delta\theta$. If $X(t)$ denotes the distance of the particles from the symmetry axis at time t, then

$$|\boldsymbol{\xi}(t)| \asymp X(t)\,\delta\theta \asymp \left[\frac{|\boldsymbol{\xi}_0|}{R}\right] X(t). \tag{28.3}$$

Next, we remind the reader of the defining geometrical property of the tractrix, illustrated in [28.5] (which merely reproduces our original figure, [5.2], page 52). This figure implies that

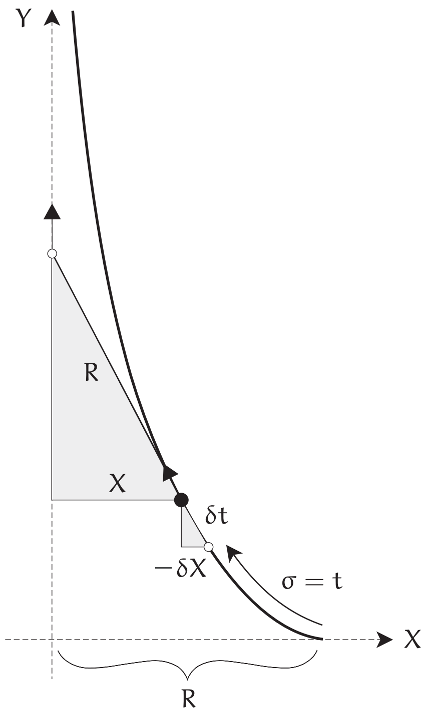
$$\frac{-dX}{dt} = \frac{X}{R}.$$

This equation leads *directly* to the Jacobi Equation, without even needing to solve it first!

To see how, we rephrase the equation in terms of the relative velocity of our particles: (28.3) implies

$$\dot{\boldsymbol{\xi}} = -\left(\frac{1}{R}\right)\boldsymbol{\xi}.$$

Thus the initial approach speed is $|\boldsymbol{\xi}_0|/R$, and therefore (as we observed earlier) in the absence of the force the initial gap $|\boldsymbol{\xi}_0|$ would shrink to zero in time $t = R$.



[28.5] **The Tractrix** *is defined by the property that the distance along its tangent to a fixed line has constant length,* R.

The fact that the particles do *not* in fact collide is *due to the repulsive force arising from the negative curvature*. This is confirmed by differentiating a second time, once again arriving at Jacobi's Equation Of Geodesic Deviation:

$$\ddot{\boldsymbol{\xi}} = -\left(\frac{1}{R}\right)\dot{\boldsymbol{\xi}} = +\left(\frac{1}{R^2}\right)\boldsymbol{\xi} = -\mathcal{K}\,\boldsymbol{\xi}.$$

The last equality follows from the fact that the pseudosphere has constant negative curvature, $\mathcal{K} = -(1/R^2)$.

The above argument can easily be generalized to prove that the Jacobi Equation holds on *all* surfaces for revolution, but we leave this for you to confirm, in Exercise 8. Instead, we now move on to establish the result in the *general* case, for a surface without any symmetries.

28.2 Two Proofs of the Jacobi Equation

28.2.1 *Geodesic Polar Coordinates*

Figure [28.3] may be viewed as providing a novel interpretation of ordinary spherical polar coordinates on $\mathbb{S}^2$.

We simultaneously launch particles in all possible directions from the north pole, so that every point of $\mathbb{S}^2$ (other than the poles) is uniquely specified by *which* particle hits it, and *when*. In greater detail, *which* particle hits the point can be specified by the longitude angle θ that the launch direction makes with some particular[2] (but arbitrary) direction, defined to be $\theta = 0$. *When* the particle hits the point is specified by the latitude angle $\phi = \sigma = t$ (assuming now that $R = 1$).

Geodesic polar coordinates are the natural generalization of this idea to an arbitrary surface, $\mathcal{S}$. See [28.6], which illustrates the idea in the case that $\mathcal{S}$ is a torus. From an arbitrary point[3] o on $\mathcal{S}$, launch particles at unit speed in all directions, so that they travel outward along geodesics, and choose one of these to be $\theta = 0$. Provided the region surrounding o is not so large that it contains any conjugate points, then every point within it is hit by a unique particle at a unique time. If that particle is launched in direction θ and hits the point at time t, then we can assign to that point the unique ***geodesic polar coordinates*** (t, θ).

Both in the plane and on the sphere, we know that the intrinsic circle $K(\sigma)$ with intrinsic radius $\sigma = t =$ *const.* is *orthogonal to its "radii,"* i.e., the geodesic segments of length σ emanating from its centre o. It turns out that this remains true on a *general* surface, as illustrated in [28.6]. Gauss (1827, §15) proved this on his way to proving other results, and so it has become known as

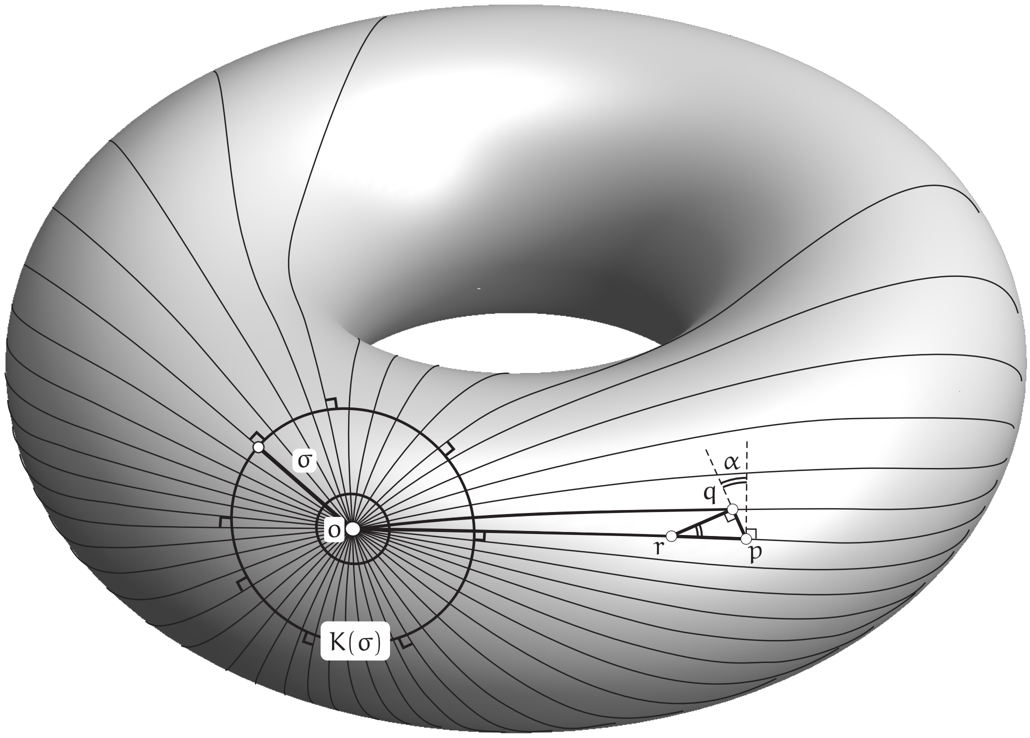
Gauss's Lemma. *If particles are launched in all directions from a point on a general surface, travelling out a distance σ along geodesics, they will form (by definition) a* geodesic circle $K(\sigma)$ *with intrinsic radius* σ. This geodesic circle $K(\sigma)$ cuts its geodesic radii at right angles. (28.4)

Vector Calculus would not be developed by Gibbs and Heaviside[4] till the 1880s, so in 1827 Gauss was forced to prove this by means of a full page of computation. (A much shorter and

[2]On planet Earth we call $\theta = 0$ the ***prime meridian***, and define it to be the one passing through the Royal Observatory at Greenwich, London, England. Since 1999, this direction has been marked by particles of *light* (from a powerful green laser) launched northward across the London night sky!

[3]We now drop the hat for points on $\mathcal{S}$, because we no longer need to distinguish them from points in the map.

[4]See the *Chronology* in Crowe (1985, pp. 256–259)



[28.6] Gauss's Lemma: A geodesic circle $K(\sigma)$ of intrinsic radius σ cuts its geodesic radii at right angles. *Gauss's second, geometrical proof of this result assumes that the two neighbouring points* p *and* q *lie on* $K(R)$*, and yet* $\alpha \neq 0$*. As explained in the text, this leads to a contradiction.*

more transparent version of this now-standard computation—based upon the *extrinsic* fact that the acceleration of the geodesics is normal to the surface—can be found in Ex. 9.) But immediately after completing his computational proof, Gauss did something completely uncharacteristic: he offered a *second* proof that was instead *intuitive and geometrical*. Furthermore, this second proof was *intrinsic*.

We can only speculate that Gauss did this because he thought this second proof was especially important, or because he was especially proud of his geometrical ingenuity. In any event, here is the evidence, in Gauss's own words: "We have thought it worthwhile to deduce this theorem from the fundamental property of shortest lines; but the truth of the theorem can be made apparent without any calculation by means of the following reasoning."

We shall now present this reasoning, which appears to have been completely lost in the mists of time—indeed, we have yet to discover a single instance of its reincarnation in any modern text. This is perhaps due to the fact that Gauss expressed his reasoning in terms of *infinitesimals*, thereby wrongfooting modern mathematicians. In place of Gauss's infinitesimals, we shall instead employ Newtonian *ultimate equalities* between small, ultimately vanishing quantities. Also, in contrast to the ever-unsolicitous Gauss (1827, §15), we have supplied a *diagram* in order to *see* what Gauss meant!

Suppose we launch two neighbouring geodesics from o, the small (ultimately vanishing) angle between them being $\delta\theta$. These are the two darker geodesic segments shown in [28.6] that terminate at p and q. Suppose that p and q both lie on $K(R)$—i.e, the geodesic segments op and oq have *the same length* $\sigma = R$. But now imagine (as illustrated) that pq is *not* ultimately orthogonal to op and oq (as $\delta\theta \to 0$), contrary to Gauss's Lemma. Following Gauss, we now show that this is impossible, thereby proving (28.4).

Suppose that $\angle \mathrm{opq} \asymp (\pi/2) - \alpha$, as illustrated. Now draw the geodesic through q orthogonal to pq, and let its intersection with op be r, as shown. Then $\angle \mathrm{prq} \asymp \alpha$, and therefore $\mathrm{rq} \asymp \mathrm{rp} \cos \alpha$. (Here, and in the following, "rq" is serving double duty, standing for both the segment, and the *length* of the segment, the context making clear which is meant.)

Therefore,

$$\begin{aligned}\text{length of } o \rightsquigarrow r \rightsquigarrow q &= \mathrm{or} + \mathrm{rq} \\ &\asymp (R - \mathrm{rp}) + \mathrm{rp} \cos \alpha \\ &= R - \mathrm{rp}(1 - \cos \alpha) \\ &< R.\end{aligned}$$

Thus this indirect route from o to q (via r) is *shorter than the shortest route*, the direct geodesic route, oq. Gauss has his contradiction!

Having established the orthogonality of the (t, θ) coordinates, the metric takes the form

$$d\widehat{s}^2 = dt^2 + \rho^2(t, \theta)\, d\theta^2.$$

Thus if $\delta\theta$ is the angular separation of two neighbouring geodesics launched from o, then

$$|\boldsymbol{\xi}| \asymp \rho\, \delta\theta \quad \Longrightarrow \quad |\ddot{\boldsymbol{\xi}}| \asymp \ddot{\rho}\, \delta\theta. \tag{28.5}$$

But if we substitute $u = t$, $v = \theta$, $A = 1$, and $B = \rho$ into (27.1), we obtain

$$\begin{aligned}\mathcal{K} &= -\frac{1}{AB}\left(\partial_v\left[\frac{\partial_v A}{B}\right] + \partial_u\left[\frac{\partial_u B}{A}\right]\right) \\ &= -\frac{1}{\rho}\left(\partial_\theta\left[\frac{\partial_\theta 1}{\rho}\right] + \partial_t\left[\frac{\partial_t \rho}{1}\right]\right) \\ &= -\frac{\ddot{\rho}}{\rho}.\end{aligned} \tag{28.6}$$

Combining this result with (28.5), we have proved the Jacobi Equation, (28.2):

$$\ddot{\boldsymbol{\xi}} = -\mathcal{K}\, \boldsymbol{\xi}.$$

Finally, we remark that (28.6) also allows us to finally prove ***Minding's Theorem*** (alluded to on p. 21), which states that if two surfaces have the same *constant* curvature, then they are locally isometric. (See Ex. 7 for the details.)

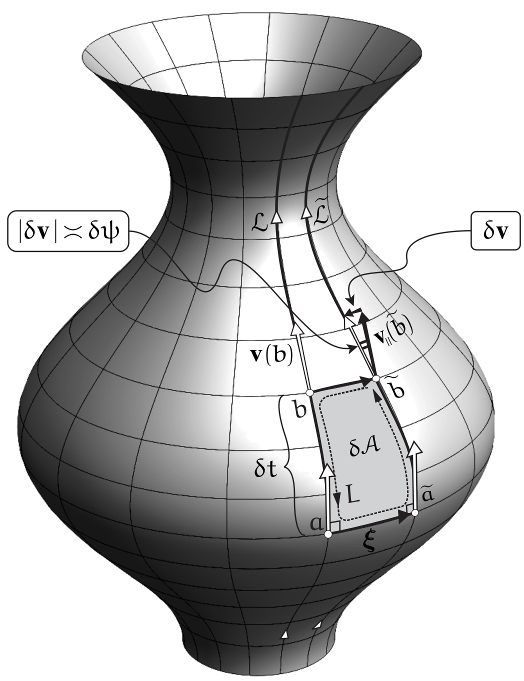
28.2.2 *Relative Acceleration = Holonomy of Velocity*

While our first proof of the Jacobi Equation was extremely brief (once Gauss's Lemma was established), it did require the full force of the ("*Star Trek* phaser") formula, (27.1). As we shall now see, there was actually no need to "set phasers to kill"! In its place, we now present a second proof that is *direct, geometrical, and intuitive*.

Indeed, the title of this subsection *is* that proof, albeit in aphoristic form. We now spell out the details, using [28.7], which uses the same notation as in [28.2]. This figure reuses the surface from our proof of Clairaut's Theorem, [11.7], but seen now with new eyes. Specifically, we see that the meridian geodesic generators $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ (through a and the neighbouring point $\widetilde{a}$) are attracted to each other initially in the region of positive curvature, but are then repelled from each as they enter the negative curvature region in the neck of the vase. (NOTE: The following argument is completely general, and does not rely on or make any use of the symmetry of this particular surface.)

Let $\boldsymbol{\xi}$ connect two neighbouring geodesics, $\mathcal{L}$ and $\widetilde{\mathcal{L}}$, launched with unit speed, perpendicular to a short geodesic segment $a\widetilde{a}$, as illustrated. This ensures that $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ start out parallel, in the sense that initially they remain the same distance apart, with zero rate of separation: $\dot{\boldsymbol{\xi}} = \mathbf{0}$. (NOTE: We have drawn the separation fairly large in order to make the subsequent geometric reasoning easier to follow, but ultimately we will let the gap $a\widetilde{a}$ go to zero, so $\boldsymbol{\xi}$ will ultimately be a genuine tangent vector that can be pictured as lying *within* the surface.)

After time δt both particles will have travelled distance δt across the surface, arriving at b and $\widetilde{b}$, respectively. By then the relative velocity $\dot{\boldsymbol{\xi}}$ between $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ has increased from zero to $\delta\mathbf{v} \asymp \ddot{\boldsymbol{\xi}}\,\delta t$, by virtue of the definition, (28.1).



[28.7] Geometric Proof of the Jacobi Equation. *Parallel transporting* $\mathbf{v}_{||}(\widetilde{b})$ *around* L *returns it to* $\widetilde{b}$ *rotated by the holonomy* $\delta\psi = \mathcal{R}(L) \asymp \mathcal{K}\,\delta\mathcal{A} \asymp \mathcal{K}\,|\boldsymbol{\xi}|\,\delta t$. *But* $\delta\psi \asymp |\delta\mathbf{v}| \asymp |\ddot{\boldsymbol{\xi}}|\,\delta t$, *and the Jacobi Equation follows.*

Presumably $\delta\mathbf{v}$ should here be interpreted as the difference $[\widetilde{\mathbf{v}}(\widetilde{b}) - \mathbf{v}(b)]$ between the new velocities. However, this does not make *intrinsic* sense, since we cannot subtract tangent vectors to $\mathcal{S}$ that emanate from different points and lie with different tangent planes. In order to compare them, let us join b to $\widetilde{b}$ with a short geodesic[5] segment $b\widetilde{b}$ and then parallel transport $\mathbf{v}(b)$ along $b\widetilde{b}$ to obtain $\mathbf{v}_{||}(\widetilde{b})$ at $\widetilde{b}$, as illustrated. We thereby obtain an *intrinsic* measure of the relative acceleration:

$$\boxed{\ddot{\boldsymbol{\xi}}\,\delta t \asymp \delta\mathbf{v} = \widetilde{\mathbf{v}}(\widetilde{b}) - \mathbf{v}_{||}(\widetilde{b}).}$$

As illustrated, let $\delta\psi$ be the angle between these two vectors at $\widetilde{b}$. Since the arc (not shown) of the unit circle centred at $\widetilde{b}$ through the tips of these vectors is ultimately equal to the chord connecting those tips, we see that

$$|\delta\mathbf{v}| \asymp \delta\psi.$$

Now we come to the crux of the matter. As illustrated, let us define L to be the counterclockwise loop around the boundary of the shaded quadrilateral: $L \equiv \widetilde{b}ba\widetilde{a}\widetilde{b}$. Starting at $\widetilde{b}$, let us parallel transport $\mathbf{v}_{||}(\widetilde{b})$ around L.

First, since $\mathbf{v}_{||}(\widetilde{b})$ was originally obtained by parallel transporting $\mathbf{v}(b)$ along $b\widetilde{b}$, it follows that parallel transporting it back again along $\widetilde{b}b$ will return it to its original state, $\mathbf{v}(b)$. Next, since the velocity of a geodesic is parallel transported along itself, parallel transport along ba yields $\mathbf{v}(a)$ at a. Next, parallel transport along the geodesic $a\widetilde{a}$ maintains orthogonality with it, so it arrives at $\widetilde{a}$ as $\widetilde{\mathbf{v}}(\widetilde{a})$. Finally, parallel transport along the geodesic $\widetilde{a}\widetilde{b}$ returns it to its starting point as $\widetilde{\mathbf{v}}(\widetilde{b})$, so parallel transport around L has rotated $\mathbf{v}$ by $\delta\psi$.

[5]This choice makes the argument easier to follow, but we will ultimately take the limit that the segment shrinks to zero, so the precise nature of the segment does not matter.

In other words,

> *The relative acceleration of the geodesics is measured by the rotation $\delta\psi$ of the velocity vector, but this is none other than the* holonomy *of the velocity when it is parallel transported around* L. *In brief,* $\delta\psi = \mathcal{R}(L)$*!*

Since the area $\delta\mathcal{A}$ inside L is given by $\delta\mathcal{A} \asymp |\boldsymbol{\xi}|\,\delta t$, the proof of the Jacobi Equation comes down to the following sequence of ultimate equalities, each of which should now be visually apparent:

$$|\ddot{\boldsymbol{\xi}}|\,\delta t \asymp |\delta\mathbf{v}| \asymp \delta\psi = \mathcal{R}(L) \asymp \mathcal{K}\,\delta\mathcal{A} \asymp \mathcal{K}|\boldsymbol{\xi}|\,\delta t.$$

Cancelling δt, and taking into account the fact that $\delta\mathbf{v}$ points in the *opposite* direction to $\boldsymbol{\xi}$ for the illustrated case of *positive* $\mathcal{K}$, we have arrived again at the Jacobi Equation:

$$\ddot{\boldsymbol{\xi}} = -\mathcal{K}\,\boldsymbol{\xi}.$$

28.3 The Circumference and Area of a Small Geodesic Circle

In Section 2.2 we showed that, on the sphere, the inhabitants of the surface could determine the curvature of their world by examining carefully either the circumference $C(r)$ or the area $\mathcal{A}(r)$ of a small circle $K(r)$ of intrinsic radius r. In both cases, the key to finding the curvature was to detect how these quantities *departed from* their Euclidean values of $C(r) = 2\pi r$ and $\mathcal{A}(r) = \pi r^2$.

Specifically, we proved that in the limit that r vanishes, the curvature is given by (2.4) and (2.5), which we repeat here, for the convenience of the reader:

$$\boxed{\mathcal{K} \asymp \frac{3}{\pi}\left[\frac{2\pi r - C(r)}{r^3}\right],} \tag{28.7}$$

and

$$\boxed{\mathcal{K} \asymp \frac{12}{\pi}\left[\frac{\pi r^2 - \mathcal{A}(r)}{r^4}\right].} \tag{28.8}$$

We claimed at the time that these formulas apply to *all* surfaces, not merely the sphere, and we are now finally in a position to prove this.

With the same geodesic polar coordinates as before, let $\xi(r)$ denote the separation of the geodesics launched at angle θ_0 and $\theta_0 + \delta\theta$, so that if we define

$$g(r) \equiv \rho(\theta_0, r) \qquad \text{then} \qquad \xi(r) \asymp g(r)\,\delta\theta.$$

As $r \to 0$, we know this reduces to the Euclidean formula, $\xi(r) \asymp r\,\delta\theta$, but it is precisely the small *departure* from this Euclidean result that we must detect in order to measure the non-Euclideanness, aka the curvature.

To that end, we expand $g(r)$ into a Maclaurin series:

$$g(r) = r + \frac{1}{2} g''(0)\, r^2 + \frac{1}{6} g'''(0)\, r^3 + \cdots .$$

But (28.6) tells us that $g''(r) = -\mathcal{K} g(r)$, so $g''(0) = 0$, and also

$$g'''(0) = [-\mathcal{K} g]'(0) = -\mathcal{K}'(0) g(0) - \mathcal{K}(0) g'(0) = -\mathcal{K}(0).$$

Thus,

$$\boxed{C(r) = 2\pi\, g(r) = 2\pi r - \tfrac{\pi}{3}\, \mathcal{K}(0)\, r^3 + \cdots ,}$$

and (28.7) follows immediately.

Finally, the area $\delta\mathcal{A}$ of an annulus of geodesic width δr is given by

$$\delta\mathcal{A} \asymp C(r)\, \delta r,$$

so integration of the formula for $C(r)$ readily yields [exercise] the area formula, (28.8).

Chapter 29

Riemann's Curvature

29.1 Introduction and Summary

In this chapter we shall see how the forgoing insights concerning 2-dimensional surfaces can be extended naturally to n-dimensional spaces, called ***manifolds***.

Just as our 2-dimensional surfaces are locally described by their tangent planes, each of which has the structure of $\mathbb{R}^2$, so the immediate vicinity of a point in an n-dimensional manifold locally resembles $\mathbb{R}^n$, but distances between neighbouring points are measured with a metric[1] that is not Euclidean.

• **First**, we shall extend to an ***n-manifold*** (shorthand for an n-dimensional manifold) our very first measure of the intrinsic curvature $\mathcal{K}(p)$ of a 2-surface, namely, the ***local angular excess per unit area*** of a small geodesic triangle Δ as it shrinks to a point p:

$$\mathcal{K}(p) = \lim_{\Delta \to p} \frac{\mathcal{E}(\Delta)}{\mathcal{A}(\Delta)}.$$

• **Second**, it turns out that the angular excess is too blunt an instrument to give direct insight into the more subtle curvature(s!) of an n-manifold. However, the *holonomy* resulting from parallel-transporting a vector around a small shrinking loop *is* able to fully reveal this more intricate curvature structure, in a very direct way. Clearly, then, the next step towards defining the curvature of an n-manifold must be to understand ***parallel transport*** in such a space.

Generalizing parallel transport from 2-surfaces to n-manifolds is *nontrivial*, and yet it is given scant attention in almost every standard text we have examined. To redress the balance, we shall provide *three* different geometrical constructions (all leading to the same result) that can be used to generalize Levi-Civita's parallel transport to such an n-manifold.

• **Third**, we use parallel transport to define the ***intrinsic*** (aka ***"covariant"***) ***derivative*** within an n-manifold. The good news is that—when expressed in *intrinsic* terms—the passage from a 2-surface to an n-manifold requires absolutely no change to our original definition (23.2); only the *notation* changes, writing $\nabla_{\mathbf{v}}$ in place of $D_{\mathbf{v}}$.

• **Fourth**—and this is the beating heart of the chapter—we use parallel transport to generalize holonomy from 2-dimensional surfaces to n-manifolds.

Riemann, pictured in [29.1], discovered that in place of Gauss's single number $\mathcal{K}$ characterizing a 2-surface, his generalized *intrinsic* curvature of an n-manifold is specified by an array of[2]

---

[1]The fully general conception of a manifold does not actually require that distance be defined within it. For our purposes, however, the metric is *the* central structure to be investigated, so we must insist that *our* manifolds have metrics! The technical name for a manifold with a positive-definite metric is a ***Riemannian manifold***, and a manifold with metric that can yield *negative* ("distance")^2, as is the case for spacetime, is called ***pseudo-Riemannian***. Rather than constantly stating explicitly that our manifolds are either Riemannian or pseudo-Riemannian, we shall merely let it be *understood* that our manifolds have metrics: they will *always* be either Riemannian or pseudo-Riemannian.

[2]This is proved in Exercise 11.

$$\boxed{\frac{1}{12}n^2(n^2-1)} \qquad (29.1)$$

distinct curvature components: these form the numerical description of a geometrical object called the ***Riemann tensor***.

In a 2-dimensional surface the Riemann tensor therefore reduces (with $n=2$) to a single component, and this is simply $\mathcal{K}$. The room in which you now sit *appears* to be 3-dimensional, and therefore the space around you (with $n=3$) is described by six curvature components. However, as you sit still in your chair, you are actually hurtling into the future, along the fourth dimension, *time!* In Einstein's 4-dimensional, curved spacetime, there are 20 curvature components, and, as we shall see in the next chapter, these describe the gravitational field!

[29.1] *Bernhard Riemann (1826–1866).*

• **Fifth**, we shall see how to generalize the ***Jacobi Equation*** to an n-manifold.

• **Sixth**, we shall describe a particularly important, geometrically meaningful *average* of Riemann's curvatures, called the ***Ricci curvature tensor***. In Einstein's curved, 4-dimensional spacetime, this captures exactly *half* of the complete curvature information (10 components out of 20 total).

29.2 Angular Excess in an n-Manifold

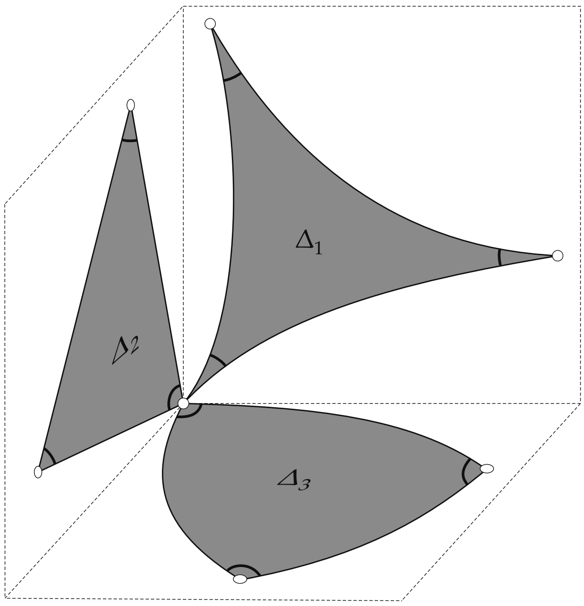
Our study of 2-surfaces has provided the essential springboard to understand n-manifolds. But a 2-surface is simply too special to manifest the full variety of phenomena that can and do occur in n-manifolds.

However, very happily, it turns out that essentially *all*[3] *the fundamentally new features and concepts of n-manifolds are revealed if we merely go up* one *dimension, to a readily-visualizable* **3***-manifold*, such as the space you currently sit in. We shall therefore focus our attention initially on this very concrete case, but in the next chapter we shall increase the dimension yet again, from three to four, in order to understand Einstein's curved spacetime.

We have gained many important insights into our 2-surfaces by viewing them as embedded in $\mathbb{R}^3$. And while it is certainly possible to conceive of our 3-manifold as likewise embedded in a higher dimensional space, there is no longer any visual advantage to doing so. Furthermore, with our eye on the prize of understanding Einstein's curved spacetime, our focus will be solely on *intrinsic* properties of an n-manifold, as determined by its intrinsic metric. Therefore, *henceforth, we shall always assume that we are creatures living within the n-manifold, and* only *intrinsic measurements within it are meaningful.*

[3]There are a few phenomena that only appear once we reach dimension $n=4$. For example, the vitally important Weyl curvature tensor, described in Exercise 15, only exists if $n \geqslant 4$.

Our very first measure of intrinsic curvature was via the angular excess $\mathcal{E}(\Delta)$ of a geodesic triangle Δ. But to construct a geodesic triangle Δ, the first thing we need are the geodesics themselves! On a 2-surface, we constructed a geodesic by stretching a string over[4] the surface. In a 3-manifold, we simply stretch the string between two points, thereby obtaining a seemingly straight line between them. Alternatively, we may shoot a laser beam from one point to the other. But we no longer assume that our 3-manifold is the Euclidean $\mathbb{R}^3$; *instead, it is equipped with a general, non-Euclidean metric*, so stetched strings and laser beams can no longer be relied upon to behave as expected. Indeed, as we shall discuss in the next chapter, the bending of "straight" light rays was one of the first experimental confirmations (in 1919) of Einstein's prediction that the geometry of physical space is curved.



[29.2] *Geodesic triangles constructed in three perpendicular coordinate planes of a 3-manifold reveal only* part *of the curvature structure of the space. Here, $\mathcal{E}(\Delta_1) < 0$, $\mathcal{E}(\Delta_2) = 0$, and $\mathcal{E}(\Delta_3) > 0$.*

In a 2-surface, there was only one possible choice for the plane of our Δ. But in our 3-manifold, there are clearly infinitely many planes to choose from, one for each choice of the normal vector. The curvature within such a slice of space is called a ***sectional curvature***, and is the subject of Section 29.5.8. Figure [29.2] illustrates three such sectional curvatures in three orthogonal coordinate planes in a 3-manifold: they are all independent, both as regards their magnitudes and their signs. Here we have illustrated three geodesic triangles Δ_i with a common vertex, but each lying in a different coordinate plane. As illustrated, here $\mathcal{K}(\Delta_1) < 0$, $\mathcal{K}(\Delta_2) = 0$, and $\mathcal{K}(\Delta_3) > 0$.

But we have already stated that the Riemann tensor of a 3-manifold has six curvature components, so the angular excess in these three coordinate planes has only been able to detect *half* of the curvature structure. Things only get worse as the dimension increases: in a 4-manifold, the angular excess in the six different coordinate planes only yields six curvatures, whereas the Riemann tensor has 20 curvature components.

To explore curvature, in all its glory, it is essential that we instead turn to holonomy. To that end, we must now grapple with the *nontrivial* task of generalizing parallel transport from a 2-surface to an n-manifold. By virtue of its fundamental importance, we now offer not one, but *three* different geometrical constructions!

29.3 Parallel Transport: Three Constructions

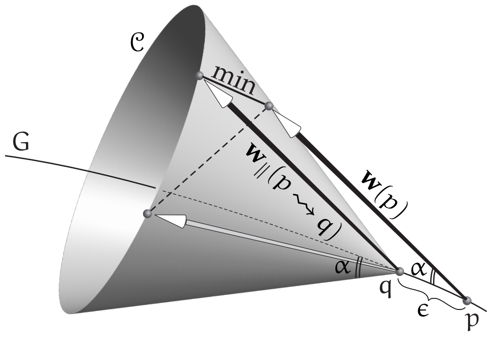
29.3.1 Closest Vector on Constant-Angle Cone

Recall that in order to parallel transport a vector $\mathbf{w}(p)$ along a curve in a 2-surface, the fundamental *intrinsic* construction takes this curve to be a *geodesic* G. (To parallel transport along a *general*

[4]To cope with negative curvature, recall that we imagined the string trapped between two parallel layers of surface.

curve, we break it down into geodesic segments.) To achieve parallel transport along G, it is only necessary to *keep the angle* α *between* $\mathbf{w}_{||}$ *and* G *constant* (and keep the length of $\mathbf{w}_{||}$ constant).

In a 3-manifold, this construction immediately runs into trouble. See [29.3]. Suppose we try to carry $\mathbf{w}(p)$ a short distance ϵ along G from p to q to create the parallel vector $\mathbf{w}_{||}(p \rightsquigarrow q)$, and all we know is that α remains constant, and the length $\mathbf{w}_{||}$ remains constant. Then $\mathbf{w}_{||}(p \rightsquigarrow q)$ could lie anywhere on the illustrated *cone* $\mathcal{C}$ of directions; which generator of $\mathcal{C}$ should we pick?



[29.3] *Suppose that* $\mathbf{w}(p)$ *makes angle* α *with the geodesic* G*. To parallel-transport* $\mathbf{w}(p)$ *distance* ϵ *along the geodesic* G *to* q*, construct the cone* $\mathcal{C}$ *of vectors at* q *making angle* α *with* G*, then pick the one whose tip is closest to the tip of* $\mathbf{w}(p)$*.*

For this construction, and for each of the other constructions that follow, it suffices to ask ourselves how we would do the analogous construction in $\mathbb{R}^3$, for as $\epsilon \to 0$ the two constructions must agree. Well, in $\mathbb{R}^3$, the generator of $\mathcal{C}$ that is parallel to $\mathbf{w_o}$ is the one that is *closest*. That is, as illustrated,

$\mathbf{w}_{||}(p \rightsquigarrow q)$ is *the* generator of $\mathcal{C}$ whose tip is *closest* to the tip of $\mathbf{w}(p)$.

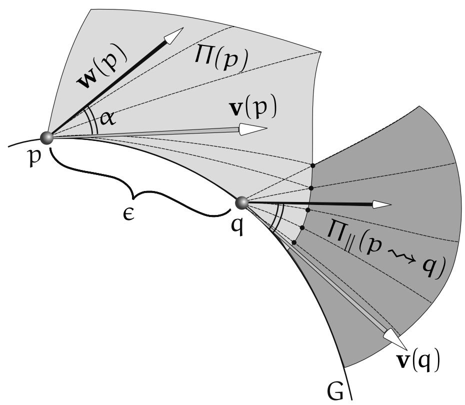
By repeating this process over and over, we may parallel-transport $\mathbf{w}(p)$ as far as we wish along G.

29.3.2 Constant Angle within a Parallel-Transported Plane

The next construction, which we have only ever seen described by Arnol'd (1989, pp. 305–306), again draws its inspiration from $\mathbb{R}^3$.

To parallel-transport $\mathbf{w}(p)$ along a Euclidean straight line G in $\mathbb{R}^3$, with tangent $\mathbf{v}$, let $\Pi(p)$ be the plane spanned by $\mathbf{w}(p)$ and $\mathbf{v}(p)$, the angle between the two vectors being α, as before. Now parallel-transport $\Pi(p)$ along G to obtain $\Pi_{||}$. Finally, parallel-transport $\mathbf{w}(p)$ along G as the unique vector in $\Pi_{||}$ making angle α with $\mathbf{v}$.

Of course in $\mathbb{R}^3$ this is a perverse waste of time, for we have an absolute parallelism that allows us to move $\mathbf{w}(p)$ along G without the assistance of $\Pi_{||}$. *However*, this now suggests a new construction in a curved 3-manifold (or n-manifold). See [29.4].



[29.4] *We construct the "plane"* $\Pi(p)$ *by launching geodesics from* p *in all directions spanned by* $\mathbf{w}(p)$ *and* $\mathbf{v}(p)$*. We then (approximately) parallel-transport* $\Pi(p)$ *to* q *to form* $\Pi_{||}(p \rightsquigarrow q)$*, generated by geodesics emanating from* q *and passing through neighbouring points of* $\Pi(p)$*.*

In $\mathbb{R}^3$, $\Pi(p)$ was made up of all the straight line emanating from p whose directions were linear combinations of $\mathbf{v}(p)$ and $\mathbf{w}(p)$. By analogy, in our curved 3-manifold, let $\Pi(p)$ now denote the "plane" made up of all the *geodesics* emanating from p whose directions are linear combinations of $\mathbf{v}(p)$ and $\mathbf{w}(p)$; of course this "plane" is actually a *curved 2-surface*. Note that this surface $\Pi(p)$ necessarily contains G, the geodesic in the direction $\mathbf{v}(p)$.

As in the previous construction, let q lie ϵ along G. By construction, $\Pi(p)$ necessarily contains q *and* the new direction of G at q, namely $\mathbf{v}(q)$. If we zoom in on the small region of $\Pi(p)$ surrounding q, it looks like a Euclidean plane. We can now construct a new surface $\Pi_{||}(p \rightsquigarrow q)$—the parallel translation of $\Pi(p)$ to q—by launching geodesics out from q in all directions within this small planar region of $\Pi(p)$ surrounding q. Figure [29.4] illustrates a few of these geodesics (dashed), launched from q and passing through nearby (black dot) points of $\Pi(p)$. To be more precise, the geodesic generators of $\Pi_{||}(p \rightsquigarrow q)$ are obtained by taking the *limit* that the black dots are pulled into q.

Repeating this construction over and over, we may approximate parallel-transport of $\Pi(p)$ down the length of G. Finally, letting $\epsilon \to 0$, we obtain a continuously varying, parallel-transported 2-surface $\Pi_{||}$ along the length of G, and then,

> *To parallel transport $\mathbf{w}(p)$ along the geodesic G, we keep its length constant, and maintain its angle α with G, while always keeping it tangent to the parallel-transported 2-surface $\Pi_{||}$.*

29.3.3 Schild's Ladder

Our final construction is also the simplest. It was first stated by Alfred Schild (1921–1977) in an (unpublished) Princeton lecture in 1970, and was later named ***Schild's Ladder*** in Misner, Thorne, and Wheeler (1973).

On a 2-surface, it has often been helpful for us to imagine that a very short tangent vector (that actually lives in the $\mathbb{R}^2$ tangent plane to the surface) instead lies *in* the surface. Likewise, in an n-manifold, a very short vector in the $\mathbb{R}^n$ tangent space can be imagined to be *in* the manifold. Therefore, since parallel-transport (by definition) preserves length, if we wish to parallel-transport a long tangent vector at p, we may first *shrink it by some large factor* N in the $\mathbb{R}^n$ tangent space at p, then picture it as living *in* the manifold, then parallel-transport it, then finally *expand it by the same large factor* N in the $\mathbb{R}^n$ tangent space at the destination, thereby restoring its original length.

Schild's Ladder construction uses this freedom to imagine that the vector $\mathbf{w}(p)$ that is to be parallel-transported along the geodesic G is very short, and can therefore be pictured as a very short geodesic segment *in* the manifold. With this understanding, the construction of *Schild's Ladder* is spelled out in the caption of [29.5], which illustrates it.

29.4 The Intrinsic (aka "Covariant") Derivative $\nabla_{\mathbf{v}}$

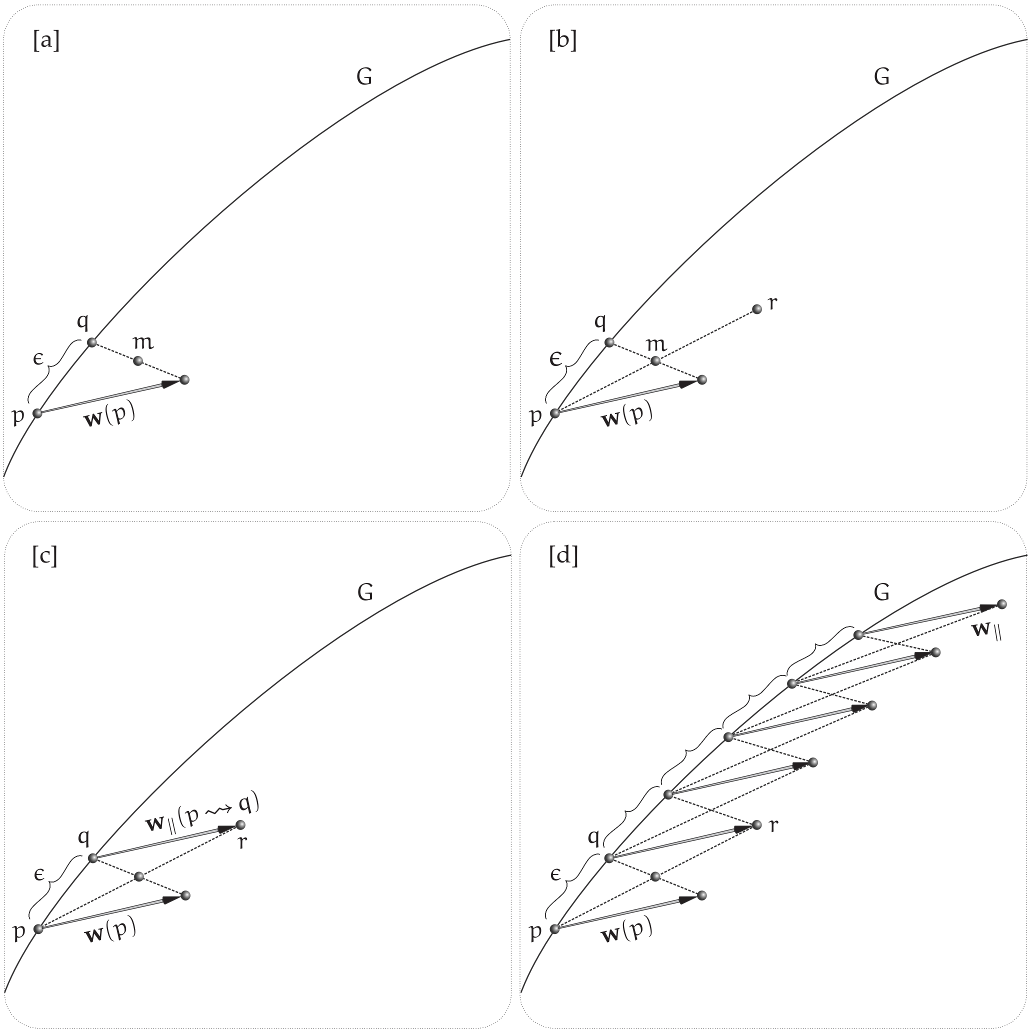
Previously, we used $\nabla_{\mathbf{v}}$ to denote the directional derivative in $\mathbb{R}^3$, and we used $D_{\mathbf{v}}$ to denote the intrinsic derivative within the 2-surface. In *extrinsic* terms, we saw in (23.3) that $D_{\mathbf{v}}$ could be viewed as the projection $\mathcal{P}$ of $\nabla_{\mathbf{v}}$ into the surface, along its normal $\mathbf{n}$:

$$D_{\mathbf{v}}\mathbf{w} = \mathcal{P}\left[\nabla_{\mathbf{v}}\mathbf{w}\right] = \nabla_{\mathbf{v}}\mathbf{w} - (\mathbf{n} \bullet \nabla_{\mathbf{v}}\mathbf{w})\,\mathbf{n}.$$

However, as we have noted, we shall *not* consider our n-manifold to be embedded in a higher-dimensional space, so no analogue of the above formula will be sought or found. Instead, we will think entirely *intrinsically*.

Fortunately, we already have an intrinsic definition of this intrinsic derivative in (23.2), page 242, and we even drew a picture of it in [23.3]. Let us briefly review that construction.

Our task is to find the rate of change of a vector field $\mathbf{w}$ as we move away from a point p in the direction of a unit vector $\mathbf{v}$. We move a small (ultimately vanishing) distance ϵ away from p along $\epsilon\mathbf{v}$, arriving at q. To see how much the new vector $\mathbf{w}(q)$ has changed from its original value $\mathbf{w}(p)$,



[29.5] Schild's Ladder. **[a]** *Travel distance* ϵ *along the geodesic* G *to construct the point* q. *Connect the tip of* $\mathbf{w}(p)$ *to* q *with a geodesic segment, and mark its midpoint* m. **[b]** *Connect* p *to* m *with another geodesic segment, and then extend it an equal distance beyond* m *to construct the segment terminating in* r. **[c]** *Join* q *to* r *to create the first rung of Schild's Ladder* $\asymp \mathbf{w}_{||}(p \rightsquigarrow q)$. **[d]** *Repeat the construction to add more rungs to Schild's Ladder. Finally, let* $\epsilon \to 0$, *thereby parallel transporting* $\mathbf{w}_{||}$ *along* G.

we parallel-transport it back from q *to* p *as* $\mathbf{w}_{||}(q \rightsquigarrow p)$. We then find the change, $[\mathbf{w}_{||}(q \rightsquigarrow p) - \mathbf{w}(p)]$, and finally we divide by ϵ to find the *rate* of change.

Now that we know how to do parallel transport [three ways!] in an n-manifold, we can define the intrinsic (aka "covariant") derivative in an n-manifold *exactly* as just described. Only the notation changes: henceforth, *we adopt the standard choice of a* ***bold*** *nabla symbol—*$\boldsymbol{\nabla}$*—to represent the* ***intrinsic derivative***,

$$\boxed{\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w} \asymp \frac{\mathbf{w}_{||}(q \rightsquigarrow p) - \mathbf{w}(p)}{\epsilon}.} \tag{29.2}$$

For our upcoming purposes, it will prove much more useful to think of the intrinsic derivative not in terms of a *rate* of change, but rather more directly in terms of the actual change itself:

$$\mathbf{w}_{||}(q \rightsquigarrow p) - \mathbf{w}(p) \asymp \epsilon \nabla_{\mathbf{v}} \mathbf{w} = \nabla_{\epsilon \mathbf{v}} \mathbf{w}.$$

Let $\delta_{pq}\mathbf{w}$ denote the small intrinsic change in $\mathbf{w}$ along the vector $\boldsymbol{\epsilon} \equiv \epsilon \mathbf{v}$ from p to q, then

$$\boxed{\delta_{pq}\mathbf{w} = \text{intrinsic change in } \mathbf{w} \text{ from tail to tip of } \boldsymbol{\epsilon} \asymp \nabla_{\boldsymbol{\epsilon}} \mathbf{w}.} \tag{29.3}$$

It follows directly from this definition, that

$$\boxed{\nabla_{\mathbf{v}} \mathbf{w}_{||} = 0 \quad \Longleftrightarrow \quad \mathbf{w}_{||} \text{ is parallel-transported along } \mathbf{v}.}$$

If $\mathbf{v}$ is the velocity of a geodesic, then it is parallel-transported along itself, so the ***Geodesic Equation*** (23.5) now takes the form,

$$\boxed{\nabla_{\mathbf{v}} \mathbf{v} = \mathbf{0}.} \tag{29.4}$$

We should note that if we allow the particle to speed up or slow down as it travels along the geodesic, then we obtain a more general form of the geodesic equation, namely, $\nabla_{\mathbf{v}} \mathbf{v} \propto \mathbf{v}$, which says that the *direction* of $\mathbf{v}$ is intrinsically constant, but its magnitude is permitted to change.

29.5 The Riemann Curvature Tensor

29.5.1 Parallel Transport Around a Small "Parallelogram"

Just as we did in the case of 2-manifolds (surfaces) we can now study curvature in an n-manifold by parallel-transporting a unit vector $\mathbf{w}_{||}$ around a small loop L. More specifically, we shall *attempt* to construct L as a *parallelogram* whose edges are built out of two (unit) vector fields $\mathbf{u}$ and $\mathbf{v}$.

Therefore, as illustrated in [29.6], starting at o, we lay out short vectors in these two directions, $\mathbf{u}(o)\,\delta u$ (connecting o to a) and $\mathbf{v}(o)\,\delta v$ (connecting o to p). In order to attempt to create a parallelogram, we now lay out $\mathbf{u}(a)\,\delta u$ (connecting a to b) and $\mathbf{v}(p)\,\delta v$ (connecting p to q). But the problem is that, in general, $q \neq b$: *the "parallelogram" fails to close!*

In the next subsection we shall find the formula for the extremely small "gap-closing" vector $\mathbf{c}$ connecting b to q. For now, though, simply suppose that we know how to close this gap, and are therefore able to parallel-transport $\mathbf{w}$ around the *closed* loop, $L = oabqpo$.

Within a 2-surface, the initial vector $\mathbf{w}_o$ that is parallel-transported can only lie in the same plane as $\mathbf{u}$ and $\mathbf{v}$. The *fundamentally new feature* in a 3-manifold (or more generally an n-manifold) is that $\mathbf{w}$ *can now stick out of the plane of the small loop around which it is parallel-transported.*

Having set the scene, let us parallel-transport $\mathbf{w}_o$ around L as $\mathbf{w}_{||}$, returning it to o as $\mathbf{w}_{||}(o)$. We can then define the ***vector holonomy***[5] to be the net change in $\mathbf{w}_{||}$ induced by the curvature:

$$\boxed{\delta \mathbf{w}_{||} \equiv \mathbf{w}_{||}(\text{upon returning to } o) - \mathbf{w}(\text{as it departs from } o)}$$

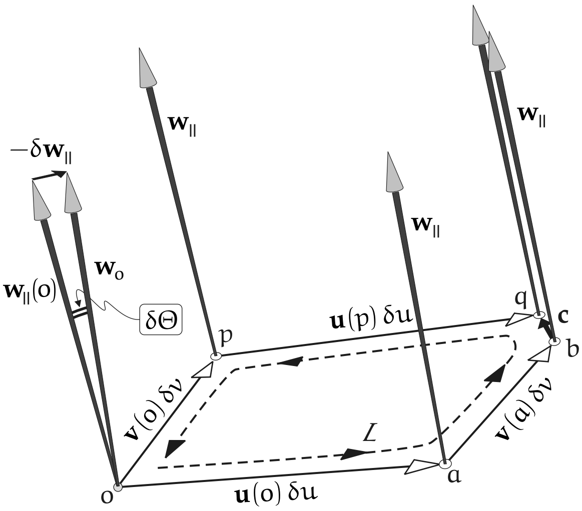
Within a 2-surface, we introduced the holonomy operator $\mathcal{R}(L)$, which (when applied to $\mathbf{w}$) gave the net rotation (holonomy) of $\mathbf{w}$ after it was parallel-transported around L. Now, in a 3-manifold, or indeed an n-manifold, there are infinitely many different planes within which L may lie. Furthermore, in a 2-surface the entire tangent plane rotated rigidly as the vectors within

[5] We have invented this term because we are not aware of a standard name for this concept in the literature.

it were parallel-transported, so we did not need to pay attention to *which* vector $\mathbf{w}$ was being transported: they all rotated the same amount, $\mathcal{R}(L)$. But in a 3-manifold, $\mathbf{w}$ can stick out of the plane of L, and in an n-manifold, there can be many independent ways it can point. Crucially, *the vector holonomy now* does *depend on which vector is parallel-transported around the loop.*

For both these reasons, we must refine and generalize our previous notation, and introduce the ***Riemann curvature operator*** $\boldsymbol{\mathcal{R}}$ associated with the edges of the parallelogram L, which then acts on the vector that is parallel-transported, yielding the vector holonomy:

$$\boxed{-\delta\mathbf{w}_{||} \equiv \boldsymbol{\mathcal{R}}(\mathbf{u}\,\delta u, \mathbf{v}\,\delta v)\,\mathbf{w}.}$$



[29.6] Geometric Meaning of Riemann's Curvature. *Starting at* o*, we create a small parallelogram out of two vector fields* $\mathbf{u}$ *and* $\mathbf{v}$*, then parallel-transport an initial vector* $\mathbf{w}_o$ *around the loop* L *as* $\mathbf{w}_{||}$*. When it returns to* o *as* $\mathbf{w}_{||}(o)$*, it has changed by the illustrated* **vector holonomy** $\delta\mathbf{w}_{||} = -\boldsymbol{\mathcal{R}}(\mathbf{u}\,\delta u, \mathbf{v}\,\delta v)\,\mathbf{w}$*, where* $\boldsymbol{\mathcal{R}}$ *is the* **Riemann curvature operator**.

As we know, the curvature $\mathcal{K}$ of a 2-surface is completely captured by the holonomy per unit area. In an n-manifold we can likewise look at the angle $\delta\Theta$ between the initial vector $\mathbf{w}_o$ and the parallel-transported vector $\mathbf{w}_{||}(o)$ when it returns to o, as illustrated in [29.6]. Since $\mathbf{w}$ is a unit vector, the distance $|\delta\mathbf{w}|$ that its tip rotates is ultimately equal to $\delta\Theta$, the angle of rotation.

For simplicity's sake, suppose that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, so that our parallelogram is a rectangle of area $\delta\mathcal{A} = \delta u\,\delta v$. We can then generalize $\mathcal{K}$ and extract a scalar curvature $\mathcal{K}(\mathbf{u}, \mathbf{v}; \mathbf{w})$, once again defined as the rotation per unit area:

$$\mathcal{K}(\mathbf{u}, \mathbf{v}; \mathbf{w}) \asymp \frac{\delta\Theta}{\delta\mathcal{A}} \asymp \frac{|\delta\mathbf{w}_{||}|}{\delta\mathcal{A}} \asymp |\boldsymbol{\mathcal{R}}(\mathbf{u}, \mathbf{v})\,\mathbf{w}|\,.$$

However, it is clear that this is no longer a satisfactory measure of the curvature, for we have completely lost the crucial information about the *direction* of the vector holonomy $\delta\mathbf{w}_{||}$. As we shall see, the Riemann tensor is the geometric object that encodes the *full* curvature information: both the angle $\delta\Theta$ by which $\mathbf{w}_{||}$ tips away from $\mathbf{w}_o$, *and* the direction in which it does so.

29.5.2 *Closing the "Parallelogram" with the Vector Commutator*

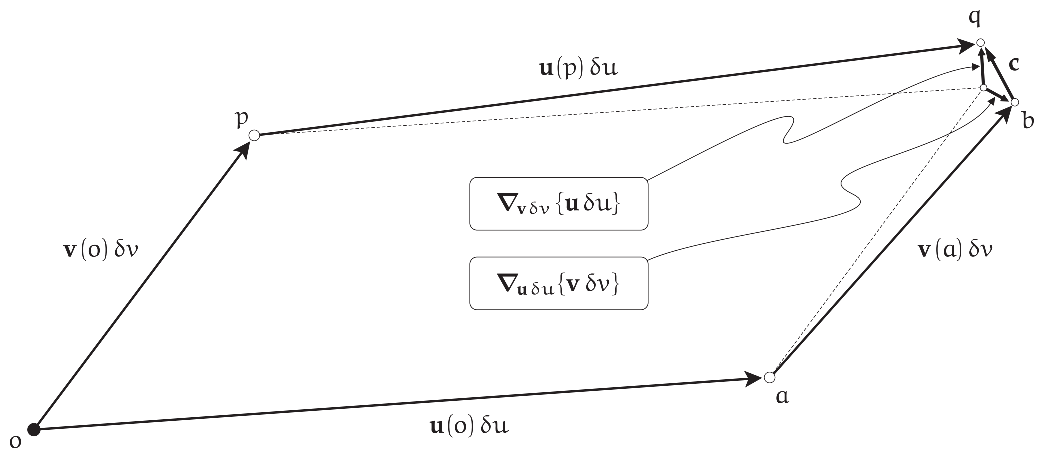
The vector $\mathbf{c}$ that closes up our near-miss parallelogram is extremely small, being (at worst[6]) of order $\delta u\,\delta v$, so ignoring it would still yield an excellent approximation to the curvature. But to obtain a *mathematically perfect* description of the Riemann tensor, we must parallel transport our vector around a *closed* loop, and therefore we *must* close the "parallelogram" with $\mathbf{c}$.

As we see geometrically, directly from [29.7], this very short vector that closes up the gap in the parallelogram can be expressed as the ***commutator***[7] of its edges:

---

[6]For vector fields tied to a coordinate grid, the gap disappears completely.

[7]Also called the ***Lie Bracket***, after the great Norwegian mathematician, Sophus Lie (pronounced "lee") (1842–1899).



[29.7] *Geometric proof that the vector* **c** *that closes up the gap in the "parallelogram" is the* **commutator** *of its edges:* $\mathbf{c} \asymp \nabla_{\mathbf{v}\,\delta v}\{\mathbf{u}\,\delta u\} - \nabla_{\mathbf{u}\,\delta u}\{\mathbf{v}\,\delta v\} = [\mathbf{v},\mathbf{u}]\,\delta u\,\delta v$.

$$\boxed{\mathbf{c} \asymp [\mathbf{v}\,\delta v, \mathbf{u}\delta u] = [\mathbf{v},\mathbf{u}]\,\delta u\,\delta v,}$$

where

$$\boxed{[\mathbf{v},\mathbf{u}] \equiv \nabla_{\mathbf{v}}\mathbf{u} - \nabla_{\mathbf{u}}\mathbf{v}.}$$

We note a simple fact which we will need very shortly, namely, that the commutator is *antisymmetric*:

$$\boxed{[\mathbf{v},\mathbf{u}] = -[\mathbf{u},\mathbf{v}].}$$

29.5.3 The General Riemann Curvature Formula

We can now return to the problem of finding the change $\delta\mathbf{w}_{||}$ in $\mathbf{w}_{||}$ when it is parallel-transported around the closed loop, $L = oabqpo$.

To motivate the following argument, let us briefly review the first crucial step in Hopf's intrinsic proof of GGB—the introduction of the "holonomy" of an *open* curve. As illustrated in [26.1], page 258, we introduced a fiducial vector field **U**, enabling us to define $\mathcal{R}_U(K)$ 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:

$$\mathcal{R}_U(K) \equiv \delta_K\,(\angle U w_{||}).$$

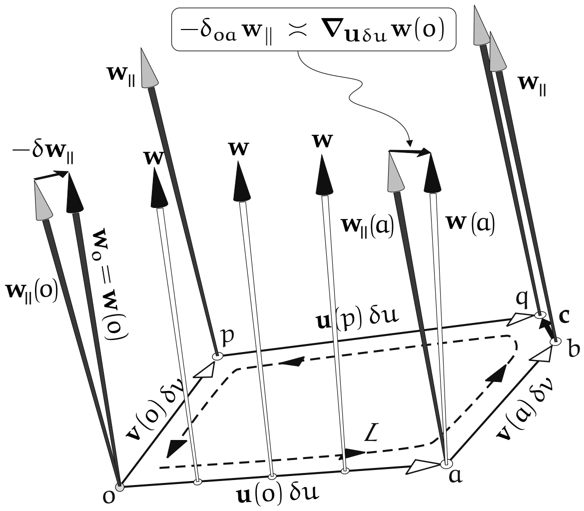
As we noted at the time, the subscript U is essential, for $\mathcal{R}_U(K)$ *does indeed depend on our arbitrary choice of* **U**. Thus $\mathcal{R}_U(K)$ has no true mathematical meaning; it was merely a stepping stone in finding the geometrically meaningful holonomy of a *closed* loop.

Indeed, the holonomy $\mathcal{R}(L)$ of a closed polygonal loop L could then be expressed as the sum of the "holonomies" of its edges, and while each of the individual holonomies *did* depend on the arbitrary choice of **U**, their sum $\mathcal{R}(L) = \mathcal{K}(L)$ measured the curvature within, and was *independent of* **U**.

In a 2-surface we did *not* need to specify the vector $\mathbf{w}$ that was parallel-transported, because the entire tangent plane rotated rigidly as we moved across the surface. (This is *no longer true* in a 3-manifold: the initial choice of the direction of $\mathbf{w}$ *does* affect its vector holonomy around a closed loop.)

Still in a 2-surface, for the moment, we may choose $\mathbf{w}$ freely, so let us choose $\mathbf{w} = \mathbf{U}$. Now consider the *negative* of this holonomy—the change in $\mathbf{U}$ relative to parallel-transported vector $\mathbf{U}_{||}$—along a very a short vector $\boldsymbol{\epsilon}$. Then the holonomy becomes none other than the *intrinsic derivative* of $\mathbf{U}$:

$$-\mathcal{R}_{\mathbf{U}}(\boldsymbol{\epsilon}) = \delta_{\boldsymbol{\epsilon}}\,(\angle \mathbf{U}_{||}\mathbf{U}) \asymp \boldsymbol{\nabla}_{\boldsymbol{\epsilon}}\,(\angle \mathbf{U}_{||}\mathbf{U}) \asymp |\boldsymbol{\nabla}_{\boldsymbol{\epsilon}}\,\mathbf{U}|.$$



[29.8] The Riemann Curvature Tensor. *To calculate the vector holonomy, we introduce a fiducial vector field* $\mathbf{w}$ *that is completely arbitrary,* except *that* $\mathbf{w}(o) = \mathbf{w}_o$*; these are the black and white vectors (only those along* oa *are shown). When summed over all five edges, the changes in* $\mathbf{w}$ *relative to* $\mathbf{w}_{||}$ *yield the* negative *of the vector holonomy of the loop.*

We can now return to the problem at hand and apply this insight to the evaluation of $-\delta\mathbf{w}_{||}$ within the 3-manifold [29.6]. As illustrated in [29.8], we now introduce a fiducial vector field $\mathbf{w}$ throughout the region containing the parallelogram—these are the black-and-white vectors, which we have only drawn along oa, to avoid clutter. This vector field can be completely arbitrary, *except* that now it *does* matter which vector $\mathbf{w}_o$ is parallel-transported, so we insist that $\mathbf{w}(o) = \mathbf{w}_o$. Then, using (29.3), the negative of the vector holonomy along the first edge of the parallelogram is[8]

$$-\delta_{oa}\mathbf{w}_{||} = \mathbf{w}(a) - \mathbf{w}_{||}(a) \asymp \boldsymbol{\nabla}_{\mathbf{u}\delta u}\mathbf{w}(o) = \delta u\,\boldsymbol{\nabla}_{\mathbf{u}}\mathbf{w}(o).$$

Let us now repeatedly apply (29.3) to sum these changes in the arbitrary vector field $\mathbf{w}$, relative the parallel-transported vector $\mathbf{w}_{||}$, along the five sides of the closed-up parallelogram. When we return to o, of course $\mathbf{w}(o)$ has the same value as when we left o, so the net change in $\mathbf{w}$ relative to $\mathbf{w}_{||}$ is *entirely* due to the absolute change in $\mathbf{w}_{||}$, and is *independent* of the arbitrary choice of the vector field $\mathbf{w}$—this net change is the vector holonomy we seek.

We spell out every step of this summation, and encourage you to refer back to [29.8] to make geometrical sense of each line. Having absorbed each separate step of this crucial argument, it is worth your time to step back and grasp it as a single, complete idea:

$$\begin{aligned}
-\delta\mathbf{w}_{||} &= -\Big[\delta_{oa}\,\mathbf{w}_{||} + \delta_{ab}\,\mathbf{w}_{||} + \delta_{bq}\,\mathbf{w}_{||} + \delta_{qp}\,\mathbf{w}_{||} + \delta_{po}\,\mathbf{w}_{||}\Big] \\
&\asymp \delta u\,\boldsymbol{\nabla}_{\mathbf{u}}\mathbf{w}(o) + \delta v\,\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}(a) + \boldsymbol{\nabla}_{\mathbf{c}}\mathbf{w}(b) - \delta u\,\boldsymbol{\nabla}_{\mathbf{u}}\mathbf{w}(p) - \delta v\,\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}(o) \\
&= \delta v\,\{\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}(a) - \boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}(o)\} - \delta u\,\{\boldsymbol{\nabla}_{\mathbf{u}}\mathbf{w}(p) - \boldsymbol{\nabla}_{\mathbf{u}}\mathbf{w}(o)\} + \delta u\,\delta v\,\boldsymbol{\nabla}_{[\mathbf{v},\mathbf{u}]}\mathbf{w}(o) \\
&\asymp \delta v\,\{\delta u\,\boldsymbol{\nabla}_{\mathbf{u}}\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}\} - \delta u\,\{\delta v\,\boldsymbol{\nabla}_{\mathbf{v}}\boldsymbol{\nabla}_{\mathbf{u}}\mathbf{w}\} - \delta u\,\delta v\,\boldsymbol{\nabla}_{[\mathbf{u},\mathbf{v}]}\mathbf{w}
\end{aligned}$$

---

[8] Here, and in the following, a more accurate approximation would be obtained by evaluating the derivative at the *midpoint* of the segment. However, since we shall shrink the entire figure down to the point o, and are dealing with ultimate equality in this limit, we think the argument may be easier to follow without taking this extra step.

$$
\begin{aligned}
&= \delta u\,\delta v \left\{ \nabla_{\mathbf{u}}\nabla_{\mathbf{v}} - \nabla_{\mathbf{v}}\nabla_{\mathbf{u}} - \nabla_{[\mathbf{u},\mathbf{v}]} \right\} \mathbf{w}(o) \\
&= \delta u\,\delta v\, \mathcal{R}(\mathbf{u},\mathbf{v})\, \mathbf{w}(o).
\end{aligned}
$$

We have arrived at a major event in our drama—**vector holonomy in terms of Riemann's curvature**:

$$
\boxed{-\frac{\delta \mathbf{w}_{||}}{\delta u\,\delta v} \asymp \mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w} = \left\{ [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]} \right\} \mathbf{w}.} \tag{29.5}
$$

We now see that the ***Riemann curvature operator*** introduced in [29.6] is, in fact,

$$
\boxed{\mathcal{R}(\mathbf{u},\mathbf{v}) = [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]}.} \tag{29.6}
$$

The minus sign on the left-hand side of (29.5) is important, so let us make a point of reminding the reader of its origin and geometrical meaning. The operator on the right-hand side tells us the net change in the (arbitrary) fiducial vector field $\mathbf{w}$ (*relative to* $\mathbf{w}_{||}$) as we traverse the loop. But the geometrically meaningful quantity we seek is the vector holonomy $\delta\mathbf{w}_{||}$, which is the *opposite*: the change in $\mathbf{w}_{||}$ *relative to* $\mathbf{w}$.

Note that $\mathcal{R}$ is *antisymmetric*:

$$
\boxed{\mathcal{R}(\mathbf{u},\mathbf{v}) = -\mathcal{R}(\mathbf{v},\mathbf{u}) \quad \Longrightarrow \quad \mathcal{R}(\mathbf{u},\mathbf{u}) = \mathbf{0}.} \tag{29.7}
$$

The truth of the first equation is geometrically clear, for it simply asserts that if we reverse the direction in which we traverse the parallelogram, the vector holonomy is reversed. Likewise, the second equation may be viewed as saying that if *any* vector is parallel-transported back and forth a short distance along an arbitrary vector $\mathbf{u}$, then it returns unchanged.

Only a minor change of notation is now needed to finally arrive at the standard definition of the famous ***Riemann curvature tensor*** **R**—it is a vector-valued function $\mathbf{R}(\mathbf{u},\mathbf{v};\,\mathbf{w})$ of three input vectors:[9]

$$
\boxed{\mathbf{R}(\mathbf{u},\mathbf{v};\,\mathbf{w}) \equiv \mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w} = \left\{ [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]} \right\} \mathbf{w}.} \tag{29.8}
$$

CONVENTION WARNING: The arrangement of vectors in the slots of **R** differs amongst authors. For example, Misner, Thorne, and Wheeler (1973) *agree* with our definition of $\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w}$ above, but, on the left-hand side, their $\mathbf{R}(\mathbf{u},\mathbf{v};\,\mathbf{w})$ rearranges the slots in such a way that their **R** turns out to be the *negative* of ours.

[9] It is *not* entirely standard to use a semicolon to separate the parallelogram vectors from the vector that is parallel-transported around it, but we think it helpful (as does Penrose!).

29.5.4 Riemann's Curvature Is a Tensor

As we shall discuss in greater detail in Act V, a ***tensor*** is (by definition) a *linear* function of its multiple vector[10] inputs. As we now explain, $\mathbf{R}$ is indeed ***multi-linear***:

> $\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w})$ *is a linear function of each of its three inputs, when the other two are held fixed.*

Furthermore, *despite* the fact that formula (29.8) is built out of derivatives, the Riemann curvature *only* depends on the values of the three vectors *at* the point at which $\mathbf{R}$ is evaluated—it is *independent* of how the vectors vary in the vicinity of that point. At the computational level, this appears downright *paradoxical*, and yet it does make geometrical sense, for $\mathbf{R}$ is telling us about the curvature of the *space itself*, and the vectors merely serve to single out particular parts (or ***components***) of this curvature.

To summarize, here are

> **The two defining properties of a *tensor*:**
>
> - *The output depends* linearly *on each input vector.*
> - *The output* only *depends on the input vectors* at *the point of evaluation.*

First, let us establish linearity in the $\mathbf{w}$-slot. Suppose that $\mathbf{w} = k_1\mathbf{w}_1 + k_2\mathbf{w}_2$, where k_1 and k_2 are constants. If $\mathbf{w}$, $\mathbf{w}_1$, and $\mathbf{w}_2$ are each parallel-transported (starting at o) around a small (ultimately vanishing) parallelogram L with edges along $\mathbf{u}$ and $\mathbf{v}$, then

$$\delta\mathbf{w}_{||} = k_1\delta[\mathbf{w}_1]_{||} + k_2\delta[\mathbf{w}_2]_{||},$$

so the geometrical interpretation (29.5) implies that

$$\mathbf{R}(\mathbf{u}, \mathbf{v};\, k_1\mathbf{w}_1 + k_2\mathbf{w}_2) = k_1\,\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w}_1) + k_2\,\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w}_2).$$

Of course this can also be instantly confirmed by computation: just use the formula (29.8) and the simple fact that the intrinsic derivatives act linearly, e.g.,

$$\boldsymbol{\nabla}_{\mathbf{v}}[k_1\mathbf{w}_1 + k_2\mathbf{w}_2] = k_1\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}_1 + k_2\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{w}_2.$$

To start to test the *second* requirement of tensorhood, suppose that we change $\mathbf{w}^{\mathrm{OLD}}$ in the vicinity of o into $\mathbf{w}^{\mathrm{NEW}} = f\,\mathbf{w}^{\mathrm{OLD}}$, where f is an arbitrary (differentiable) function of position, *except* that $f(o) = 1$, so that $\mathbf{w}^{\mathrm{NEW}}(o) = \mathbf{w}^{\mathrm{OLD}}(o)$.

At this point, we strongly encourage you to pause in order to actually *carry out* the somewhat lengthy (but straightforward) *computation* that proves that, when evaluated at o,

$$\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w}^{\mathrm{NEW}}) = \mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w}^{\mathrm{OLD}}).$$

If you just did this calculation, you saw that four terms arose that involved derivatives of f, but then, "miraculously," they all cancelled each other out! (NOTE: This is *not* cause for celebration—see the definition of a "false miracle" in the Prologue!)

[10]In Act V we will generalize this to the *standard* definition, whereby we allow the input of both vectors *and* 1-forms (introduced at the start of Act V).

Fortunately, this can instead be understood *directly* from the geometrical perspective, because varying the fiducial vector field $\mathbf{w}$ along L in *any* manner (not merely scaling it by f) has absolutely no effect on the geometrically meaningful parallel transport of the initial vector around L, and hence absolutely no effect on the vector holonomy! Thus $\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w})$ does indeed *only* depend on the value of $\mathbf{w}$ *at* o.

Now let us turn our attention to the first two slots, and let us start with the second defining property of a tensor. In the limit that loop shrinks to o, only the values $\mathbf{u}(o)$ and $\mathbf{v}(o)$ are needed to construct the loop around which $\mathbf{w}$ is parallel-transported, so the vector holonomy, and therefore $\mathbf{R}$, *only* depends on the values of these input vectors *at* o.

To witness the decisive advantage of the geometrical perspective, but also as a character-building exercise in the computational use of formula (29.8), let us play the same game with $\mathbf{u}$ that we previously did with $\mathbf{w}$. That is, let us change $\mathbf{u}^{\mathrm{OLD}}$ in the vicinity of o into $\mathbf{u}^{\mathrm{NEW}} = f\,\mathbf{u}^{\mathrm{OLD}}$, where f is an arbitrary function. Now try your hand at proving (directly from the formula) that

$$\mathbf{R}(\mathbf{u}^{\mathrm{NEW}}, \mathbf{v};\, \mathbf{w}) = \mathbf{R}(f\,\mathbf{u}^{\mathrm{OLD}}, \mathbf{v};\, \mathbf{w}) = f\,\mathbf{R}(\mathbf{u}^{\mathrm{OLD}}, \mathbf{v};\, \mathbf{w}).$$

Thus, if we again insist that $f(o) = 1$, to ensure that $\mathbf{u}^{\mathrm{NEW}}(o) = \mathbf{u}^{\mathrm{OLD}}(o)$, then

$$\mathbf{R}(\mathbf{u}^{\mathrm{NEW}}, \mathbf{v};\, \mathbf{w})(o) = \mathbf{R}(\mathbf{u}^{\mathrm{OLD}}, \mathbf{v};\, \mathbf{w})(o),$$

as required of a tensor.

Finally, let us turn to the linearity in the first two slots. Because of the antisymmetry (29.7), it suffices to prove linearity in either slot. Here there is essentially no difference between the geometrical and computational perspectives, and we leave the very short calculation to you, which establishes that,

$$\mathbf{R}(k_1\mathbf{u}_1 + k_2\mathbf{u}_2, \mathbf{v};\, \mathbf{w}) = k_1\,\mathbf{R}(\mathbf{u}_1, \mathbf{v};\, \mathbf{w}) + k_2\,\mathbf{R}(\mathbf{u}_2, \mathbf{v};\, \mathbf{w}).$$

GRAND CONCLUSION: Riemann's curvature is indeed a *tensor*!

29.5.5 Components of the Riemann Tensor

In order to give a numerical description of geometrical objects, we introduce a set of *orthonormal basis vectors*, $\{\mathbf{e}_i\}$, for the $\mathbb{R}^n$ tangent space at each point of the n-manifold. Then, for example, a geometrical vector $\mathbf{u}$ is represented by its numerical *components*, $\{u^i\}$, where $\mathbf{u} = \sum_i u^i \mathbf{e}_i$. WARNING: When dealing with components, we must remember that the superscripts are labels, *not* powers!

More complex geometrical objects, such as the Riemann tensor, require *multiple* distinct indices to describe their components. This leads to a profusion of summations. But it turns out that we can always arrange for the index that is being summed over to appear once as a superscript and once as a subscript.

Einstein therefore introduced a simple, clutter-clearing convention, called the ***Einstein summation convention***, whereby the sum over such a pair of indices is simply *understood*, and the summation sign is omitted. For example,

$$\boxed{\mathbf{u} = \sum_{i=1}^{n} u^i \mathbf{e}_i \quad \Longleftrightarrow \quad \textbf{\textit{Einstein summation convention}}\text{:} \quad \mathbf{u} = u^i \mathbf{e}_i.}$$

To find the components of the Riemann tensor, we decompose each of its three vector inputs into components:

$$\mathbf{u} = u^i \mathbf{e}_i, \qquad \mathbf{v} = v^j \mathbf{e}_j, \qquad \text{and} \qquad \mathbf{w} = w^k \mathbf{e}_k.$$

Thus,

$$\begin{aligned}\mathbf{R}(\mathbf{u},\mathbf{v};\,\mathbf{w}) &= \mathbf{R}(u^i\mathbf{e}_i,\, v^j\mathbf{e}_j;\, w^k\mathbf{e}_k)\\ &= \mathbf{R}(\mathbf{e}_i,\,\mathbf{e}_j;\,\mathbf{e}_k)\,u^i\,v^j\,w^k\end{aligned}$$

We can now define $R_{ijk}{}^l$, the ***components of the Riemann tensor***, as the components of the vector that results when the Riemann tensor acts on three basis vectors:

$$\boxed{\mathbf{R}(\mathbf{e}_i,\,\mathbf{e}_j;\,\mathbf{e}_k) \equiv R_{ijk}{}^l\,\mathbf{e}_l.}$$

Thus the effect of **R** on three general vectors can easily be expressed in terms of these components:

$$\mathbf{R}(\mathbf{u},\mathbf{v};\,\mathbf{w}) = \left[R_{ijk}{}^l\,u^i\,v^j\,w^k\right]\,\mathbf{e}_l.$$

For later use, we also define,

$$\boxed{R_{ijkm} \equiv \mathbf{R}(\mathbf{e}_i,\,\mathbf{e}_j;\,\mathbf{e}_k)\bullet\mathbf{e}_m.} \tag{29.9}$$

In our chosen *orthonormal basis*, $\mathbf{e}_l \bullet \mathbf{e}_m = 0$ if $l \neq m$, and it equals 1 when $l = m$, so

$$\boxed{R_{ijkm} = R_{ijk}{}^m.} \tag{29.10}$$

29.5.6 *For a Given* $\mathbf{w}_o$*, the Vector Holonomy* Only *Depends on the* Plane *of the Loop and Its* Area

Within a 2-surface, the holonomy $\mathcal{R}(L) \asymp \mathcal{K}\,\delta\mathcal{A}$ of a small loop L *only* depended on its area, $\delta\mathcal{A}$, and was independent of it shape. In an n-manifold, there are now many, independent choices for the plane of L. Let us therefore define,

$$\boxed{\Pi(\mathbf{u},\mathbf{v}) \equiv \text{The plane spanned by } \mathbf{u} \text{ and } \mathbf{v}.}$$

This choice of $\Pi(\mathbf{u},\mathbf{v})$ certainly dictates the vector holonomy. However, for a given $\mathbf{w}_o$, the vector holonomy *only* depends on Π itself, and on the area of the loop:

> *If we parallel transport* $\mathbf{w}_o$ *around a small (ultimately vanishing) parallelogram in a plane* Π*, then the vector holonomy is proportional to the* area $\delta\mathcal{A}$ *of the parallelogram, and is* independent *of its shape.* (29.11)

We will prove this by calculation, in order to illustrate the power of our brand-new tool of tensor components. Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be chosen within the plane of L. Then, using the Einstein summation convention,

$$\mathbf{u}\,\delta u = \delta u^1\,\mathbf{e}_1 + \delta u^2\,\mathbf{e}_2 = \delta u^i\,\mathbf{e}_i \qquad \text{and} \qquad \mathbf{v}\,\delta v = \delta v^1\,\mathbf{e}_1 + \delta v^2\,\mathbf{e}_2 = \delta v^j\,\mathbf{e}_j.$$

Thus the area of the loop is

$$\delta\mathcal{A} = \begin{vmatrix} \delta u^1 & \delta v^1 \\ \delta u^2 & \delta v^2 \end{vmatrix} = \delta u^1\,\delta v^2 - \delta u^2\,\delta v^1.$$

Let $\mathbf{w}_o = w_o^k\,\mathbf{e}_k$, and $\delta\mathbf{w}_{||} = \delta w_{||}^l\,\mathbf{e}_l$. Then

$$\begin{aligned} -\delta\mathbf{w}_{||} &= -\delta w_{||}^l\,\mathbf{e}_l \\ &= \mathbf{R}(\mathbf{u}\,\delta u, \mathbf{v}\,\delta v;\, \mathbf{w}_o) \\ &= \mathbf{R}(\delta u^i\,\mathbf{e}_i, \delta v^j\,\mathbf{e}_j;\, w_o^k\,\mathbf{e}_k) \\ &= \delta u^i\,\delta v^j\,w_o^k\,R_{ijk}{}^l\,\mathbf{e}_l. \end{aligned}$$

But, by virtue of (29.7),

$$R_{11k}{}^l = 0 = R_{22k}{}^l \qquad \text{and} \qquad R_{21k}{}^l = -R_{12k}{}^l.$$

Therefore, the components of the vector holonomy are given by

$$\boxed{-\delta\mathbf{w}_{||} = -\delta w_{||}^l\,\mathbf{e}_l \asymp \delta\mathcal{A}\left[R_{12k}{}^l\,w_o^k\right]\,\mathbf{e}_l,} \tag{29.12}$$

completing the proof of (29.11).

29.5.7 Symmetries of the Riemann Tensor

Naively, one would expect the number of independent components of the Riemann tensor R_{ijkm} in a 3-manifold, defined in (29.9), to be $3^4 = 81$. And yet we have already stated that the true number is only *six!* This dramatic reduction results from the fact that the Riemann tensor possesses *four* remarkable algebraic symmetries.

The first one we have already met in (29.7): $\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w})$ is antisymmetric in its first two slots, and this implies that

$$\boxed{R_{jikm} = -R_{ijkm}.}$$

This immediately reduces the number of independent components [exercise] from 81 to 54.

We now prove a second, far less obvious symmetry, namely, that R_{ijkm} is also antisymmetric in its *last* two slots:

$$\boxed{R_{ijmk} = -R_{ijkm}} \tag{29.13}$$

This further reduces the number of independent components [exercise] from 54 to 36.

Since parallel transport preserves length, the vector $\mathbf{w}_o$ returns to o as the *slightly rotated vector* $\mathbf{w}_{||}(o)$, and therefore the connecting vector $\delta\mathbf{w}_{||}$ between the tips of these vectors is ultimately *orthogonal* to both of them:

$$\delta\mathbf{w}_{||} \bullet \mathbf{w}_o = 0 \qquad \Longrightarrow \qquad [\mathcal{R}(\mathbf{u}, \mathbf{v})\,\mathbf{w}_o] \bullet \mathbf{w}_o = 0.$$

Now let $\mathbf{x}$ and $\mathbf{y}$ be arbitrary vectors, and set $\mathbf{w}_o = \mathbf{x}+\mathbf{y}$. Then,

$$\begin{aligned} 0 &= [\mathcal{R}(\mathbf{u},\mathbf{v})\,(\mathbf{x}+\mathbf{y})] \bullet (\mathbf{x}+\mathbf{y}) \\ &= [\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}] \bullet \mathbf{x} + [\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}] \bullet \mathbf{y} + [\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{y}] \bullet \mathbf{x} + [\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{y}] \bullet \mathbf{y} \\ &= 0 + [\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}] \bullet \mathbf{y} + [\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{y}] \bullet \mathbf{x} + 0. \end{aligned}$$

Therefore,

$$\boxed{[\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}] \bullet \mathbf{y} = -\,[\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{y}] \bullet \mathbf{x},} \tag{29.14}$$

and so (29.13) follows immediately from the definition (29.9).

For the sake of completeness, we now state the remaining symmetries of the Riemann tensor.

First, the ***Algebraic Bianchi Identity***[11] states that if the first three vectors are permuted cyclically, their sum vanishes:

$$\boxed{\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w} + \mathcal{R}(\mathbf{v},\mathbf{w})\,\mathbf{u} + \mathcal{R}(\mathbf{w},\mathbf{u})\,\mathbf{v} = \mathbf{0} \quad\Longleftrightarrow\quad R_{ijkm} + R_{jkim} + R_{kijm} = 0.} \tag{29.15}$$

Exercise 10 provides both a computational and a geometric proof of this result.

Next, the Riemann tensor is *symmetric under interchange of the first and second* pairs *of vectors*:

$$\boxed{[\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}] \bullet \mathbf{y} = [\mathcal{R}(\mathbf{x},\mathbf{y})\,\mathbf{u}] \bullet \mathbf{v} \qquad\Longleftrightarrow\qquad R_{ijkm} = R_{kmij}.} \tag{29.16}$$

Although this is very useful, it is not a truly new symmetry, but rather a consequence of the previous symmetries. For a computational proof, see Exercise 10.

Using these symmetries, you may confirm [exercise][12] that in a 3-manifold the Riemann tensor has only six independent components. More generally, Exercise 11 shows that in an n-manifold the number of components is given by $\frac{1}{12}n^2(n^2-1)$, as previously claimed in (29.1).

Finally, in addition to the four algebraic symmetries above, there exists a fifth, *differential* symmetry, of fundamental importance to Einstein's theory of gravity—it is called the ***Differential Bianchi Identity:***[13]

$$\boxed{\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w} + \nabla_{\mathbf{u}}\mathcal{R}(\mathbf{v},\mathbf{x})\,\mathbf{w} + \nabla_{\mathbf{v}}\mathcal{R}(\mathbf{x},\mathbf{u})\,\mathbf{w} = \mathbf{0},} \tag{29.17}$$

in which the first three vectors are permuted cyclically, as they were in the Algebraic Bianchi Identity. The proof of this result will be deferred until we can derive it elegantly, using ***curvature 2-forms***, in Section 38.12.4.

---

[11] Also called the ***First Bianchi Identity*** or the ***Bianchi Symmetry***, but actually discovered by Ricci!

[12] Try simply listing six independent components, and then convince yourself that all other components can be found using the symmetries.

[13] Also called the ***Second Bianchi Identity***. According to Pais (1982, pp. 275–276), this was first discovered by Aurel Voss (1880), then Ricci (1889), and then Bianchi (1902). However, even Voss was not the first—see [29.10]!

29.5.8 Sectional Curvatures

In a 2-surface, the Gaussian curvature $\mathcal{K}$ emerged as the holonomy per unit area of a vector that *necessarily resided within the plane of the loop* around which it was parallel-transported. But in our 3-manifold (or n-manifold), $\mathbf{w}_o$ will typically stick *out* of this plane. Nevertheless, we are certainly free to *choose* it to lie within $\Pi(\mathbf{u},\mathbf{v})$, in the hope of recovering a concept of curvature akin to the 2-dimensional case.

However, at first we seem to hit a road block:

> *Even if we insist that the initial vector $\mathbf{w}_o$ lie within $\Pi(\mathbf{u},\mathbf{v})$, then parallel-transporting it around a loop within Π will* not *return it to o merely rotated within Π—the returning vector $\mathbf{w}_{||}(o)$ will typically stick out of Π.*

To overcome this, let us focus our attention on the orthogonal *projection* $\mathcal{P}[\mathbf{w}_{||}]$ of $\mathbf{w}_{||}$ into Π. More formally, if $\Pi=\Pi(\mathbf{e}_1,\mathbf{e}_2)$ then this ***orthogonal projection operator*** $\mathcal{P}$ is given by,

$$\boxed{\mathcal{P}[a^1\mathbf{e}_1 + a^2\mathbf{e}_2 + a^3\mathbf{e}_3] = a^1\mathbf{e}_1 + a^2\mathbf{e}_2.}$$

We will prove the following:

> *Let $\mathbf{w}_o$ be* any *vector in Π, and let $\mathbf{w}_{||}$ be parallel-transported around a loop in Π. Define $\mathcal{K}(\mathbf{u},\mathbf{v})$ to be the rotation per unit area of the* projection *$\mathcal{P}[\mathbf{w}_{||}]$ of $\mathbf{w}_{||}$ into Π. Then $\mathcal{K}(\mathbf{u},\mathbf{v})$ is* independent *of the specific choice of $\mathbf{w}_o$.*

The curvature $\mathcal{K}(\mathbf{u},\mathbf{v})=\mathcal{K}(\Pi)$ therefore depends *only* on the plane Π, and is called the ***sectional curvature*** of $\Pi(\mathbf{u},\mathbf{v})$.

To verify this, we simply take the general vector holonomy formula (29.12) and insist that $\mathbf{w}_o$ lie in $\Pi(\mathbf{u},\mathbf{v})=\Pi(\mathbf{e}_1,\mathbf{e}_2)$:

$$\mathbf{w}_o = w_o^1\,\mathbf{e}_1 + w_o^2\,\mathbf{e}_2,$$

and for simplicity's sake let us again assume that it is unit vector. Writing (29.12) in column-vector form, and using (29.10), we see that $\mathcal{P}[\delta\mathbf{w}_{||}]$ is

$$-\begin{bmatrix}\delta w_{||}^1\\ \delta w_{||}^2\end{bmatrix} = \begin{bmatrix}R_{12k}{}^1\,w_o^k\\ R_{12k}{}^2\,w_o^k\end{bmatrix}\delta\mathcal{A} = \begin{bmatrix}R_{121}{}^1\,w_o^1 + R_{122}{}^1\,w_o^2\\ R_{121}{}^2\,w_o^1 + R_{122}{}^2\,w_o^2\end{bmatrix}\delta\mathcal{A} = \begin{bmatrix}R_{1211}\,w_o^1 + R_{1221}\,w_o^2\\ R_{1212}\,w_o^1 + R_{1222}\,w_o^2\end{bmatrix}\delta\mathcal{A}.$$

But the antisymmetry (29.13) yields

$$R_{1211} = 0 = R_{1222} \qquad \text{and} \qquad R_{1221} = -R_{1212}.$$

Therefore, we now formally define the ***sectional curvature*** as

$$\boxed{\boxed{\mathcal{K}(\Pi) \equiv \mathcal{K}(\mathbf{e}_1,\,\mathbf{e}_2) \equiv [\boldsymbol{\mathcal{R}}(\mathbf{e}_1,\,\mathbf{e}_2)\,\mathbf{e}_2]\bullet\mathbf{e}_1 = R_{1221},}} \tag{29.18}$$

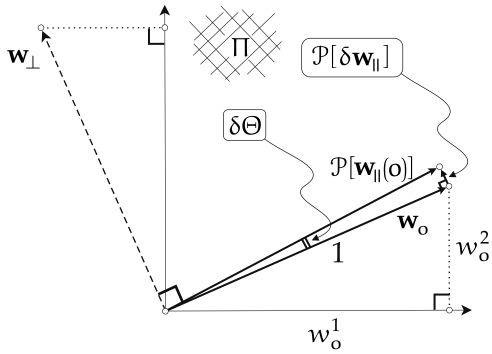
which is to say, the holonomy of $\mathbf{e}_2$, after parallel-transporting it around the $\{\mathbf{e}_1,\mathbf{e}_2\}$-loop, projected onto $\mathbf{e}_1$. Then

$$\boxed{\mathcal{P}[\delta\mathbf{w}_{||}] = \begin{bmatrix} \delta w_{||}^1 \\ \delta w_{||}^2 \end{bmatrix} \asymp \begin{bmatrix} -w_o^2 \\ w_o^1 \end{bmatrix} \mathcal{K}(\Pi)\,\delta\mathcal{A} = \mathbf{w}_\perp\,\mathcal{K}(\Pi)\,\delta\mathcal{A},} \tag{29.19}$$

where $\mathbf{w}_\perp$ is simply $\mathbf{w}_o$ rotated through a right angle within Π, as illustrated in [29.9].

Thus, as the figure explains, the angle $\delta\Theta$ through which the *projection* $\mathcal{P}[\mathbf{w}_{||}]$ has been rotated by parallel-transport around the loop of area $\delta\mathcal{A}$ is $\delta\Theta \asymp \mathcal{K}(\Pi)\,\delta\mathcal{A}$, *independent* of $\mathbf{w}_o$. We have come full circle to our original view of curvature as the holonomy per unit area within a 2-surface:

$$\boxed{\frac{\delta\Theta}{\delta\mathcal{A}} \asymp \mathcal{K}(\Pi) = R_{1221}.} \tag{29.20}$$



[29.9] *If* $\mathbf{w}_o$ *lies* within *the plane* Π *of the loop, then the* projection *into* Π *of the parallel-transported vector returns rotated by* $\delta\Theta \asymp |\delta\mathcal{P}[\mathbf{w}_{||}]| \asymp \mathcal{K}(\Pi)\,\delta\mathcal{A}$, *where* $\mathcal{K}(\Pi)$ *is the* **sectional curvature** *of* Π.

Only now there are *many* such sectional curvatures, depending upon *which* plane Π we parallel transport within.

As we mentioned earlier, knowing the sectional curvatures in just the orthogonal coordinate planes, for example, only gives a very incomplete[14] picture of the Riemann curvature of the space. But what if we are allowed to know $\mathcal{K}(\Pi)$ for *all possible* planes Π?

It turns out that this *is* sufficient to reconstruct the complete Riemann tensor! It is not easy to find this discussed in standard texts, but Kühnel (2015, p. 247) gives an *explicit formula* for $[\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}]\bullet\mathbf{y}$ in terms of a sum of *eighteen* sectional curvatures of eighteen planes constructed from linear combinations of the four input vectors of the Riemann curvature!

29.5.9 Historical Notes on the Origin of the Riemann Tensor

Riemann announced the discovery of his tensor, albeit obliquely, on the 10th of June, 1854, in a lecture at Göttingen University—as a mandatory part of his application to become an *unsalaried* faculty member!—delivered to a mainly non-mathematical audience.

The lecture made oracular pronouncements (without providing a shred of evidence) and it contained only a single formula: the metric for an n-manifold of constant curvature. But there was certainly *one* mathematician in the audience who grasped his meaning, the one who had *forced* Riemann to undertake this investigation in the first place: Carl Friedrich Gauss!

As was required by the rules, Riemann had submitted *three* possible topics for the mandatory lecture, the first of which was his recently completed, ground-breaking doctoral work on what are now called ***Riemann surfaces.*** Convention required that Gauss choose this one, which Riemann was well-prepared to address. Gauss did not! Instead, he picked Riemann's *third* choice: "The Foundations of Geometry."

Riemann was thereby forced into a brand new, major investigation, coming right on the heels of his doctoral masterpiece. The strain of this work, perhaps exacerbated by living in poverty, led

[14] In the final section of this chapter we shall see that (surprisingly!) the mere *sum* of these sectional curvatures nevertheless contains critical geometrical information!

Riemann to suffer another of his recurring mental breakdowns. But he recovered, and, following seven more weeks of intense effort, he was ready to deliver his June 10th lecture.

Whether Gauss knew of or cared about Riemann's plight in the run-up to the lecture is not known; what *is* known is how Gauss *reacted* to Riemann's lecture. Dedekind later recalled that Gauss sat at the lecture "which surpassed all his expectations, in the greatest astonishment, and on the way back from the faculty meeting he spoke to Wilhelm Weber, with the greatest appreciation, and with an excitement rare for him, about the depth of the ideas presented by Riemann."

In 1861 Riemann submitted a Prize Essay to the French Academy of Sciences that included a more mathematically explicit account of the Riemann tensor. But the essay—which did not win the prize!—concerned *heat conduction*, and Riemann gave only the briefest of hints regarding its significance to *geometry*.

Neither this essay nor the original 1854 lecture were published until after Riemann's death in 1866 (at the age of 39, from tuberculosis). Felix Klein (1928) summed up the reaction to their eventual publication in 1868: "I still have vivid memories of the extraordinary impression that Riemann's trains of thought made on young mathematicians. Much of it seemed obscure and hard to understand and yet of unfathomable depth."

Now we come face to face with the *same* mystery that opened Act IV, there regarding Einstein, but now regarding Riemann: Levi-Civita discovered parallel transport fully *63 years* after Riemann's lecture, so the modern geometrical interpretation of the Riemann tensor that we have provided would have been out of reach for Riemann himself. So how *did* Riemann discover his tensor?

The answer is that nobody knows! Spivak (1999) provides an English translation of the 1854 lecture, along with explanatory notes in which Spivak speculates on how Riemann *might* have done it. Spivak also includes an analysis of Riemann's Prize Essay of 1861.

What *is* clear from Riemann's own words is that his *initial* interpretation of his tensor was that *it measures the deviation of the* n*-manifold's metric from the Euclidean metric.* More precisely, Riemann discovered the following remarkably direct link between his tensor and the metric. If ds is the distance between the point p with Cartesian coordinates x_i and the neighbouring point with coordinates $(x_i + dx_i)$, then the difference between the actual metric and the Euclidean metric is given by this remarkable formula:

$$ds^2 - [dx_1^2 + \cdots + dx_n^2] \asymp \frac{1}{12} \sum_{i,j,k,m} R_{ijkm}(p)\,(x_i\,dx_j - x_j\,dx_i)\,(x_k\,dx_m - x_m\,dx_k).$$

For more on this, see Spivak (1999) and Berger (2003, §4.4).

The Mystery of Riemann's Curvature (Darrigol 2015) is a fascinating and thought-provoking investigation of these questions. Darrigol's analysis is based, in part, on photographs of previously unpublished, private notes by Riemann, which had been sitting in the archives of Göttingen University for more than 150 years.

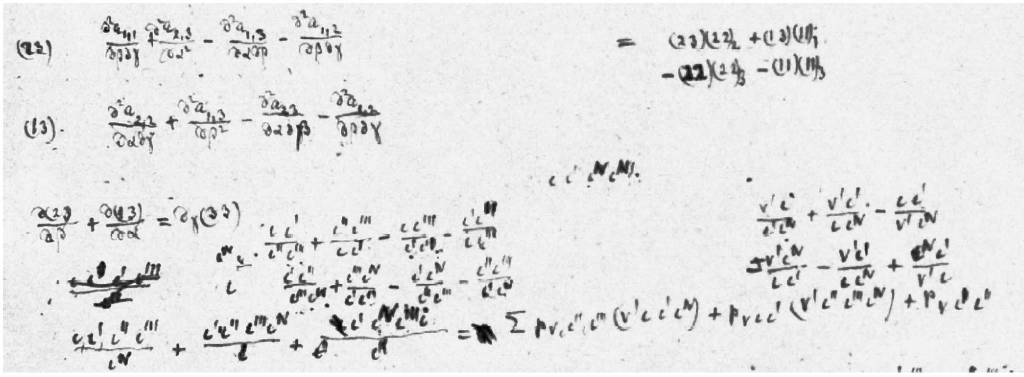
One such scrap of paper is shown in [29.10]. Darrigol's painstaking decoding of Riemann's notation allows him to enter a fascinating correction into the annals of the history of mathematics. For that long, unintelligible final line turns out to be none other than the *Differential Bianchi Identity*, (29.17), which would be rediscovered decades later—first[15] by Aurel Voss (1880), then Ricci (1889), and then Bianchi (1902)—becoming one of the cornerstones[16] of Einstein's General Relativity!

For more on the life and science of Riemann, we recommend Monastyrskiĭ (1999), and Laugwitz (1999).

---

[15]See Pais (1982, pp. 275–276).

[16]See Exercise 14.



[29.10] *Extract of folio 4 of Cod. Ms. B. Riemann 9; courtesy of Göttingen University Archive. The final line is the* Differential Bianchi Identity, *(29.17), anticipating others by decades! Extract of folio 4 of Cod. Ms. B. Riemann 9; courtesy of Göttingen University.*

29.6 The Jacobi Equation in an n-Manifold

29.6.1 Geometrical Proof of the Sectional Jacobi Equation

Within a 2-surface, the Jacobi Equation (the Equation of Geodesic Deviation) (28.2) describes the attraction (or repulsion) of neighbouring geodesics that pass through a region of positive (or negative) curvature, respectively.

In a 2-surface, this attraction or repulsion *only* depends on the curvature of the surface *at* the point through which the geodesics pass: it does *not* depend on the *direction* in which the geodesics pass through the point.

Within a 3-manifold, the attraction or repulsion again only depends on the (instantaneous) *plane* within which the neighbouring geodesics are travelling, not the direction of the geodesics within that plane. This plane $\Pi = \Pi(\mathbf{v}, \boldsymbol{\xi})$ is spanned by the velocity vector $\mathbf{v}$ and the connecting vector $\boldsymbol{\xi}$. But whereas there was only one possible "choice" for the plane for a 2-surface, now there infinitely many such planes within our 3-manifold (or n-manifold), and, as already illustrated in [29.2], the sectional curvature in these planes can be positive in one and negative in another, causing attraction in one, and repulsion in the other.

It is therefore natural to guess that the generalization of the Jacobi Equation (28.2) to an n-manifold simply replaces the Gaussian curvature $\mathcal{K}$ with the *sectional curvature* $\mathcal{K}(\Pi)$ of the instantaneous plane within which the two neighbouring geodesics are travelling, and we shall now give two proofs that this is correct, but with a twist: *we must focus on the component of the relative acceleration that lies within* Π.

Whereas before we used a Newtonian superscript dot to denote the derivative along the geodesics, we now use the more standard $\nabla_{\mathbf{v}}$. Therefore, the relative velocity is now written $\nabla_{\mathbf{v}}\,\boldsymbol{\xi}$, and the relative acceleration, $\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}$.

Reconsider our original proof of the Jacobi Equation in [28.7], but now imagine that the rectangle around which we parallel-transport the velocity lies within an arbitrary plane Π in our 3-manifold (or n-manifold). As before, parallel-transporting $\mathbf{v}_{||}$ around the loop with orthogonal edges $\boldsymbol{\xi}$ and $\mathbf{v}\delta t$, yields

$$\delta\mathbf{v}_{||} \asymp \delta t\, \nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}.$$

Since the speed of the particles is constant, $\delta\mathbf{v}_{||}$ is orthogonal to $\mathbf{v}$. In the 2-surface [28.7], this implied that it pointed along $\pm\boldsymbol{\xi}$, but in a 3-manifold *this is no longer true*: it can *also* have a component[17] *orthogonal to* Π. In summary, the relative acceleration $\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}$ has a component

[17]In an n-manifold there will be $(n-2)$ components orthogonal to Π.

within Π that is responsible for pulling the geodesics together (or pushing them apart), *and* a component orthogonal to Π that does not affect the separation of the geodesics, but that causes them to *rotate* about one another.

We now focus attention on the attraction/repulsion component within Π, namely $\mathcal{P}[\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}]$. As we have seen in (29.20), *if we project the parallel-transported vector into* Π then the holonomy is once again given by the curvature times the area of the rectangle, only now it is the *sectional curvature*, $\mathcal{K}(\Pi)$.

Thus, our original argument in [28.7] goes through essentially unchanged, except that now (29.19) *only* applies to the *projection* of $\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}$ into Π, yielding,

$$\delta t\,\mathcal{P}[\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}] \asymp \mathcal{P}[\delta \mathbf{v}_{||}] \asymp \mathbf{v}_{\perp}\,\mathcal{K}(\Pi))\,\delta\mathcal{A} = \left[-\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|}\right]\,\mathcal{K}(\Pi)\,|\boldsymbol{\xi}|\,\delta t,$$

and so we obtain what we shall christen[18] the

Sectional Jacobi Equation:
$$\boxed{\boxed{\mathcal{P}\,[\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}] = -\mathcal{K}(\Pi)\,\boldsymbol{\xi}.}} \tag{29.21}$$

NOTE: This is *not* the standard "Jacobi Equation," which will be derived shortly (see (29.24)).

29.6.2 Geometrical Implications of the Sectional Jacobi Equation

To gain an intuitive grasp of the import of this equation, consider [29.11]. A set of particles are arranged around the perimeter of a small circle in a plane. A fiducial particle is added at the centre of the circle, the vector $\boldsymbol{\xi}$ connecting this central particle to a typical particle on the circle.

Now simultaneously fire these particles perpendicularly to the plane, with velocity $\mathbf{v}$. Initially, the particles will move in rigid unison, maintaining the same-size circular pattern. But then the curvature of the space begins to make itself felt, via the Sectional Jacobi Equation, (29.21). The motion of each individual particle on the circle, relative to the central one, is now governed by the *particular* $\mathcal{K}(\Pi)$ for the *particular* plane $\Pi(\mathbf{v}, \boldsymbol{\xi})$ for that *particular* particle's $\boldsymbol{\xi}$.

Thus as we move from a 2-surface to a 3-manifold (or n-manifold), a *fundamentally new feature* emerges: as the plane $\Pi(\mathbf{v}, \boldsymbol{\xi})$ rotates about the central axis (consisting of the central particle's trajectory) the sectional curvature $\mathcal{K}(\Pi)$ *varies*, both in magnitude and (possibly) sign.

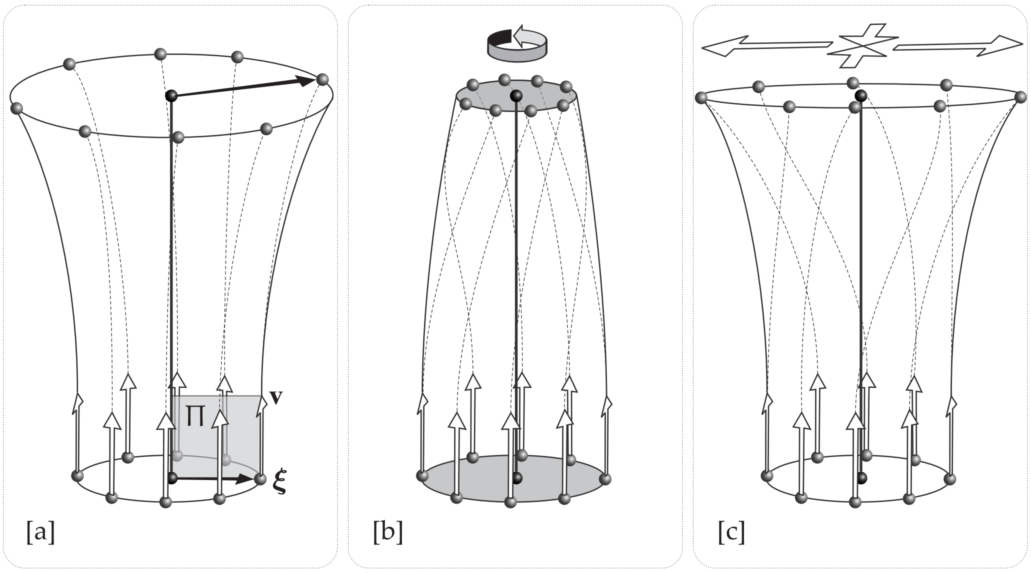
Figure [29.11a] illustrates the case in which all sectional curvatures $\mathcal{K}(\mathbf{v}, \boldsymbol{\xi})$ are *negative*, regardless of the direction of $\boldsymbol{\xi}$, causing the particles to be repelled, accelerating apart on an expanding circle.[19] To use an optical analogy, negative curvature acts like a diverging lens.

Next, [29.11b] illustrates the case in which all curvatures are *positive*. In addition to the focusing attraction, governed by (29.21), here we have illustrated a case in which the relative acceleration has an appreciable component *orthogonal* to Π, causing the bundle of geodesics to rotate about the central one. This effect is intentionally ignored in our Sectional Jacobi Equation, (29.21), but it *is* taken into account by the (upcoming) full Jacobi Equation, (29.24). This is akin to the positive focussing of light by a magnifying lens.

Finally, [29.11c] depicts the most interesting case, and, as we shall see in the next chapter, the one of greatest importance to gravitational physics. Here the sectional curvature is positive in some planes (squeezing the circle of particles together) and negative in others (pulling the circle apart). The originally circular array of particles is therefore deformed into an ellipse, with minor

[18] Despite its twin virtues—(1) of providing a very direct interpretation of the sectional curvature, and (2) of being formally identical to the full Jacobi equation (28.2) in a 2-surface—we have not found this result written down in any standard text. We therefore felt the need to invent a logical name for this formula.

[19] If the negative curvatures were to vary in magnitude, then the circle would become oval as it grew.



[29.11] *A small circle of particles are fired perpendicularly to the circle. Some planes* $\Pi(\mathbf{v}, \boldsymbol{\xi})$ *may have* $\mathcal{K}(\Pi) > 0$, *resulting in* attraction *towards the central particle, while other planes may have* $\mathcal{K}(\Pi) < 0$, *resulting in* repulsion *from the central particle.* **[a]** $\mathcal{K}(\Pi) < 0$ *in all planes.* **[b]** $\mathcal{K}(\Pi) > 0$ *in all planes.* **[c]** *Attraction in some planes squeezes the circle, and repulsion in others stretches it out, so that the circle is deformed into an ellipse.*

axis smaller than the original circle, and major axis larger than it. This is akin to the effects of an astigmatic lens.

In the next chapter we shall see that the optical analogy is not an analogy at all—it is reality! That is, the gravitational curvature of spacetime really *does* act like a lens as light passes through it!

29.6.3 *Computational Proofs of the Jacobi Equation and the Sectional Jacobi Equation*

For our orthonormal basis for Π, let us take $\mathbf{e}_1 = \widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}/|\boldsymbol{\xi}|$ and $\mathbf{e}_2 = \mathbf{v}$. Then $\mathbf{e}_3 \equiv \mathbf{e}_1 \times \mathbf{e}_2$ will be the unit normal to Π. The definition (29.18) of the ***sectional curvature*** can now be written,

$$\mathcal{K}(\Pi) = R_{1221} = [\mathcal{R}(\mathbf{e}_1,\, \mathbf{e}_2)\, \mathbf{e}_2] \bullet \mathbf{e}_1 = [\mathcal{R}(\widehat{\boldsymbol{\xi}},\, \mathbf{v})\, \mathbf{v})] \bullet \widehat{\boldsymbol{\xi}}. \tag{29.22}$$

By definition, $\boldsymbol{\xi}$ always connects the two neighbouring geodesics, so the parallelogram with edges $\mathbf{v}$ and $\boldsymbol{\xi}$ closes up:

$$[\mathbf{v}, \boldsymbol{\xi}] = \mathbf{0} \qquad \Longleftrightarrow \qquad \nabla_{\mathbf{v}}\boldsymbol{\xi} = \nabla_{\boldsymbol{\xi}}\mathbf{v}. \tag{29.23}$$

Note (for use in just a moment) that this implies that the Riemann curvature operator (29.6) simplifies to

$$\mathcal{R}(\mathbf{v}, \boldsymbol{\xi}) = [\nabla_{\mathbf{v}}, \nabla_{\boldsymbol{\xi}}].$$

Taking the intrinsic derivative ($\nabla_{\mathbf{v}}$) of (29.23) along the geodesics, and recalling that $\mathbf{v}$ satisfies the geodesic equation, $\nabla_{\mathbf{v}}\mathbf{v} = \mathbf{0}$, we find that

$$\begin{aligned} \nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi} &= \nabla_{\mathbf{v}}\nabla_{\boldsymbol{\xi}}\,\mathbf{v} \\ &= [\nabla_{\mathbf{v}}, \nabla_{\boldsymbol{\xi}}]\,\mathbf{v} + \nabla_{\boldsymbol{\xi}}(\nabla_{\mathbf{v}}\mathbf{v}) \\ &= \mathcal{R}(\mathbf{v}, \boldsymbol{\xi})\,\mathbf{v} + \nabla_{\boldsymbol{\xi}}\,(\mathbf{0}). \end{aligned}$$

Thus we have obtained the ***Equation of Geodesic Deviation***, or the

$$\textit{Jacobi Equation:} \qquad \boxed{\boxed{\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi} = -\mathcal{R}(\boldsymbol{\xi},\,\mathbf{v})\,\mathbf{v}.}} \tag{29.24}$$

This is the standard form of the Jacobi equation, which is found in all texts.

To instead recover our more intuitive "sectional" form, we decompose the vector on the right-hand side of the equation into its components, and find,

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\xi},\,\mathbf{v})\,\mathbf{v} &= \left\{[\mathcal{R}(\boldsymbol{\xi},\,\mathbf{v})\,\mathbf{v}]\bullet\widehat{\boldsymbol{\xi}}\right\}\widehat{\boldsymbol{\xi}} + \{[\mathcal{R}(\boldsymbol{\xi},\,\mathbf{v})\,\mathbf{v}]\bullet\mathbf{v}\}\,\mathbf{v} + \left\{\left[\mathcal{R}(|\boldsymbol{\xi}|\widehat{\boldsymbol{\xi}},\,\mathbf{v})\,\mathbf{v}\right]\bullet\mathbf{e}_3\right\}\mathbf{e}_3 \\
&= \left\{\left[\mathcal{R}(\widehat{\boldsymbol{\xi}},\,\mathbf{v})\,\mathbf{v}\right]\bullet\widehat{\boldsymbol{\xi}}\right\}\boldsymbol{\xi} + \mathbf{0} + \{[\mathcal{R}(\mathbf{e}_1,\,\mathbf{e}_2)\,\mathbf{e}_2]\bullet\mathbf{e}_3\}\,|\boldsymbol{\xi}|\,\mathbf{e}_3 \\
&= \mathcal{K}(\Pi)\,\boldsymbol{\xi} + R_{1223}|\boldsymbol{\xi}|\mathbf{e}_3,
\end{aligned}$$

by virtue of (29.22).

The second term, $R_{1223}|\boldsymbol{\xi}|\mathbf{e}_3$, is the component of the relative acceleration orthogonal to Π, which is responsible for the *rotation* of the bundle of geodesics, as illustrated in [29.11b].

The first term represents the *attraction* or *repulsion* within Π. Indeed, applying the projection operator $\mathcal{P}$ to both sides of the Jacobi Equation, (29.24), we recover the Sectional Jacobi Equation, (29.21):

$$\boxed{\mathcal{P}\,[\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\,\boldsymbol{\xi}] = -\mathcal{P}\,[\mathcal{R}(\boldsymbol{\xi},\,\mathbf{v})\,\mathbf{v}] = -\mathcal{K}(\Pi)\,\boldsymbol{\xi}.}$$

29.7 The Ricci Tensor

29.7.1 Acceleration of the Area Enclosed by a Bundle of Geodesics

Reconsider [29.11b]. The shaded area $\delta\mathcal{A}$ of the disc (inside the circle of particles) clearly shrinks as the particles are pulled together by the positive sectional curvatures. But according to what law?

For simplicity's sake, let us initially suppose that as Π rotates about the central particle's trajectory, all the sectional curvatures have the *same* positive value $\mathcal{K}$. Now let $r(t) = |\boldsymbol{\xi}(t)|$, so that $\delta\mathcal{A}(t) = \pi[r(t)]^2$ is the area at time t after launch. Reverting to Newtonian dots for time derivatives, it is clear that the initial rate of change of the area *vanishes*, because the particles head off in rigid unison, so that $\dot{r}(0) = 0$, and therefore

$$(\dot{\delta\mathcal{A}})(0) = 2\pi r(0)\,\dot{r}(0) = 0.$$

The sectional Jacobi Equation (29.21) tells us that the acceleration of the particles towards the centre is $\ddot{r}(0) = -\mathcal{K}r(0)$. Therefore the curvature manifests itself in the initial *acceleration* of the area:

$$(\ddot{\delta\mathcal{A}})(0) = 2\pi\,[\dot{r}^2(0) + r(0)\,\ddot{r}(0)] = -2\pi\,\mathcal{K}[r(0)]^2 = -2\mathcal{K}\,\delta\mathcal{A}(0).$$

Therefore, for small t, the Maclaurin expansion shows that the area decreases in proportion to the curvature and as the square of the time:

$$\text{change in the area after time } t = \delta\mathcal{A}(t) - \delta\mathcal{A}(0) \asymp -\mathcal{K}\,\delta\mathcal{A}(0)t^2.$$

Of course if $\mathcal{K}$ were *negative*, as in [29.11a], then the right-hand side would be positive, corresponding accelerating *growth* of the area.

Now that we have a rough idea of what is happening, let us move on to the general case, in which $\mathcal{K}(\Pi)$ varies in magnitude (and possibly sign) as Π rotates about the central axis.

Let θ denote the angle around this axis, starting from an arbitrary initial $\boldsymbol{\xi}$-direction. Consider the narrow (ultimately vanishing) sector of the disc located at angle θ and of angular width $\delta\theta$. The analysis above still applies to this sector, provided we replace $\mathcal{K}$ with the sectional curvature $\mathcal{K}(\theta)$ of the *particular* plane $\Pi(\theta)$ in this direction.

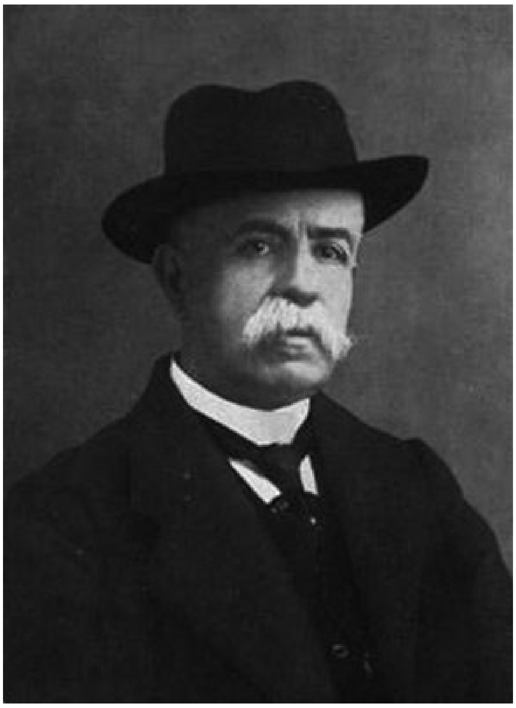
It follows that, in the general case, the previous equation becomes

$$\delta\mathcal{A}(t) - \delta\mathcal{A}(0) \asymp -\mathcal{K}_{\text{mean}}\, \delta\mathcal{A}(0) t^2,$$

where

$$\mathcal{K}_{\text{mean}} \equiv \frac{1}{2\pi}\int_0^{2\pi} \mathcal{K}(\theta)\, d\theta. \tag{29.25}$$

Here, $\mathcal{K}_{\text{mean}}$ is the *average* (or "mean") of the sectional curvatures in all the planes containing the central particle's trajectory, in the direction $\mathbf{v}$.

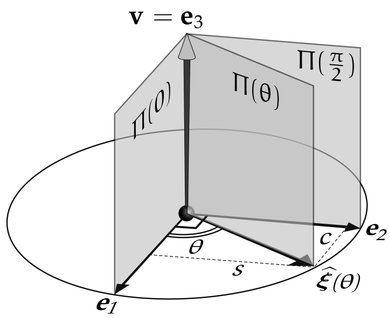


[29.12] *Gregorio Ricci-Curbastro (1853–1925), who later in life became known simply as* Gregorio Ricci *(pronounced "reechee").*

Next, we shall find a formula for $\mathcal{K}(\theta)$ in terms of the Riemann tensor. To do so, let us choose $\mathbf{e}_3 = \mathbf{v}$ to be the direction of the central axis, in the direction that the particles are launched. See [29.13]. Then $\mathbf{e}_1$ and $\mathbf{e}_2$ are orthogonal unit vectors in the plane of the circle of particles. Having chosen the direction of $\mathbf{e}_1$ *arbitrarily*, let us agree to measure θ from this direction, as shown.

The plane $\Pi(\theta)$ is then spanned by $\mathbf{v}$ and the unit vector $\widehat{\boldsymbol{\xi}}(\theta)$ in the direction of the particle at angle θ:

$$\widehat{\boldsymbol{\xi}}(\theta) = \cos\theta\, \mathbf{e}_1 + \sin\theta\, \mathbf{e}_2 \equiv c\, \mathbf{e}_1 + s\, \mathbf{e}_2,$$



[29.13]

for short. Then the sectional curvature of $\Pi(\theta)$ is given by (29.22):

$$\begin{aligned}
\mathcal{K}(\theta) &= [\mathbf{R}(\widehat{\boldsymbol{\xi}}, \mathbf{v}; \mathbf{v})] \cdot \widehat{\boldsymbol{\xi}} \\
&= [\mathbf{R}(c\,\mathbf{e}_1 + s\,\mathbf{e}_2, \mathbf{v}; \mathbf{v})] \cdot (c\,\mathbf{e}_1 + s\,\mathbf{e}_2) \\
&= [c\,\mathbf{R}(\mathbf{e}_1, \mathbf{v}; \mathbf{v}) + s\,\mathbf{R}(\mathbf{e}_2, \mathbf{v}; \mathbf{v})] \cdot (c\,\mathbf{e}_1 + s\,\mathbf{e}_2) \\
&= c^2\,\mathbf{R}(\mathbf{e}_1, \mathbf{v}; \mathbf{v}) \cdot \mathbf{e}_1 + s^2\,\mathbf{R}(\mathbf{e}_2, \mathbf{v}; \mathbf{v}) \cdot \mathbf{e}_2 + sc\,[\mathbf{R}(\mathbf{e}_1, \mathbf{v}; \mathbf{v}) \cdot \mathbf{e}_2 + \mathbf{R}(\mathbf{e}_2, \mathbf{v}; \mathbf{v}) \cdot \mathbf{e}_1] \\
&= c^2\,\mathcal{K}(0) + s^2\,\mathcal{K}(\tfrac{\pi}{2}) + 2scR_{1332},
\end{aligned}$$

where the last term follows from the symmetry [exercise], $R_{2331} = R_{1332}$.

To compute $\mathcal{K}_{\text{mean}}$, first verify that

$$\int_0^{2\pi} \cos^2\theta \, d\theta = \int_0^{2\pi} \sin^2\theta \, d\theta = \pi, \qquad \text{and} \qquad \int_0^{2\pi} \sin\theta \, \cos\theta \, d\theta = 0.$$

Then,

$$\boxed{\mathcal{K}_{\text{mean}} = \frac{\mathcal{K}(0) + \mathcal{K}(\frac{\pi}{2})}{2}.} \tag{29.26}$$

Recall that the direction $\theta = 0$ was chosen *arbitrarily*, so we have proved the remarkable fact that the average of the sectional curvatures over *all* directions is obtained by averaging it in *any two orthogonal* directions.

Combining (29.25) with (29.26), we may sum up our findings as follows:

If a small circle of particles are fired with velocity **v** *perpendicularly to the circle, the acceleration of the area* $\delta\mathcal{A}$ *of the circle is controlled by the average of the sectional curvatures in all planes containing* **v***, but this is given by* the sum of the sectional curvatures in *any* two orthogonal planes containing **v***:*

$$(\ddot{\delta\mathcal{A}})(0) = -[\mathcal{K}(0) + \mathcal{K}(\tfrac{\pi}{2})] \, \delta\mathcal{A}(0), \tag{29.27}$$

and so

$$\delta\mathcal{A}(t) - \delta\mathcal{A}(0) \asymp -\frac{1}{2} [\mathcal{K}(0) + \mathcal{K}(\tfrac{\pi}{2})] \, \delta\mathcal{A}(0) t^2$$

29.7.2 Definition and Geometrical Meaning of the Ricci Tensor

Let us focus our attention on this sum of sectional curvatures in orthogonal coordinate planes containing the velocity $\mathbf{v} = \mathbf{e}_3$ of the particles. Recalling that $R_{3333} = 0$, by antisymmetry, we may write this as

$$\begin{aligned} \mathcal{K}(0) + \mathcal{K}(\tfrac{\pi}{2}) &= \mathbf{R}(\mathbf{e}_1, \mathbf{e}_3; \mathbf{e}_3) \cdot \mathbf{e}_1 + \mathbf{R}(\mathbf{e}_2, \mathbf{e}_3; \mathbf{e}_3) \cdot \mathbf{e}_2 \\ &= R_{1331} + R_{2332} + R_{3333} \\ &= R_{m33}{}^{m}. \end{aligned}$$

In a moment, this expression will guide us to the introduction of a brand new tensor, the *Ricci curvature tensor*. Whereas the Riemann tensor takes three vectors as input, and outputs a vector, the new Ricci tensor takes *two* vectors as input, and outputs a *scalar*.

NOTE ON NOTATION: It is an unfortunate accident of history that both the names Riemann and Ricci begin with the letter "R"! This has resulted in the universally accepted notation whereby the components of the Ricci tensor are denoted R_{jk}. We dare not defy this tradition—and do not even wish to!—but please be on high alert that an R with *four* indices is a component of the *Riemann* tensor, while an R with *two* indices is a component of the *Ricci* tensor. That said, it would be simply *too* confusing to represent these two (very different) geometrical tensors with the *same* symbol, **R**!

Therefore, we shall adopt (essentially[20]) the notation employed by Misner, Thorne, and Wheeler (1973), reserving **R** for the Riemann tensor, and writing the geometrical Ricci tensor itself as **R**icci.

The definition of the ***Ricci curvature tensor*** is this:

$$\boxed{\mathbf{R}\mathrm{icci}(\mathbf{v}, \mathbf{w}) \equiv \sum_{m=1}^{n} \mathbf{R}(\mathbf{e}_m,\, \mathbf{v};\, \mathbf{w}) \cdot \mathbf{e}_m \quad \Longleftrightarrow \quad R_{jk} = \mathbf{R}\mathrm{icci}(\mathbf{e}_j, \mathbf{e}_k) = R_{mjk}{}^{m}.}$$

CONVENTION WARNING: Many authors instead define $R_{jk} \equiv R_{jmk}{}^{m} = -R_{mjk}{}^{m}$, i.e., the negative of ours. Notably, this is true of Misner, Thorne, and Wheeler (1973). However, as we noted earlier, their definition of **R** is *also* the opposite of ours, and therefore the two sign-convention disagreements *cancel*: the sign of their ***Ricci*** is the *same* as our **R**icci! As we shall see in a moment, *positive* Ricci curvature corresponds to *attraction* (both for us and for Misner, Thorne, and Wheeler). On the other hand, to cite just one other important example, Penrose (2005) shares our definition of **R** but uses the opposite definition of **R**icci, so for Penrose it is *negative* Ricci curvature that causes attraction.

The fact that **R**icci is indeed a tensor in its own right follows immediately [exercise] from the the fact that **R** is a tensor. It also follows from the symmetries of **R** that **R**icci is *symmetric*:

$$\boxed{\mathbf{R}\mathrm{icci}(\mathbf{w}, \mathbf{v}) = \mathbf{R}\mathrm{icci}(\mathbf{v}, \mathbf{w}) \quad \Longleftrightarrow \quad R_{kj} = R_{jk}.} \tag{29.28}$$

To see this, swap the *first pair* of vectors in its Riemann-tensor definition with the *second pair*, then swap the vectors within each pair.

We can now see that

$$\mathcal{K}(0) + \mathcal{K}(\tfrac{\pi}{2}) = R_{m33}{}^{m} = \mathbf{R}\mathrm{icci}(\mathbf{v}, \mathbf{v}) = R_{jk} v^j v^k.$$

We have therefore arrived at the beautifully simple, Jacobi-type equation that governs the acceleration of the area of the bundle of geodesics launched with velocity **v**:

$$\boxed{\ddot{\delta\mathcal{A}} = -\mathbf{R}\mathrm{icci}(\mathbf{v}, \mathbf{v})\, \delta\mathcal{A}.}$$

The role of the Gaussian curvature in the Jacobi equation of a 2-surface (or sectional curvature in an n-manifold) is here taken over by the Ricci curvature. We stress that *positive* Ricci curvature causes *attraction*, resulting in *shrinking* area as the geodesics converge.

Now imagine the generalization of the foregoing analysis to a 4-manifold. Again let **v** be the velocity with which we will launch a group of particles. This group is again chosen to be equidistant from the central particle, and located in the space orthogonal to **v**. But instead of yielding a *circle* in a *plane*, this construction now yields a small *sphere* of particles, enclosing a *volume* $\delta\mathcal{V}$, in the *3-dimensional space* orthogonal to **v**!

Although we shall not work out the details, the foregoing analysis goes through unchanged, and, specifically, the Ricci curvature now dictates the acceleration of the *volume* of the small sphere of particles that are simultaneously launched with velocity **v**:

---

[20]Verily, in the Bible it is written, ***Ricci*** (in all bold, and italics). By instead only putting the "**R**" in (nonitalic) bold, we hope to mentally connect **R**icci to its components, R_{jk}.

$$\boxed{\ddot{\delta\mathcal{V}} = -\mathbf{R}\text{icci}(\mathbf{v}, \mathbf{v})\, \delta\mathcal{V}.} \tag{29.29}$$

We stress for future use that *positive* Ricci curvature causes *attraction*, resulting in *shrinking* volume as the geodesics converge.

As before, this implies that (initially) the *volume changes in proportion to the Ricci curvature, and as the square of the time*:

$$\boxed{\delta\mathcal{V}(t) - \delta\mathcal{V}(0) \asymp -\tfrac{1}{2}\,\mathbf{R}\text{icci}(\mathbf{v}, \mathbf{v})\, \delta\mathcal{V}(0)t^2} \tag{29.30}$$

29.8 Coda

This chapter has introduced a wealth of new ideas and results, but one formula is conspicuous by its absence: an analogue of (27.1) for expressing the Riemann curvature of an n-manifold in terms of its metric. The principal reason for this omission was a disciplined focus on providing the *minimal* set of concepts and results that will be needed to understand Einstein's theory of gravity. A second reason, however, is that the formula for **R** in terms of the metric coefficients turns out to be extremely complicated, at least when expressed in standard tensor formalism. In Act V, we shall see how the calculation of **R** can be greatly simplified—at both the theoretical level and the pragmatic, computational level—by the use of Forms.

As you began this chapter, you may have been preemptively flummoxed by the very notion that space could be *curved*—and who could blame you! But we very much hope that the multiple geometrical interpretations we have provided (especially via the Sectional Jacobi Equation and the Ricci tensor) have not only put you at your ease regarding Riemann's curvature within a 3-manifold, but have actually given you a very concrete, *tangible* grasp of exactly what this curvature *does*.

But 3-manifolds were merely the warm-up act. For it is a psychological *illusion* that you are sitting in a 3-dimensional room. In *reality*, you and your room are hurtling into the future, within Einstein's *4*-dimensional spacetime. It is to the curvature of *that* manifold—what we call "reality"—that we now turn.

Study formula (29.29) well, for we shall see that it is *the key to the Universe!*

Chapter 30

Einstein's Curved Spacetime

30.1 Introduction: *"The Happiest Thought of My Life."*

This chapter lowers the curtain on Act IV—*it is the end of our self-contained introduction to (Visual) Differential Geometry.* We mark this milestone by returning to the subject that *began* Act IV: Einstein's extraordinarily beautiful *General Theory of Relativity*, which declares that gravity *is* the curvature of the 4-dimensional spacetime that comprises the framework of what we call "reality."

The great American physicist John Archibald Wheeler (1911–2008) famously distilled the theory down to a single sentence: *"Space tells matter how to move and matter tells space how to curve."* In greater detail, free-falling matter moves along the *geodesics* of the curved spacetime: these are still the *straightest* paths, but they now *maximize* (instead of minimize) the "distance" as measured by the spacetime metric. The second half of Wheeler's aphorism is the grand finale of Act IV: the precise law that Einstein discovered in 1915 that describes *how* matter and energy curve spacetime—the *Gravitational Field Equation* of General Relativity.

We begin at the historical beginning, with the (probably oversimplified[1]) tale of Newton's discovery of the inverse-square law of gravity, which would have us picture him sitting in his garden, in 1666, watching an apple fall from a tree. Suddenly, so the story goes, Newton realized that the *same* force that had pulled the apple to Earth might reach out to the Moon, pulling it towards the Earth, holding it in its orbit.

The genesis story of Einstein's geometrical theory of gravity is remarkably similar, but Einstein's falling object was not an apple, it was a *man!*

In 1907, fully two years after his discovery of Special Relativity, Einstein still held no academic position. He had spent the previous *five years* as a patent clerk (third class, promoted to second class) at the Federal Office for Intellectual Property, Bern, Switzerland.

As we noted at the outset of Act IV, Einstein knew that despite the superb accuracy of Newton's gravitational law, its instantaneous action-at-a-distance was fundamentally incompatible with the finite speed of light, and hence with his 1905 discovery of Special Relativity. Thus began Einstein's long struggle to understand gravitation, which would end almost exactly eight years later, on the 25th of November, 1915, when he finally wrote down the *Gravitational Field Equation*.

Einstein took the first, crucial step of that long journey in November of 1907, when he had what he later described as *"the happiest thought of my life"*:[2]

> I was sitting in a chair in the patent office at Bern when all of a sudden a thought occurred to me: If a person falls freely he will not feel his own weight . . . I was startled. This simple thought made a deep impression on me. It impelled me towards a theory of gravitation.

When we watch astronauts on TV, floating about within their orbiting space station, it appears that gravity has been totally eliminated—this is precisely the import of Einstein's insight. This *seemingly* total elimination of gravity arises from the fact that *the astronauts and their spacecraft are*

[1]Newton himself did, however, tell this story as related here to four different people on four different occasions. See Westfall (1980, pp. 154–155).

[2]See (Pais 1982 §9).

falling freely together in the Earth's gravitational field—it does *not* arise from the spacecraft being so far from Earth that it has escaped the force of gravity!

Yes, Newton's inverse-square law tells us that gravity is weaker up there than down here, but a quick calculation shows that the strength of the Newtonian gravitational field pulling on the International Space Station (orbiting at an altitude of 250 miles) is only about 12% weaker than it is for us down here. No, the seemingly *complete* elimination of gravity has nothing to do with this modest effect.

Instead, the *total* elimination of gravity is the result of the remarkable empirical fact—first discovered by Galileo around 1590, by (supposedly) dropping objects from the Leaning Tower of Pisa—that *all objects are accelerated equally by gravity, regardless of their mass or composition*.[3] Thus the astronauts and their spacecraft all follow the *same*, free-falling orbit *together*.

This empirical fact is simultaneously enshrined in and explained by Newton's Law of Gravitation. If a particle of very small mass m lies at distance r from a particle with a very large mass M, then the gravitational effect of m on M can be neglected, and we may think of m as simply being pulled towards M (with force F) by the latter's gravitational field. Then the acceleration a of m towards M is given by Newton's Second Law of Motion:

$$F = ma.$$

But if r is the distance of m from M, then ***Newton's Inverse-Square Law of Gravitation*** states that

$$\boxed{\boxed{F = \frac{GmM}{r^2},}}$$

where G is the ***gravitational constant***. Combining these facts,

$$m\ddot{r} = -\frac{GmM}{r^2} \quad \Longrightarrow \quad \boxed{\ddot{r} = -\frac{GM}{r^2},} \tag{30.1}$$

which is independent of m.

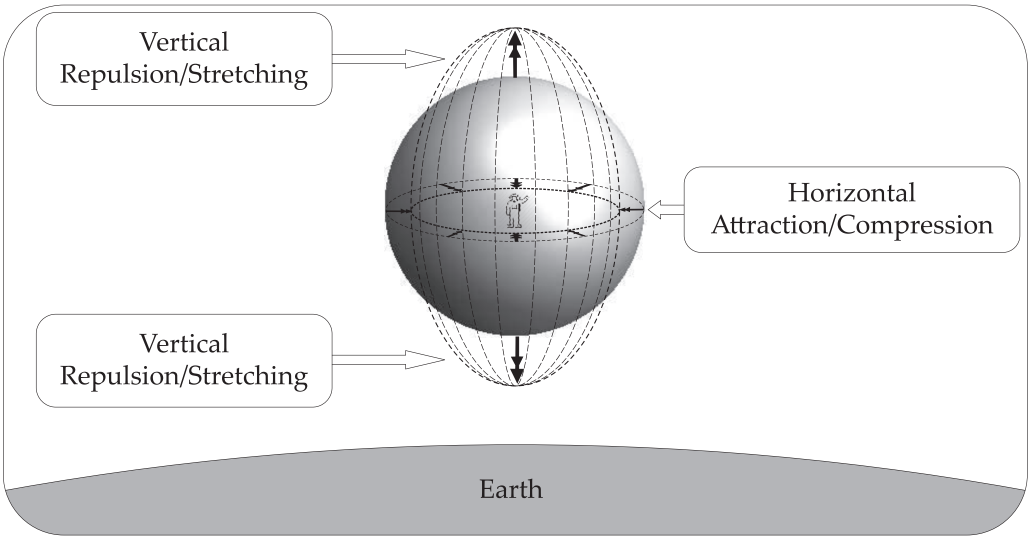
So Newton's gravitational law *explains* Galileo's otherwise deeply mysterious empirical fact that all objects (launched from the same place in the same manner) follow the *same* trajectory, regardless of their composition.

But is it possible to go even deeper? Can we in turn *explain* this aspect of Newton's gravitational law? Einstein's geometrical theory does exactly that! If a particle is destined to travel along the *geometrically* determined geodesics of spacetime—*which have nothing whatsoever to do with the particle!*—this *explains* Galileo's empirical fact, and likewise provides the *reason* that the gravitational force F *must* be proportional to the mass, m!

30.2 Gravitational Tidal Forces

But what exactly *remains* of gravity if it seemingly vanishes for the astronauts within their spacecraft, all free-falling together? Surely it must leave *some* trace of its existence?! This is the fundamentally new question to which we are led by Einstein's startling insight.

[3]This is only true if air resistance can be ignored. There is a wonderful film (from the 1971 Apollo 15 mission) of astronaut David Scott simultaneously dropping a feather and a hammer while standing on the surface of the Moon—they both hit the lunar surface at the same time!



[30.1] The gravitational tidal forces of the Earth act on a free-falling sphere of particles*, resulting in* compression in the horizontal "equatorial" plane, parallel to the Earth's surface, and stretching in the vertical direction of the gravitational field itself. *As the sphere falls, it therefore starts to deform into an ellipsoidal egg. For the inverse-square law, the vertical stretching turns out to be exactly* twice *as powerful as the equatorial compression.*

Well, imagine that you yourself are that astronaut, high above the Earth, but rather than being safely in orbit inside the space station, you are outside, in your spacesuit, using your jetpack to hover at a fixed location. Further imagine that you are oriented straight up and down, with your head furthest from Earth, and your feet closest to it. Surrounding you, evenly spread over the surface of a sphere centred on you, are hundreds of shiny ball bearings, momentarily at rest, like you. See [30.1].

Now you turn off your jetpack. You and the sphere of ball bearings surrounding you begin to free-fall *together*, from rest, all accelerating downward towards the Earth. What *changes*, if any, will you observe in the sphere of particles as you fall?

If the gravitational field were *perfectly uniform*, with constant magnitude and fixed direction, your answer would be, "I see no changes at all!" Gravity would be truly invisible to you. And since the actual gravitational field of the Earth is *approximately* uniform over the small sphere, you won't see any changes, either . . . at least not *at first*.

However, slowly, you will start to notice that the parts of the sphere above your head and below your feet are accelerating *away* from you, while the "equatorial" particles are accelerating *towards* to you. The net result is that the sphere starts turning into an egg!

To understand why this is happening, see [30.1], and let us immediately clarify that this diagram shows the *forces* that are going to *drive* the deformation of the sphere into the egg, not the egg itself, which is the new shape of the sphere after it has fallen down towards the Earth for a period of time.

First, realize that the equatorial particles, at the same altitude as you, are not accelerating in precisely the same *direction* that you are. Yes, you and the particles are accelerating downward, and at the same rate, but you are all *accelerating towards the centre of the Earth*,[4] and therefore your

[4]In reality, you are being pulled towards every little part of the Earth, with a force that is inversely proportional to the square of the distance of that particular part from you. But it turns out that the *net* effect of this horrendously complex sum of forces is *precisely* as though all the mass of the Earth were concentrated at its centre! Newton himself was shocked to discover this genuinely miraculous property of his inverse-square law, and he provided an elegant geometrical proof of it in the *Principia* (Newton 1687, Theorem 31). Note that this is another *signature* of the inverse-square law—it is *not* shared by other force laws.

trajectories are slowly *converging*. Since you only see the *relative acceleration*, the equatorial particles appear to be accelerating directly towards you, as if *attracted* to you, as illustrated in [30.1].

Now look up at the particles above your head at the top of the sphere, and below your feet at the bottom of the sphere. Unlike the previous case, you and the particles *are* now all heading in the same direction, straight down towards the centre of the Earth. However, the acceleration of the particles below your feet is *greater* than your acceleration, because they are closer to the centre of attraction than you are. Thus you see them start to accelerate *away* from you downward towards the Earth. Likewise, the acceleration of the particles above your head is *smaller* than your acceleration, because they are further away from the centre of attraction, so you are accelerating downward faster than they are, so they too seem to accelerate away from you, *upward*. Thus these particles appear to be *repelled* from you, as illustrated in [30.1].

This phenomenon, of compression in one plane and stretching in the orthogonal direction, deforming a sphere into an egg, is called the gravitational ***tidal force***. It is the *answer* to our Einstein-inspired question, "What remains of gravity in free-fall?"

Let us connect these observations more directly to our original astronauts in orbit around the Earth. Instead of falling from rest, suppose that you and your sphere of particles all blast off together with an arbitrary but shared velocity; the egg-inducing forces will be exactly the same as before! In particular, suppose this velocity is chosen to point horizontally, parallel to the Earth's surface, and suppose the launch speed is chosen so as to set you and your sphere into a *circular orbit* around the Earth. Finally, imagine that you and your sphere are inside the orbiting space station, and the ball bearings are replaced with fellow astronauts! If you look very closely, we will see that gravity has *not* been completely eliminated: a vertically (i.e., radially) separated astronaut will gradually accelerate *away* from you, while a horizontally separated astronaut will gradually accelerate *towards* you. However, as we shall see in the next section, these tidal forces are *proportional to the separation* of the particles, and over the small distances within the space station they are undetectable. But over larger scales, the impact of these tidal forces becomes obvious, which brings us to our next topic. . . .

Why is this gravitational distortion force field called "tidal"? What is the connection between this phenomenon and the *tides* of our oceans?

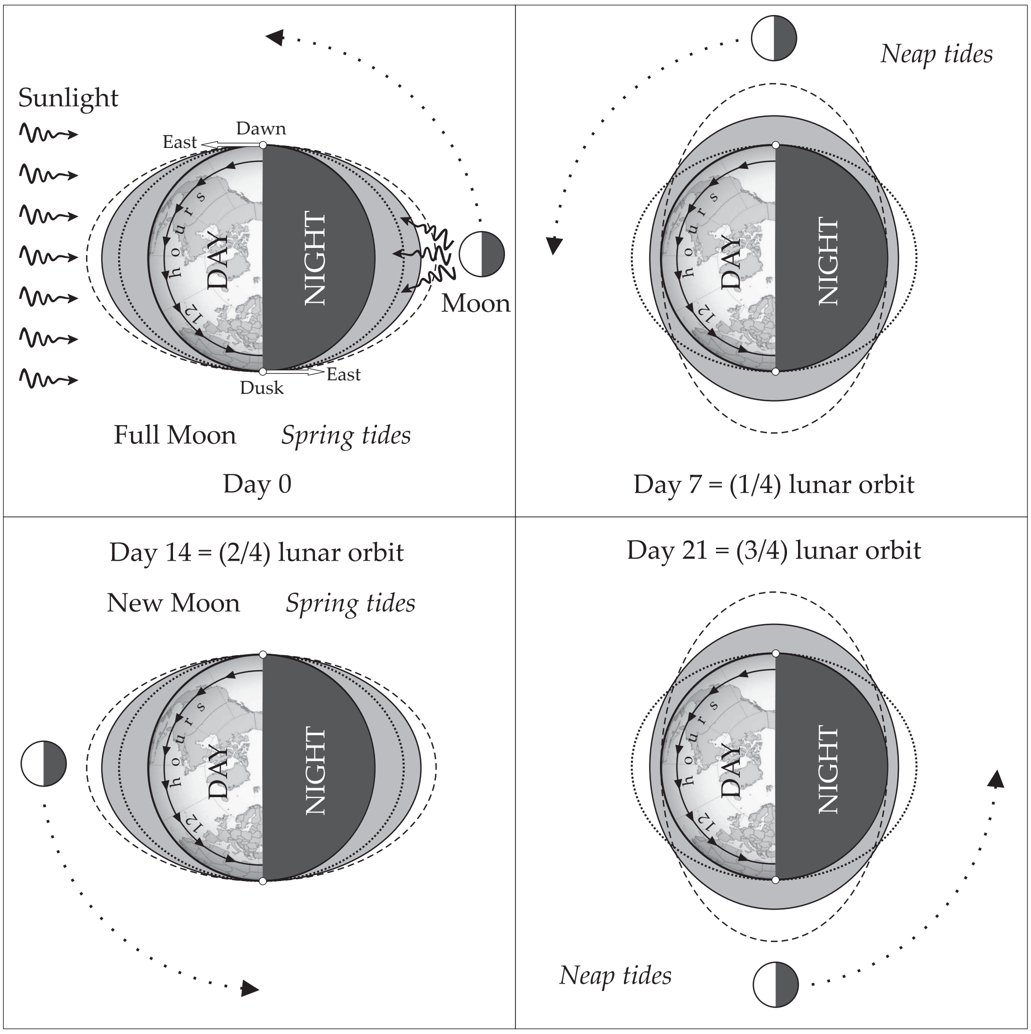
Imagine in [30.1] that the Earth is instead the *Sun*, and our small sphere is instead the *Earth*, or rather the thin skin of the Earth—its *oceans*. The tidal force of the Sun on the Earth now causes a bulging of the oceans towards the Sun on the side of the Earth facing the Sun, and a bulging of the oceans away from the Sun on the opposite side, representing a rise in sea level in these two opposite directions. This is the fixed, *dotted* ellipsoidal egg shown in all four panels of [30.2].

The directions of the bulges remain fixed as the Earth rotates once every 24 hours, so each coastal area encounters one of these bulges every 12 hours—*these are the high tides!* Well, as illustrated in [30.2], it's actually much more complicated than that, because the Moon exerts exactly the same kind of gravitational distortion on our oceans, and, despite its minuscule mass compared to the Sun, its proximity to Earth induces tidal forces that are roughly *twice*[5] as great as the Sun's! The actual tides are the complex[6] result of the superposition of these twin tidal forces of the Sun and Moon; their interaction over one lunar orbit is illustrated and explained by [30.2]. This was first understood by Newton, and was another triumph of the *Principia* (Newton, 1687).

The geometry is simplest when we see a full moon (top left of [30.2]), or a new moon (bottom left of [30.2]), for this means that the Moon, Earth, and Sun are lined up, in which case the tidal bulges induced by the Moon and Sun on Earth's oceans are lined up, too, reinforcing each other, producing the greatest high tides, and also the lowest low tides; these are called ***Spring tides***. (NOTE: These operate year-round, and have nothing to do with the season we call "Spring.")

[5] See Exercise 13.

[6] See Schutz (2003, §5).



[30.2] Explanation of the Ocean Tides. *The tidal forces of the Moon (dashed line) and of the Sun (dotted oval) combine to distort the oceans (shaded oval areas), creating the tides. (NOTE: Nothing is to scale! The oceans and the tides are massively exaggerated, and the distance to the Moon is massively shrunken!) Top left, the illuminated half of the Moon faces the Earth, the reflected sunlight being seen by us as a Full Moon. Since the tidal forces of the Sun and Moon are now aligned, they reinforce each other to produce the greatest high tides, and the lowest low tides; these are the* Spring tides. *Top right, the more powerful lunar tidal forces are partially cancelled out by the orthogonal solar tidal forces, resulting in the smallest difference between high and low tides; these are the* Neap tides.

When the Moon and Sun form a right angle at the Earth (top right and bottom right of [30.2]), they produce bulges in orthogonal directions, so the smaller solar bulge partially cancels out the larger lunar bulge, resulting in the smallest difference between high and low tide; these are called ***Neap tides***.

Since—by definition!—the Moon executes one orbit of Earth each month, it is lined up with the Earth and the Sun *twice* per month: the New Moon corresponds to the Moon lying between the Sun and the Earth, while the Full Moon corresponds to the Earth lying between the Sun and the Moon. Thus the Spring tides occur every two weeks, and the same is true of the Neap tides, the alternation between the two being weekly—the time it takes for the Moon to execute one-quarter of an orbit around the Earth, as shown in [30.2].

To learn much more about the physics of gravity (and the tides in particular) we strongly recommend *Gravity From The Ground Up* (Schutz 2003).

30.3 Newton's Gravitational Law in Geometrical Form

(NOTE: The following exposition of Einstein's geometrical theory of gravity is inspired by Penrose's beautiful essay, *The Geometry of the Universe* (Penrose 1978) and also by the three relevant chapters of his remarkable work, *The Road to Reality* (Penrose 2005, §§17–19). However, we shall explicitly prove several crucial results that Penrose states but leaves as exercises for the reader, and we shall also include some additional geometrical observations that Penrose does not mention.)

The tidal forces we have just described would be *qualitatively* the same for *any* conceivable gravitational law of attraction that decreased with distance. But the *actual* law of gravitation, discovered by Newton and published in his *Principia* of 1687, *decreases as the square of the distance.* As we will now show, this law, and this law *alone*, results in tidal forces that have a very specific and very beautiful geometrical signature.

Let r be your distance from the centre of the Earth, let ξ be the radius of the horizontal, "equatorial" circle of the sphere of particles that surround you, and let $\delta\varphi$ be the small angle subtended by ξ at the centre of the Earth. Then $\xi \asymp r\,\delta\varphi$. Since $\delta\varphi$ remains fixed as you and the equatorial particles fall radially towards the centre of the Earth, the inward acceleration of this circle towards you is given by

$$\ddot{\xi} \asymp \ddot{r}\,\delta\varphi = -\frac{GM}{r^2}\,\delta\varphi = -\frac{GM}{r^3}\,r\,\delta\varphi \asymp -\frac{GM}{r^3}\,\xi,$$

by virtue of (30.1).

If the previous two chapters have done their job, your immediate, Pavlovian reaction to the mere sight[7] of this formula is to exclaim, *"That's a Jacobi equation!"* We shall make the connection explicit before long, but this attraction of horizontally separated particles is indeed due to *positive sectional curvature*, $\mathcal{K}_+$, where

$$\boxed{\ddot{\xi} = -\mathcal{K}_+\,\xi \quad \Longrightarrow \quad \mathcal{K}_+ = +\frac{GM}{r^3}.} \tag{30.2}$$

Let Ξ denote the distance from you to the particles above your head and below your feet—we chose a Greek letter that depicts this! As illustrated in [30.1], and as previously discussed, the relative acceleration of these particles *away* from you is given by the difference $\delta\ddot{r}$ in the acceleration $\ddot{r}$—given by (30.1)—at your height r and at height $r+\delta r = r + \Xi$. Thus,

$$\ddot{\Xi} = \delta\ddot{r} \asymp [\partial_r\,\ddot{r}]\,\delta r = \partial_r\left[-\frac{GM}{r^2}\right]\Xi = \left[+\frac{2GM}{r^3}\right]\Xi.$$

This repulsion of vertically separated particles is due to *negative sectional curvature*, $\mathcal{K}_-$, where

$$\boxed{\ddot{\Xi} = -\mathcal{K}_-\,\Xi \quad \Longrightarrow \quad \mathcal{K}_- = -\frac{2GM}{r^3}.} \tag{30.3}$$

[7] See (28.2) and (29.21).

Thus *the vertical repulsion is exactly double the horizontal attraction*, as illustrated in [30.1]. We are now poised to recognize the true geometrical significance of this fact—the geometrical *signature* of Newton's inverse-square tidal forces. The key is to follow the evolution of the *volume* $\delta\mathcal{V}$ of the sphere as it falls to Earth, the tidal forces deforming it into an ellipsoidal egg.

Imagine that the initial sphere of particles is enclosed in a cube. As the sphere falls and deforms into an egg, the enclosing cube and its contents undergo a linear transformation that deforms it into a box with edges of lengths 2ξ, 2ξ, and 2Ξ, and hence of volume $8\xi^2\,\Xi$. The volume occupied by the egg is a fixed proportion of this box, so

$$\boxed{\delta\mathcal{V} = \tfrac{4\pi}{3}\,\xi^2\,\Xi.}$$

Since you and the sphere of particles start from rest, clearly the rate of change of this volume must initially vanish: $\delta\dot{\mathcal{V}} = 0$. This can be confirmed by calculation:

$$\delta\dot{\mathcal{V}} = \frac{4\pi}{3}\left[2\xi\dot{\xi}\,\Xi + \xi^2\,\dot{\Xi}\right] = 0,$$

since $\dot{\xi} = 0 = \dot{\Xi}$.

But the tidal forces do instantly *accelerate* the particles of the sphere, so let us now calculate the *acceleration* of the volume of the egg:

$$\delta\ddot{\mathcal{V}} = \frac{4\pi}{3}\left[2(\dot{\xi})^2\,\Xi + 2\xi\ddot{\xi}\,\Xi + 2\xi\dot{\xi}\dot{\Xi} + 2\xi\dot{\xi}\dot{\Xi} + \xi^2\,\ddot{\Xi}\right] = \frac{4\pi}{3}\left[2\xi\ddot{\xi}\,\Xi + \xi^2\,\ddot{\Xi}\right].$$

But, initially, $\xi = \delta r = \Xi$, and we know that the horizontal attraction and vertical repulsion are governed by (30.2) and (30.3), respectively, so,

$$\delta\ddot{\mathcal{V}} = \frac{4\pi}{3}\left[2\ddot{\xi} + \ddot{\Xi}\right](\delta r)^2 = -\frac{4\pi}{3}\,(2\mathcal{K}_+ + \mathcal{K}_-)\,(\delta r)^3 = 0\,!$$

We have arrived at the beautiful

Geometrical Signature of the Inverse-Square Tidal Forces:
The tidal forces generated by the Inverse-Square Law—and only *the Inverse-Square Law—are* **volume-preserving**, *in the precise sense that the acceleration of the volume vanishes, so that the volume remains constant to order* t^2*:*

$$\delta\ddot{\mathcal{V}} = 0. \tag{30.4}$$

We have not yet proved the "only" part of this proposition. To do so, suppose that $\ddot{r} = f(r)$, where f is an unknown function. Substituting this in place of (30.1) in the above analysis, we find [exercise],

$$\delta\ddot{\mathcal{V}} = 0 \quad\Longrightarrow\quad \frac{df}{dr} + \frac{2f}{r} = 0 \quad\Longrightarrow\quad f(r) \propto \frac{1}{r^2},$$

as was to be shown.

In the foregoing analysis we have, understandably, neglected *your* microscopic gravitational field and its influence on the particles that surround you. But suppose we remove you from the interior of the sphere, and instead fill it with extremely dense matter of density ρ. Then the sphere of particles of radius $\xi = \Xi$ will be accelerated inward towards the centre, just as if there were a point mass $\rho\,\delta\mathcal{V}$ at the centre, pulling them in with acceleration,

$$\ddot{\xi} = -\frac{G\rho\,\delta\mathcal{V}}{\xi^2}.$$

Now let us compute the acceleration of the volume in this new circumstance. Since now $\delta\mathcal{V} = \frac{4\pi}{3}\xi^3$,

$$\delta\dot{\mathcal{V}} = 4\pi\xi^2\,\dot{\xi} \qquad \Longrightarrow \qquad \delta\ddot{\mathcal{V}} = 8\pi\xi(\dot{\xi})^2 + 4\pi\xi^2\,\ddot{\xi} = 8\pi\xi(0)^2 + 4\pi\xi^2\left(-\frac{G\rho\,\delta\mathcal{V}}{\xi^2}\right).$$

Thus we have discovered this

> **Geometrical Signature of the Inverse-Square Attraction:**
> *Consider a sphere of volume $\delta\mathcal{V}$ that is filled with matter of density ρ, and suppose that just above its surface are tiny test particles, released from rest. Instantly they begin to accelerate towards the centre, and the inverse-square law causes the volume they enclose to implode with an acceleration governed by this geometrical law:*
> $$\delta\ddot{\mathcal{V}} = -4\pi G\,\rho\,\delta\mathcal{V}. \tag{30.5}$$

If we now imagine this material sphere (and its skin of test particles) to be launched with arbitrary velocity in the Earth's gravitational field, then we obtain a superposition of these two effects: the Earth's tidal forces start to deform the sphere into an egg of equal volume, but the attraction of the matter within the sphere also exerts its volume-reducing effect. The net result is that the initial sphere evolves into an egg, but with its volume shrinking in proportion to the mass within, and in proportion to the square of the time.

30.4 The Spacetime Metric

Soon we will be able to recast the foregoing analysis in terms of Einstein's 4-dimensional spacetime, and we will then be led very naturally to Einstein's Gravitational Field Equation. But before we can do that, we must first discuss the metric structure of spacetime, and we must also learn how to draw *pictures* of spacetime.

The first thing to understand is that the local tangent-space structure is *not* $\mathbb{R}^4$ with the standard Euclidean metric; it is instead *Minkowski* spacetime with the ***Minkowski metric*** given by the spacetime interval we introduced in (6.15), on page 75. (We remind the reader that we have chosen units such that light travels one unit of distance in one unit of time, so that its speed is $c = 1$. In these units, Einstein's famous equation, $E = mc^2$, becomes, simply, $E = m$, i.e., "energy is mass"!)

Thus,

$$ds^2 = dt^2 - (dx^2 + dy^2 + dz^2) = \left[dx^0\right]^2 - \left(\left[dx^1\right]^2 + \left[dx^2\right]^2 + \left[dx^3\right]^2\right). \tag{30.6}$$

This is the analogue of distance within the tangent plane to a surface at a point, and just as the tangent plane is flat, so too is Minkowski spacetime—its Riemann tensor vanishes identically.

The analogue of the true distance between neighbouring points of the surface itself is given by the ***metric tensor*** of the curved spacetime, which is usually denoted **g**—it takes two vectors as input, and outputs a scalar. It generalized the ***dot product*** (or ***scalar product***), and it is symmetrical:

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) \equiv \mathbf{u} \bullet \mathbf{v} = \mathbf{v} \bullet \mathbf{u} = \mathbf{g}(\mathbf{v}, \mathbf{u}).$$

Just as in a 2-surface, the metric constitutes the most *fundamental* information about spacetime: it defines *distance*, and once we know distance, we know *everything*: geodesics, parallel transport, and Riemann curvature. If ϵ is a small (ultimately vanishing) connecting vector between two neighbouring events in spacetime, then the metric tells us the Einsteinian distance between them:

$$ds^2 = \mathbf{g}(\epsilon, \epsilon)$$

But recall that $ds^2 > 0$ only when ϵ connects two points on the spacetime trajectory (aka ***world-line***) of a material particle with nonzero mass, which is constrained to travel at less than the speed of light; in this case we say that the separation is ***timelike***. As we stated in (6.16) on page 76, in this case ds simply measures wristwatch time elapsed for the traveller of the world-line. If ϵ is instead along a light ray (the world-line of a photon) then $ds^2 = 0$; in this case we say the separation is ***null***. Finally, if $ds^2 < 0$ we say that the separation is ***spacelike***.

If $\{\mathbf{e}_i\}$ is a set of four basis vectors—called a ***tetrad***—which need *not* be orthonormal, then we can obtain the *components* of the metric tensor in exactly the same way as we obtained the components of the Ricci tensor—simply apply the tensor to pairs of basis vectors:

$$g_{ij} \equiv \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{g}(\mathbf{e}_j, \mathbf{e}_i) = g_{ji}.$$

Thus,

$$\boxed{ds^2 = \mathbf{g}(\epsilon, \epsilon) = \mathbf{g}(dx^i\, \mathbf{e}_i,\, dx^j\, \mathbf{e}_j) = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j)\, dx^i\, dx^j = g_{ij}\, dx^i\, dx^j.} \tag{30.7}$$

For example, in Minkowski spacetime, as expressed in (30.6), $g_{00} = +1$, while $g_{11} = g_{22} = g_{33} = -1$, and $g_{ij} = 0$ if $i \neq j$. In general, however, these g_{ij} are *functions* in spacetime, and $g_{ij} \neq 0$ if $i \neq j$. It is important to realize that, even in flat Minkowski spacetime, these components can look drastically different, depending on the choice of coordinates. For example, if we use standard spherical polar coordinates for the spatial coordinates, then

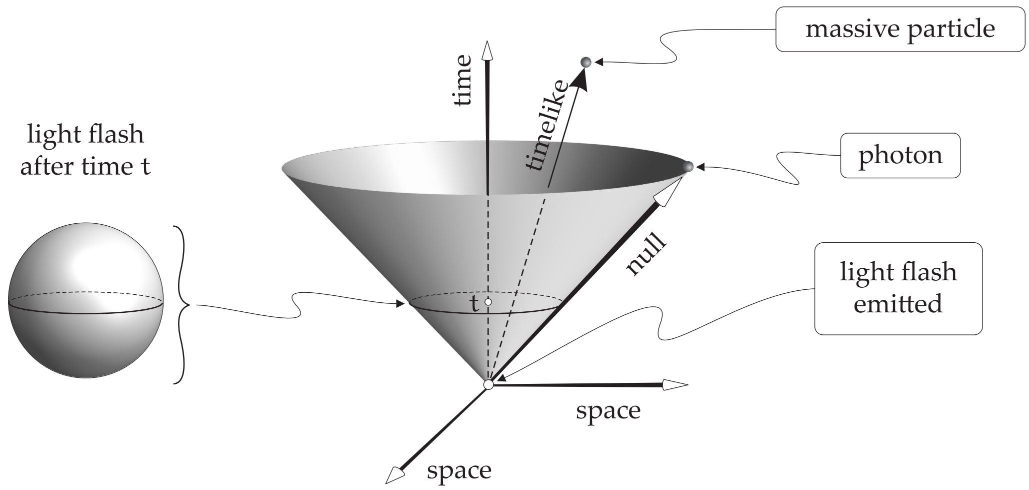
$$ds^2 = dt^2 - dr^2 - r^2(d\varphi^2 + \sin^2\varphi\, d\theta^2), \tag{30.8}$$

so, $g_{tt} = 1$, $g_{rr} = -1$, $g_{\varphi\varphi} = -r^2$, and $g_{\theta\theta} = -r^2 \sin^2\varphi$, and yet these functions describe exactly the same, flat geometry as before.

NOTE ON NOTATION: This g_{ij}-notation is *universally* accepted: it can be found in *all* modern texts, whether they be on mathematics or on physics. So we too shall employ this notation, which is the n-dimensional generalization of Gauss's original (E, F, G)-notation for a 2-surface; see (4.8), page 37. On the other hand, at least in the 2-dimensional case, we hope we have convinced you—several times over!—of the conceptual and computational advantages of our *alternative* (A^2, B^2)-notation. That said, into n dimensions our notation dare not tread. . . .

30.5 Spacetime Diagrams

The vast majority of humans (including the author) cannot directly visualize 4-dimensional spacetime. How, then, can we draw diagrams that will enable us to reason geometrically about spacetime?! The most common and useful answer to this conundrum is to simply *suppress one of the three spatial directions*. More specifically, we shall draw time going *up* the page, and we shall represent two (out of the three) spatial directions as forming a horizontal plane, perpendicular to the vertical time direction.



[30.3] The Null Cone (Light Cone) in Spacetime. *Time is represented by the vertical direction, and two (out of three) spatial directions are represented by the orthogonal, horizontal directions. A flash of light is emitted at an event in spacetime, and the expanding sphere of light is represented at each time* t *by its circular cross section. The entire future of the flash is therefore represented by a cone, the world-lines of the photons being the null generators of the cone. Material particles are constrained to travel at less than the speed of light, along timelike world-lines,* inside *the cone.*

As drastic as this loss of information may seem, symmetry often comes to our rescue: if two of the spatial directions are physically equivalent, then discarding one of them costs us nothing. For example, and of great upcoming importance, in [30.1] all horizontal directions—in the plane parallel to the Earth's surface—are indistinguishable, so if we picture this plane spanned by two orthonormal vectors, then we can safely suppress one of them, with no loss of information. In stark contrast to this, the vertical direction is the direction of the gravitational field, and it is physically quite distinct from the horizontal directions, so in a spacetime diagram we must not throw *that* direction away!

But let us begin with an even more symmetrical situation: empty, flat, Minkowski spacetime. Here *all* spatial directions are physically indistinguishable from one another. One of the most useful objects to depict in a spacetime diagram is a flash of light emitted from a particular point at a particular time—an ***event***. This results in a sphere of light that expands one unit of distance for every unit of time. Since Einstein tells us that material particles cannot travel faster than light, any massive particles ejected from the same event as the light flash must remain *inside* this expanding sphere. How does all this look in a spacetime diagram?

Suppressing one spatial dimension, the expanding sphere of light is now represented by the expanding *circle* shown in [30.3], which generates a cone. For this reason, the spacetime depiction of the light flash is called the ***light cone*** or ***null cone***. The world-lines of material particles stay inside this cone, as illustrated. The fundamental importance of the null cone is that its interior therefore represents the set of events that can be influenced by the original event—it tells us about the so-called ***causal structure*** of spacetime, i.e., can this event, here and now, cause something to happen at that event, there and then?

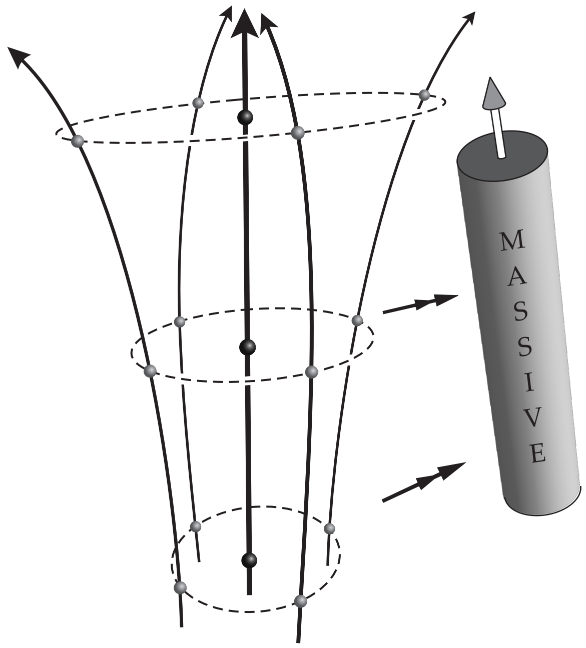
The tangent vector to the world-line of a particle is called the ***4-velocity*** of the particle. A particle at rest has a *nonzero* 4-velocity: it points straight up the time axis! Note that for a massive particle it is possible to normalize the 4-velocity in the same way that we have usually insisted that our particles travel over a 2-surface at unit speed, but this is *not* possible for the 4-velocity of a photon, because the "length" of a photon's 4-velocity is always *zero*.

30.6 Einstein's Vacuum Field Equation in Geometrical Form

In order to "derive"[8] Einstein's Vacuum Field Equation, let us return to the vacuum of space above the Earth's surface, and to the Newtonian inverse-square tidal forces that operate there, as depicted in [30.1]. But now let us analyze this from the perspective of *spacetime*.

We have already noted in [30.1] that all horizontal directions are indistinguishable, so there is no loss of information if we draw a spacetime diagram that only retains one such direction. Therefore, consider any vertical great circle of the sphere, spanned by one such horizontal direction, and the vertical direction of Earth's gravity. As the sphere of particles falls, the tidal forces distort the sphere into an ellipsoidal egg, but this evolution is *completely and faithfully captured* by looking at the evolution of merely this single vertical, great-circle cross section of the original sphere. If we follow just these particles as they fall, the tidal forces will distort their circle in to an ellipse, but then the *full*, physical ellipsoid can be recovered simply by rotating the ellipse about its vertical axis of symmetry.

Imagine, once again, that you are the astronaut in the centre of this sphere. Let us draw a spacetime diagram from *your* perspective. Having turned off your jetpack, you are in free-fall, as are the particles surrounding you. Therefore, so long as you don't look down at the Earth, you feel that you are simply *floating, motionless*, and the sphere of particles centred on you is likewise almost motionless . . . except that you gradually witness the tidal evolution of the sphere into an egg. If you *do* look down at the Earth, you see it moving towards you, but you are still at liberty to consider yourself motionless—surely, it is the *Earth* that is moving towards *you!*



[30.4] Spacetime Depiction of the Tidal Distortion of a Free-Falling Sphere into an Egg. *As usual, time runs vertically.*

It is from this perspective that we have drawn [30.4], with you being the darker world-line in the centre. It depicts the spacetime evolution of the sphere (represented by a vertical circle) into an ellipsoidal egg, as it falls to the Earth. Here the Earth is schematically represented by the object labelled MASSIVE.

This diagram should ring a bell! It looks almost exactly like [29.11c], depicting the evolution of a circular bundle of geodesics in the presence of sectional curvature that is positive in one direction (causing attraction), and negative in the perpendicular direction (causing repulsion). This analogy becomes precise if we consider the sectional curvature in spacetime, and compare (30.2) and (30.3) with the (now 4-dimensional) Sectional Jacobi Equation, (29.21).

The 4-velocity **v** of the circle of particles (starting from rest) is vertical, along the time axis. As the vertical spacetime plane containing this 4-velocity rotates around it, it passes through the direction parallel to the Earth's surface, with the positive, attractive sectional curvature (30.2):

---

[8]In our present state of understanding, the physical laws are irreducible: they cannot be logically derived from anything more primitive. However, from Newton onward, consistency with other physical laws, combined with mathematical *beauty*, have served as remarkably effective guiding principles that have allowed us to correctly *guess* some of Nature's deepest secrets.

$$\mathcal{K}_{+} = +(GM/r^3).$$

Rotating through another right angle, the spacetime plane contains the direction of the Earth's gravitational field, yielding negative, repulsive sectional curvature (30.3):

$$\mathcal{K}_{-} = -(2GM/r^3).$$

But, of course, there is a *third* spatial direction that is orthogonal to these two directions, but it is suppressed in our spacetime diagram—it is the third direction orthogonal to the time axis, lying in the plane parallel to the Earth's surface.

Let us spell this out in terms of an orthonormal tetrad adapted to this situation. Let $\mathbf{e}_0 = \mathbf{v}$ point along the (vertical) time axis, let $\mathbf{e}_1$ and $\mathbf{e}_2$ span the plane parallel to the Earth's surface, and let $\mathbf{e}_3$ point radially away from the Earth, in the direction of the gravitational field.

The generalization of (29.26) from 3-manifolds to 4-manifolds is that the average $\mathcal{K}_{\text{mean}}$ of the sectional curvatures in spacetime over *all* planes containing $\mathbf{e}_0 = \mathbf{v}$ can be obtained by averaging over just these three orthogonal planes containing $\mathbf{e}_0 = \mathbf{v}$:

$$\Pi_1 \equiv \Pi(\mathbf{e}_0, \mathbf{e}_1), \qquad \Pi_2 \equiv \Pi(\mathbf{e}_0, \mathbf{e}_2), \qquad \text{and} \qquad \Pi_3 \equiv \Pi(\mathbf{e}_0, \mathbf{e}_3), \qquad .$$

So,

$$\mathcal{K}_{\text{mean}} = \frac{\mathcal{K}(\Pi_1) + \mathcal{K}(\Pi_2) + \mathcal{K}(\Pi_3)}{3} = \frac{1}{3}\mathbf{Ricci}(\mathbf{e}_0, \mathbf{e}_0) = \frac{1}{3}\, R_{00}.$$

Very remarkably, when viewed in this geometrical way, *the passage from Newton's inverse-square tidal forces to Einstein's curved spacetime is completely seamless!*

We saw in (30.4) that the signature of Newton's inverse-square tidal forces is that it is *volume-preserving*, and we saw in (29.29) that the acceleration of the volume is governed by the *Ricci curvature*. Combining these two results, we find that

$$\boxed{\mathbf{Ricci}(\mathbf{v}, \mathbf{v})\, \delta\mathcal{V} = -\ddot{\delta\mathcal{V}} = 0.} \tag{30.9}$$

More explicity, from (30.2) and (30.3),

$$R_{00} = \mathcal{K}(\Pi_1) + \mathcal{K}(\Pi_2) + \mathcal{K}(\Pi_3) = \mathcal{K}_{+} + \mathcal{K}_{+} + \mathcal{K}_{-} = \frac{GM}{r^3} + \frac{GM}{r^3} - \frac{2GM}{r^3},$$

so

$$\boxed{R_{00} = 0.} \tag{30.10}$$

Here we have imagined that the sphere of particles starts from rest, so that $\mathbf{v} = \mathbf{e}_0$ is a purely vertical 4-velocity, along the time axis. But we know that the volume-preserving signature of Newton's tidal forces continues to hold if we launch the sphere with any velocity, and therefore (30.9) holds for *any* timelike 4-velocity $\mathbf{v}$.

The final step in the derivation of Einstein's Vacuum Field Equation depends on the *symmetry*, (29.28), of $\mathbf{Ricci}$:

$$\mathbf{Ricci}(\mathbf{x}, \mathbf{y}) = \mathbf{Ricci}(\mathbf{y}, \mathbf{x}).$$

Now let $\mathbf{v} = \mathbf{x} + \mathbf{y}$, for *arbitrary* (timelike) $\mathbf{x}$ and $\mathbf{y}$. Then,

$$\begin{aligned} 0 &= \mathbf{R}\mathrm{icci}(\mathbf{v}, \mathbf{v}) \\ &= \mathbf{R}\mathrm{icci}([\mathbf{x}+\mathbf{y}]\,,\,[\mathbf{x}+\mathbf{y}]) \\ &= \mathbf{R}\mathrm{icci}(\mathbf{x}, \mathbf{x}) + \mathbf{R}\mathrm{icci}(\mathbf{x}, \mathbf{y}) + \mathbf{R}\mathrm{icci}(\mathbf{y}, \mathbf{x}) + \mathbf{R}\mathrm{icci}(\mathbf{y}, \mathbf{y}) \\ &= 0 + \mathbf{R}\mathrm{icci}(\mathbf{x}, \mathbf{y}) + \mathbf{R}\mathrm{icci}(\mathbf{y}, \mathbf{x}) + 0 \\ &= 2\,\mathbf{R}\mathrm{icci}(\mathbf{x}, \mathbf{y}). \end{aligned}$$

We have thus arrived at the

$$\boxed{\textit{\textbf{Einstein Vacuum Field Equation:}}\qquad \mathbf{R}\mathrm{icci} = \mathbf{0} \quad \Longleftrightarrow \quad R_{ik} = 0.} \tag{30.11}$$

Let us not forget where this came from and what it means:

In vacuum, in order for the tidal forces to preserve volume, positive sectional curvatures must be perfectly balanced by negative sectional curvatures, so as to exactly cancel each other out on average.

While the Ricci curvature vanishes identically in vacuum, this is only the volume-reducing *average* of the positive and negative sectional curvatures. The Riemann tensor itself certainly does not vanish in general; witness the positive and negative tidal curvatures in (30.2) and (30.3).

In general, it is possible to split the Riemann curvature into a volume-reducing Ricci part, plus a purely tidal, volume-preserving part, called the ***Weyl curvature***; see Exercise 15. This split is most elegantly and naturally accomplished via Penrose's 2-spinor formalism; see Penrose and Rindler (1984, §4.6) or Wald (1984, §13.2).

Lest you doubt that (30.11) is in fact Einstein's equation, [30.5] shows Einstein himself proudly writing it down, and while he has modestly added a question mark, his equation now encompasses a remarkable array of startling, testable predictions, *all of which have been confirmed experimentally*, often with extraordinary precision.

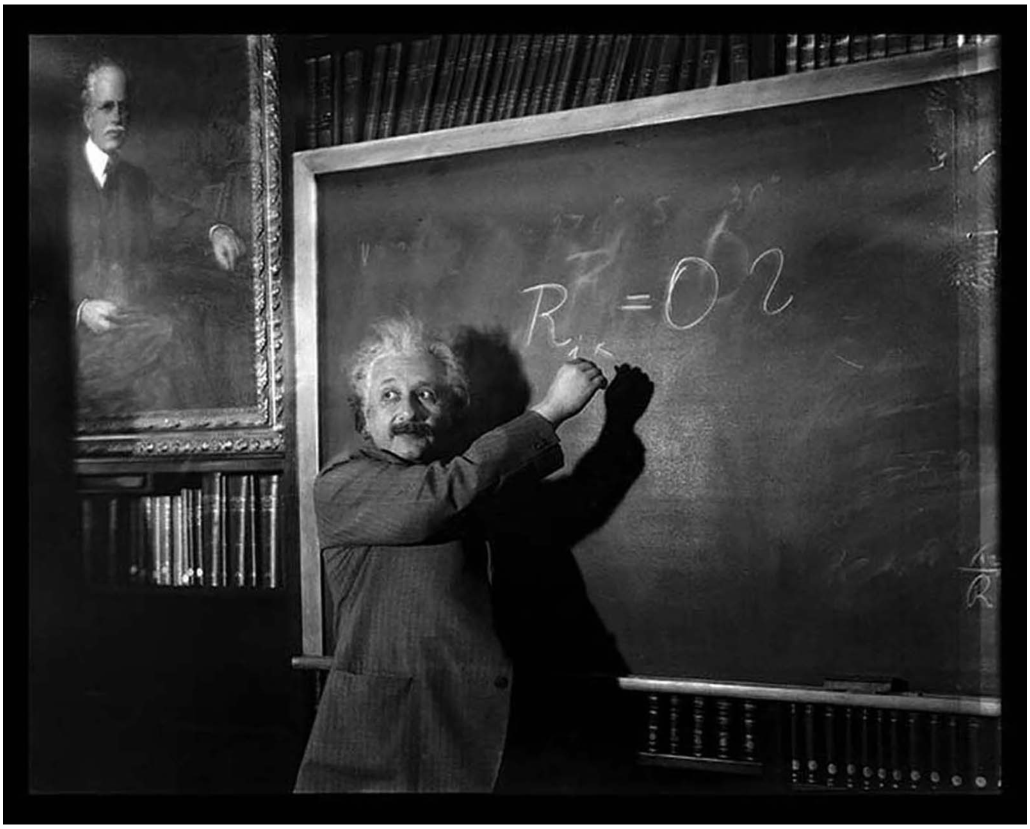
We shall outline several of these experimental triumphs, but their mathematical derivations must be left to the technical references we will provide, as well as to those in the General Relativity section of *Further Reading*, at the end of this book.

30.7 The Schwarzschild Solution and the First Tests of the Theory

When we speak of a *solution* to Einstein's equation, we mean a *geometry* of spacetime (defined by its *metric*) that satisfies the equation.

The single most important such solution (at least historically) describes the spacetime in the vacuum region outside a spherically symmetric (nonspinning) mass, M. This solution, the *first* exact solution ever found, was discovered by Karl Schwarzschild[9] (1873–1916)—shown in [30.6]—almost *instantly* after Einstein announced his theory. Schwarzschild discovered it while on active duty on the Russian front, as a 42-year-old artillery leftenant in the German army during World War I. He sent the solution to Einstein in a letter dated the 22nd of December, 1915:

[9]Schwarzschild was a child prodigy, publishing papers on celestial mechanics when he was 16 years old. Later, from 1901 to 1909, he was a professor at Göttingen, where his colleagues included Hilbert, Klein, and Minkowski.



[30.5] *Albert Einstein (1879–1955) circa 1931, delivering a talk at the Mount Wilson Observatory's Hale Library, Pasadena, California. On the blackboard, Einstein has just written his vacuum field equation:* $R_{ik} = 0$ (plus a question mark). *Image courtesy of the Observatories of the Carnegie Institution for Science Collection at the Huntington Library, San Marino, California.*

> As you see, the war treated me kindly enough, in spite of the heavy gunfire, to allow me to get away from it all and take this walk in the land of your ideas.

Decades later, physicists would slowly come to realize that Schwarzschild's solution not only describes the geometry surrounding a spherical mass, such as the Sun or the Earth, it *also* describes the pure vacuum gravitational field of a black hole!

Schwarzschild's solution builds on the spherical polar metric formula of Minkowski spacetime, (30.8):

Schwarzschild Solution

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \frac{dr^2}{\left(1 - \frac{2GM}{r}\right)} - r^2(d\varphi^2 + \sin^2\varphi\, d\theta^2). \tag{30.12}$$

NOTES:

- Here the radial coordinate r is defined in such a way as to maintain its Euclidean relation to the area $\mathcal{A}(r)$ of a sphere of radius r, namely, $r = \sqrt{\mathcal{A}(r)/4\pi}$. Thus dr does *not* measure radial distance; radial distance is instead equal to $dr/\sqrt{\left(1 - \frac{2GM}{r}\right)}$. Only as we recede from the massive body (or black hole) does the radial coordinate asymptotically regain its Euclidean interpretation as radial distance.
- If we do not choose units in which the speed of light $c = 1$, then $(2GM/r)$ must be replaced by $(2GM/c^2 r)$.
- At the conclusion of our drama, we will use ***curvature 2-forms*** to confirm that this geometry does indeed satisfy the Einstein Vacuum Equation, (30.11).

[30.6] *Karl Schwarzschild (1873–1916).*

- The g_{rr} metric component blows up when

$$r = r_s \equiv \textbf{\textit{Schwarzschild radius}} \equiv \frac{2GM}{c^2}. \tag{30.13}$$

  Einstein and others were initially confused by this, and thought this meant that there was something singular about spacetime itself when $r = r_s$. However, this ultimately turned out to be a nonphysical artifact of the particular *choice of coordinates*. If an object of length l is located at the Schwarzschild radius, it experiences ordinary, nonsingular tidal forces of order $(GM/r_s^3)\, l$, in other words, of order $(c^6 l/G^2)\,(1/M^2)$.
- If the mass M is extremely *large*, then the tidal forces at the Schwarzschild radius are extremely *small*.
- If the spherical star or planet has radius R, then the solution only applies the vacuum region *outside* it, i.e., $r > R$. In the interior, $r < R$, a quite different metric applies, which satisfies the *full* Einstein Field Equation (with matter), which we shall derive shortly. In the case of a sphere of uniform density, the same Karl Schwarzschild found the exact solution to this nonvacuum equation, almost immediately after finding the vacuum solution; it is called, logically enough, the ***Interior Schwarzschild Solution***.
- For the Sun, which has a radius of 430,000 miles, the Schwarzschild radius is only 2 miles! The Schwarzschild radius of the Earth is only about a third of an inch! Thus we must switch from the vacuum Schwarzschild solution to a different interior solution long before we get anywhere near the Schwarzschild radius.

Although the Earth and the Sun *do* spin, Schwarzschild's solution is nevertheless an excellent approximation[10] to the spacetime geometry of both. Indeed this solution suffices to give a complete analysis of our solar system, and to make three crucial predictions about phenomena in our Solar System that *deviate* from Newton's theory.

In the interest of historical accuracy, we must point out that all three of these initial predictions were made by Einstein *before* he knew Schwarzschild's exact solution. Einstein did so by means of an *approximate* solution to his equation, which—bizarrely and inelegantly—he obtained using *rectangular* coordinates!

Here were Einstein's three initial predictions/tests of his theory:

1. When light travels upward through a gravitational field, its frequency should be reduced—we say it is ***redshifted***. Einstein made this prediction in 1907, long before had discovered the field equation, and thus it is not a direct test of that equation. Nevertheless, had experiments *denied* this prediction, the entire framework of Einstein's theory would have been invalidated. Sadly, the first definitive, terrestrial confirmation of this prediction was not carried out by Pound and Rebka until 1959, four years after Einstein's death.
2. The major axis of the elliptical orbit of Mercury rotates a minuscule amount with each orbit. Newtonian theory predicted that the rotation per *century* should be 532 arcseconds—less than one degree. However, Urbain Le Verrier and later Simon Newcomb conducted extremely accurate analyses of observations collected since 1697, and by 1882 they had established that the *actual* rotation of the orbit was 575 arcseconds per century—a discrepancy of 43 arcseconds per century: a fantastically small (but also indelible and mysterious) blemish on Newton's theory.

   When Einstein calculated the correction according to his own law of gravitation, it made a definite, unambiguous prediction—whatever the verdict, there could be no appeal.

   Einstein's formula yielded 43 arcseconds per century! In that moment, Einstein realized that Nature had spoken to him—his law was built into the eternal structure of the world. Mercury's orbit had been rotating at this rate 4 billion years before mankind existed, and it would continue to do so billions of years into the future.

   Einstein told one friend that, in that moment, he had experienced physical heart palpitations; he told another friend that he felt that something actually *"snapped"* inside him. See Pais (1982, p. 253).
3. Einstein also calculated that light passing close to the rim of the Sun should be bent by 1.75 arcseconds. But starlight grazing the Sun can only be seen during a total solar eclipse, so the prediction could not be tested until one occurred, which happened on the 29th of May, 1919. In anticipation of the eclipse, Sir Arthur Eddington organized expeditions to Brazil and west Africa, in order to take the critical photographs during the brief eclipse, and to perform the measurement of the deflection, indeed, to see if there *were* any deflection.

   The dramatic *confirmation* of Einstein's prediction of the bending of light, and the precise amount by which it was bent, was an international sensation, splashed across the front pages of newspapers around the world. Overnight, the unknown German scientist became a household name, and "Einstein" became a synonym for genius.

After these early triumphs, there followed several fallow decades, in which physicists shifted their attention to all things quantum. However, beginning in the 1960s, General Relativity underwent something of a renaissance, which has continued to the present day, attracting the

[10]Rapidly rotating stars and black holes must be described by a different, vitally important solution, discovered in 1963 by the New Zealand mathematician, Roy Kerr (born 1934); accordingly, it is called the ***Kerr solution***. For Kerr's personal account of his remarkable discovery, see Kerr (2008) and Wiltshire et al. (2009).

finest theoreticians and experimentalists, with many new theoretical predictions, and many new experimental confirmations.

Some of these newer tests still depend on the simple Schwarzschild solution of 1915. Our successful daily use of GPS navigation—see Taylor and Wheeler (2000, §A-1)—is one of these newer confirmations of both Einstein's theory and Schwarzschild's solution!

Tragically, Schwarzschild himself did not live to see any of these triumphs: he died of a rare autoimmune disease on the 11th of May, 1916, mere months after discovering his two remarkable solutions of Einstein's equation. In remembering him, Sir Arthur Eddington wrote, "... his joy was to range unrestricted over the pastures of knowledge, and, like a guerrilla leader, his attacks fell where they were least expected."

30.8 Gravitational Waves

Other confirmations of Einstein's equation have arisen from completely different kinds of solutions to the vacuum field equation, (30.11). As we first outlined on page 231, the existence of gravitational waves was predicted by Einstein in 1916, and the very first experimental detection of a gravitational wave occurred almost exactly a century later, on the 14th of September, 2015.

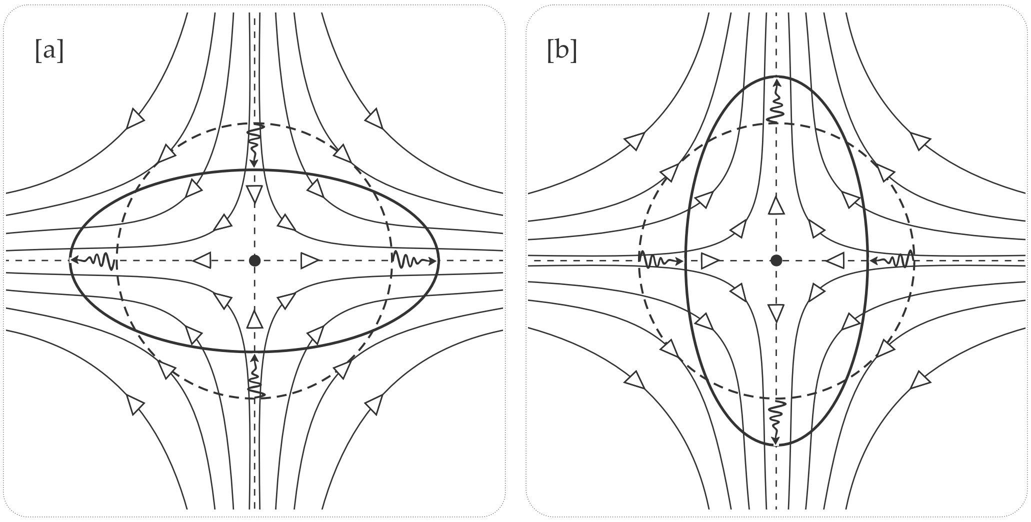
We are accustomed to the idea that electromagnetic waves—such as the light by means of which we see the world—are photons travelling *through* space. Gravitational waves are something astonishingly different: they are ripples of curvature in the fabric *of* spacetime itself! But not just any old ripples are possible: like the Schwarzschild solution, they must satisfy the Einstein vacuum equation. They are oscillating waves of pure tidal force that have broken free from the violent events that created them, travelling out across space (*as* space!) at the speed of light.

As a gravitational wave passes through a sphere of particles at a fixed location, it has no effect on the particles whose separation lies in the travel direction of the wave. But in the plane *orthogonal* to direction of the wave, it causes *oscillating*, tidal, egg deformation, stretching the sphere in one direction, and compressing it in the orthogonal direction. See [30.7a], which shows the ***field lines*** (i.e., streamlines of the tidal force field) at one particular moment.

The tidal force field of the gravitational wave is *different* in character from the tidal force field in the vacuum above the Earth, in two important respects:

- First, whereas the gravitational wave only compresses in *one* direction orthogonal to the stretching direction, the Earth's tidal force field compresses in *both* orthogonal directions.
- Second, the Einstein vacuum equation tells us that the positive curvatures in any two orthogonal directions lying within the plane parallel to the Earth's surface must exactly *cancel* the negative curvature in the radial direction of the Earth's gravitational field. Therefore these two positive curvatures must each be *half* as strong as the radial negative curvature, as illustrated in [30.1]. But in the case of the gravitational wave, the curvature associated with the direction of travel *vanishes*. Therefore the Einstein equation tells us that the remaining two curvatures must cancel, and must therefore be of *equal* strength. In other words, in [30.7a] the stretching force in the horizontal direction must *equal* the compressing force in the vertical direction, as illustrated.

Returning to [30.7a], the dashed circle of particles is initially deformed into the solid ellipse. But, half a wavelength later, the tidal forces have completely flipped, as illustrated in [30.7b]: now there is compression in the direction where the wave previously caused stretching, and visa versa. As the wave passes, the force field oscillates back and forth between these two opposite patterns of force.



[30.7] The Oscillating Tidal Force Field of a Gravitational Wave. *As the gravitational wave passes through a sphere of particles, it has no effect on the particles whose separation lies in the travel direction of the wave, which is here taken to be perpendicular to the page. In the plane orthogonal to this direction of travel, the wave causes oscillating expansion and contraction in two perpendicular directions, here shown as horizontal and vertical. Figure* **[a]** *shows the deformation of the (dashed) circle of test particles into the (solid) ellipse. Figure* **[b]** *shows the reversal of the tidal force field of the wave, half a wavelength later, now with the stretching orthogonal to the original stretching. As the wave passes, the force field oscillates back and forth between these two opposite patterns.*

As we discussed in Section 19.2, in general a phase portrait cannot represent the *magnitude* of the underlying vector field. In [30.7a] we have merely drawn a few random field lines, and there is no way to tell how strong the tidal forces are by looking at this. However, suppose the vector field is *divergence-free*, like an electric field in vacuum. If we draw the field lines in such a way that their *density/crowding* is proportional to the magnitude of the electric field in one place, then, as we follow the field lines away from that place, the density of field lines in the new place will automatically *continue*[11] to faithfully represent the magnitude of the electric field in the new place. So, although we have not done so in [30.7a], the streamlines for the gravitational wave's tidal force field can *also* be drawn in this way, because (as we now explain) it too is divergence-free.

If we take the plane [30.7] to be the *complex* plane, then the tidal force field in [30.7a] looks like the saddle point (aka crosspoint) first encountered in [19.4h] (p. 200), representing the complex function $(1/z)$. However, the Jacobi Equation tells us that the tidal force of the gravitational wave actually *increases linearly* with distance from the origin, and so it is represented by $\bar{z} = x - iy$, i.e., by the "Pólya vector field of z"; see page 201. Thus,

$$\nabla \bullet \begin{bmatrix} x \\ -y \end{bmatrix} = 0.$$

NOTE: One can also understand this from a more advanced point of view: the Pólya vector fields of *all* complex-analytic functions are *automatically* divergence-free (*and* curl-free); see VCA, page 483.

But the above interaction between the wave and the sphere of test particles is *unbelievably weak*. Einstein himself doubted that we would *ever* be able to detect his waves. While Einstein did indeed underestimate our future technological prowess, more significantly, he could not have imagined the *extraordinary violence* of cosmic events capable of generating gravitational waves. We

[11] See (VCA, pp. 494–502) and (Thorne and Blandford, 2017, §27.3.2)

will now describe one such event, which converted *astonishing* quantities of energy (in a fraction of a second!) into hugely powerful gravitational waves—veritable "tidal waves" [*sic*] in spacetime, that were capable of triggering detectors on Earth after spreading out and travelling for more than a billion years!

This 2015 detection of the first gravitational wave (at two independent detectors) was epoch-making in and of itself, but the scientists were able to go much further: they used Einstein's equation to work backwards from the *details* of the detected signal to determine, in remarkable detail, the cataclysmic event that had *created* the gravitational waves.

We will let one of these Nobel Prize–winning scientists, Kip Thorne, speak for himself. (Note that $M_\odot$ denotes the mass of our Sun.)

> On September 14, 2015, the advanced LIGO gravitational wave detectors made their first detection: a wave burst named GW150914 with amplitude 1.0×10^{-21}, duration ~ 150ms, and frequency chirping up from ~ 50Hz (when entering the LIGO band) to 240Hz. By comparing the observed waveform with those from numerical relativity simulations, the LIGO-VIRGO scientists deduced that the waves came from the merger of a $29M_\odot$ Black Hole with a $36M_\odot$ Black Hole, 1.2 billion light years from Earth, to form a $62M_\odot$ Black Hole, with a release of $3M_\odot c^2$ of energy in gravitational waves.—(Thorne and Blandford 2017, p. 1346)

We cannot allow the enormity of the energy release to pass without comment. Einstein's $E = mc^2$ translates this statement as saying that *during the last fraction of a second of the black hole collision, the equivalent of three times the mass of our Sun was converted into gravitational wave energy.* What does that mean? Let us begin with a grim but familiar instantiation of $E = mc^2$: the atomic bomb that destroyed Hiroshima in 1945, killing 75,000 people. In that atomic blast, the total amount of matter that was converted into energy weighed *less than a single raisin*. Let that sink in.

How, then, can we even *conceive* of the entire mass of our *Sun* being converted into gravitational wave energy?! Thorne and Blandford (2017, §27.5.5) offer this interpretation: in the last tenth of a second of the observed black hole collision, *the power output generating the GW150914 gravitational wave was 100 times the luminosity of all the stars in the observable Universe combined!*

For the fullest, most up-to-date—as of this writing!—technical treatment of the generation, propagation, and detection of gravitational waves, see Thorne and Blandford (2017, §27), and Schutz (2021).

Several more gravitational waves have been detected since GW150914.[12] In 2017 the same scientists achieved another first: they detected a burst, GW170817, which Einstein's equation allowed them to decode as arising from the merger of two *neutron stars*! Using the prescient theoretical work of Schutz (1986),[13] from 30 years earlier, these gravitational wave scientists were then able to use this data (in combination with optical data) to calculate the Hubble constant in a totally novel way, shining new light on the expansion of the Universe! Clearly, we have entered a new era of ***gravitational wave astronomy***.

Likewise, the gravitational bending and focusing of light, which Sir Arthur Eddington was the first to witness in 1919, has now led to ***gravitational lensing***: a technique that harnesses the gravitational focusing of light (and radio waves) as a *tool* for the study of very faint and distant objects.

Thus, in hindsight, our testing of Einstein's theory has been akin to looking down the wrong end of a telescope, checking to see if its lenses have been correctly configured. Scientists are now looking through the theory from the *correct* end, gazing out into spacetime, and using Einstein's

[12]The naming of these events is simple: GW stands for "gravitational wave"; 15 denotes 2015; 09 denotes the month, September; and 14 denotes the day of the month.

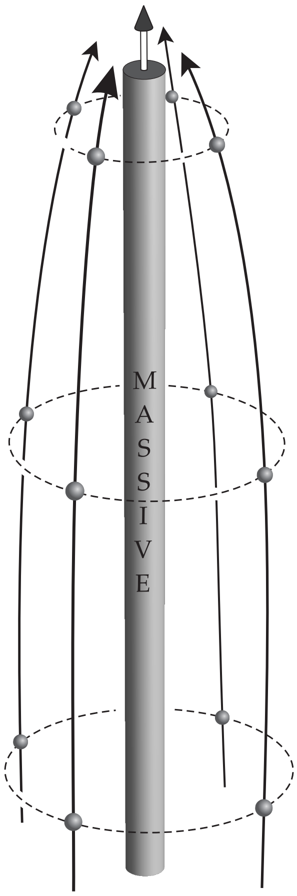
[13]For this work, the Royal Astronomical Society awarded Bernard Schutz the 2019 Eddington Medal.

equation to witness and decipher phenomena that would otherwise have remained invisible and inexplicable.

Everything we have described thus far has depended only on Einstein's *vacuum* field equation, (30.11). Even the black holes that generated GW150914 were describable by solutions of the vacuum equation, as were the gravitational waves themselves. But without the *matter* within the stars that collapsed billions of years ago, there would have been no black holes, no black hole collision, and no emitted gravitational waves!

We therefore now turn to the *full* Einstein Field Equation, with matter.

30.9 The Einstein Field Equation (with Matter) in Geometrical Form



[30.8] Spacetime Depiction of the Volume Reduction of a Free-Falling Sphere Containing Matter. *As usual, time runs vertically.*

Figure [30.4] illustrated the volume-preserving effect of the tidal forces when our sphere of particles enclosed only a vacuum. Figure [30.8] is instead a spacetime depiction of the *volume-reducing effect* when matter is present within the sphere. Again, this picture should ring a bell! It is almost exactly like [29.11b]! And, once again, this is no accident: [30.8] shows the effect—via the Sectional Jacobi Equation, (29.21)—of purely *positive* sectional curvatures in the presence of matter.

If the volume $\delta\mathcal{V}$ is filled with matter of density ρ, the Newtonian, volume-reducing acceleration is given by (30.5), but, in spacetime terms, this is again described by the average of the sectional curvatures, i.e, by the Ricci curvature formula, (29.29). Therefore, combining these results, we find that

$$\mathbf{R}\mathrm{icci}(\mathbf{v},\mathbf{v})\,\delta\mathcal{V} = -\ddot{\delta\mathcal{V}} = 4\pi G\rho\delta\mathcal{V},$$

and therefore

$$\mathbf{R}\mathrm{icci}(\mathbf{v},\mathbf{v}) = 4\pi G\rho. \qquad (30.14)$$

To make further progress, we must express the right-hand side of this equation in terms of the tensor that describes matter and energy in both Special and General Relativity. In common with the Ricci tensor, this new tensor, denoted $\mathbf{T}$, takes two vectors as input, and outputs a scalar. It is called the ***energy–momentum tensor*** or, equally commonly, the ***stress–energy tensor***. Like the Ricci tensor, it is *symmetric*:

$$\boxed{\mathbf{T}(\mathbf{w},\mathbf{v}) = \mathbf{T}(\mathbf{v},\mathbf{w}) \iff T_{ki} = T_{ik}.} \qquad (30.15)$$

Despite appearances to the contrary, this is not (at least primarily!) a physics book, so for a full description of $\mathbf{T}$ we must refer the reader to Misner, Thorne, and Wheeler (1973) or Thorne and Blandford (2017). For our purposes, the key feature of $\mathbf{T}$ is that for an observer with (normalized) 4-velocity $\mathbf{v}$, the total density of matter *and* energy is given by

$$\mathbf{T}(\mathbf{v},\mathbf{v}) = \rho_{\text{matter}} + \rho_{\text{energy}} \equiv \rho_{\text{total}}.$$

Recall that Einstein discovered that matter and energy are equivalent, so both of them curve spacetime. For example, the electromagnetic field has an energy density that Maxwell discovered to be proportional to $(|\mathbf{E}|^2 + |\mathbf{B}|^2)$, where $\mathbf{E}$ is the electric field, and $\mathbf{B}$ is the magnetic field. This electromagnetic energy curves spacetime just as surely as a planet does. Other fields, too, must be added in to obtain the complete energy–momentum tensor.

Thus, taking these relativistic effects into account, (30.14) becomes

$$\mathbf{R}\mathrm{icci}(\mathbf{v},\mathbf{v}) = 4\pi G\rho_{\mathrm{total}} = 4\pi\,\mathbf{T}(\mathbf{v},\mathbf{v}). \tag{30.16}$$

Note, for future comparison, that if we use the same tetrad as before, with $\mathbf{v} = \mathbf{e}_0$, then

$$\boxed{R_{00} = 4\pi G\rho_{\mathrm{total}}.} \tag{30.17}$$

But (30.16) is true if $\mathbf{v}$ is an *arbitrary* timelike vector. So, using the same trick as before, writing $\mathbf{v} = \mathbf{x} + \mathbf{y}$ and appealing to the symmetry of $\mathbf{T}$, we deduce [exercise] that

$$\boxed{\mathbf{R}\mathrm{icci} = 4\pi G\,\mathbf{T} \qquad \Longleftrightarrow \qquad R_{ik} = 4\pi G\,T_{ik}.} \tag{30.18}$$

This was in fact one of Einstein's original proposals for his field equation, beginning in 1913. (Of course if $\mathbf{T} = \mathbf{0}$ then it reduces to the (physically correct) vacuum equation, (30.11).) Indeed, Einstein still believed in (30.18) as late as the *11th* of November, 1915, exactly two weeks prior to his discovery of the *correct*, final field equation of General Relativity, on the 25th of November. For more on Einstein's decade-long struggle towards the correct equation, see Misner, Thorne, and Wheeler (1973, §17.7), or, better yet, Pais (1982, §14), which chronicles (essentially day by day) Einstein's mental thrashings throughout that fateful November.

The *modern* way to recognize the fatal flaw inherent in (30.18) is to observe that if this equation *were* true, the Differential Bianchi Identity, (29.17), would imply that *energy is not conserved!* To fully explain this would take us too far afield, but the essential *mathematical* fact is derived in Exercise 14. To properly understand the *physics*, we refer you to Misner, Thorne, and Wheeler (1973, §17), or Penrose (2005, §19.6), or Thorne and Blandford (2017, §25.8), or Wald (1984, §4.3). As we noted earlier, we should also point out that this is *not* how Einstein finally arrived at the truth, because *in 1915 he did not know the Differential Bianchi Identity!* See Pais (1982, p. 256).

The upshot is that the Differential Bianchi Identity almost uniquely determines the correction that *must* be made to (30.18) in order for conservation of energy to be restored.

To make this correction, we introduce the ***trace*** of the energy–momentum tensor:

$$T \equiv T^m{}_m = \rho_{\mathrm{total}} - (P_1 + P_2 + P_3)\,,$$

where P_1, P_2, P_3 are the *pressures* within the matter in our three orthogonal spatial directions, $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. Note that while pressure and density may seem very different concepts, a connection can be seen as follows:

$$\text{pressure} = \frac{\text{force}}{\text{area}} = \frac{\text{force} \times \text{distance}}{\text{volume}} = \frac{\text{energy}}{\text{volume}} = \frac{\text{mass}}{\text{volume}} = \text{density}.$$

We can now state the *correct*, final form of the ***Einstein Field Equation:***

$$\boxed{\mathbf{R}\mathrm{icci} = 8\pi G\left[\mathbf{T} - \tfrac{1}{2}T\mathbf{g}\right] \qquad \Longleftrightarrow \qquad R_{ik} = 8\pi G\left[T_{ik} - \tfrac{1}{2}Tg_{ik}\right].} \tag{30.19}$$

We note that this is *not* how you will find the equation written in almost all texts on General Relativity. Nevertheless, this is how *Einstein himself* originally wrote his equation, on the 25th of November, 1915! See Pais (1982, p. 256).

Let us state the more standard, mathematically equivalent form, though we leave the short proof to the technical references above:

$$\boxed{\mathbf{R}\mathrm{icci} - \tfrac{1}{2}R\mathbf{g} = 8\pi G\,\mathbf{T} \quad\Longleftrightarrow\quad R_{ik} - \tfrac{1}{2}Rg_{ik} = 8\pi G\,T_{ik},} \tag{30.20}$$

where $R \equiv R^m{}_m$ is the trace of $\mathbf{R}\mathrm{icci}$, called the ***curvature scalar***. The tensor on the left-hand side,

$$\boxed{\mathbf{G} \equiv \mathbf{R}\mathrm{icci} - \tfrac{1}{2}R\,\mathbf{g},} \tag{30.21}$$

is called the ***Einstein tensor***.[14] This notation leads to (by far) the most common way of writing Einstein's equation:[15,16]

$$\boxed{\mathbf{G} = 8\pi G\,\mathbf{T}.} \tag{30.22}$$

The final Einstein Field Equation, (30.19), appears very different from the original equation, (30.18), to which we were naturally led by our geometrical reasoning regarding Newton's Inverse-Square Law. In reality, they differ very little. Indeed, we now show that, physically, the Einstein equation only adds a correction to the original Newtonian equation, a correction that is extremely small under normal circumstances.

To see this, again choose the time axis along the observer's 4-velocity: $\mathbf{e}_0 = \mathbf{v}$. Then, since $g_{00} = 1$, we find that

$$R_{00} = 8\pi G\left[T_{00} - \frac{1}{2}Tg_{00}\right] = 8\pi G\left[\rho_{\text{total}} - \frac{1}{2}\Big\{\rho_{\text{total}} - (P_1 + P_2 + P_3)\Big\}\right],$$

so

$$\boxed{R_{00} = 4\pi G\left(\rho_{\text{total}} + P_1 + P_2 + P_3\right).} \tag{30.23}$$

Comparing this to our original, Newtonian-based equation, (30.17), we see that Einstein's equation *only* differs by the addition of the three pressure terms, and for matter under normal circumstances these terms are tiny compared to the mass/energy term, ρ_{total}.

It is important to remember the *significance* of the Ricci tensor, and hence of (30.23)—it tells us the *volume-compressing effect of the spacetime curvature*, via (29.29). We have finally arrived at the ***Einstein Field Equation in geometrical form***:

$$\boxed{\ddot{\delta\mathcal{V}} = -R_{00}\,\delta\mathcal{V} = -4\pi G\left(\rho_{\text{total}} + P_1 + P_2 + P_3\right)\delta\mathcal{V}.} \tag{30.24}$$

[14] For more on the geometry of the Einstein tensor, see Frankel (2011, §4).

[15] We remind the reader that the definition of the Ricci curvature varies, affecting its sign; thus, for example, Penrose (2005, §19.6) instead writes $\mathbf{G} = -8\pi G\,\mathbf{T}$.

[16] It is also common to use ***geometrized units***, in which both $c = 1$ *and* $G = 1$, in which case Einstein's equation becomes, simply, $\mathbf{G} = 8\pi\,\mathbf{T}$.

30.10 Gravitational Collapse to a Black Hole

During the normal life of a star, it is the continuous hydrogen bomb–like explosion within it that makes the star shine and that keeps the material of the star from collapsing inward under its own weight. But the star eventually exhausts its supply of its primary nuclear fuel, hydrogen, and as the nuclear fire goes out, the star starts to lose its power to resist gravity. Let us briefly consider the fate of normal star as its endgame begins.

In the course of gravitational collapse, the temperature and pressure can rise to the point that *new* nuclear reactions begin, renewing the fight against gravity by burning helium, and heavier elements. Also, the extreme circumstances of gravitational collapse can even cause the atoms themselves to be crushed, at which point, new *quantum-mechanical* forces resist the collapse, by virtue of the Pauli Exclusion Principle.

Precisely which nuclear reactions and which quantum-mechanical forces become dominant, and in what sequence, is the complicated dance[17] that governs the death throes of a star. But the *key* piece of information that largely determines this ultimate path is the *initial mass* M of the star.

Small stars (like our Sun) with $M < 8\ M_{\odot}$ may end up as stable white dwarfs, in which quantum-mechanical electron degeneracy pressure prevents further collapse. But much larger stars, with $M > 10\ M_{\odot}$, may ultimately go supernova, with much of the material of the star being exploded out into space, leaving only a core behind that then undergoes gravitational collapse. If the original star has $M < 30\ M_{\odot}$, then it is possible that the collapse of the postsupernova core will be halted by neutron degeneracy pressure (and other forces), leading to the formation of a stable, rapidly spinning neutron star.

But Einstein's equation predicts that if the collapsing core is sufficiently massive, something counterintuitive and almost paradoxical-seeming can happen, resulting in the creation of a *black hole*.

In the extreme circumstances of such a gravitational collapse, the pressure terms in Einstein's equation, (30.24), can become very significant as the matter is compressed and the speed of the atoms approaches the speed of light. Ordinary physical intuition tells us that these internal pressures will fight the collapse and perhaps even halt it. But, on the contrary, the Einstein equation tells us that the mounting pressures only *increase* the volume-crushing power of gravity! The harder the star fights its collapse, the more gravity tightens its stranglehold.

If, ultimately, the postsupernova core that undergoes gravitational collapse has a *mass that is greater than about*[18] $2\ M_{\odot}$—the ***Tolman–Oppenheimer–Volkoff limit***, TOV for short—this purely Einsteinian effect leads to *a point of no return*: now *no* force in the Universe can halt the collapse. See [30.9]. Gravity will relentlessly crush the entire core down to a point of infinite density and infinite tidal forces at $r = 0$, called a ***spacetime singularity***. After the collapse has finished, what remains is a pure, vacuum gravitational field.[19]

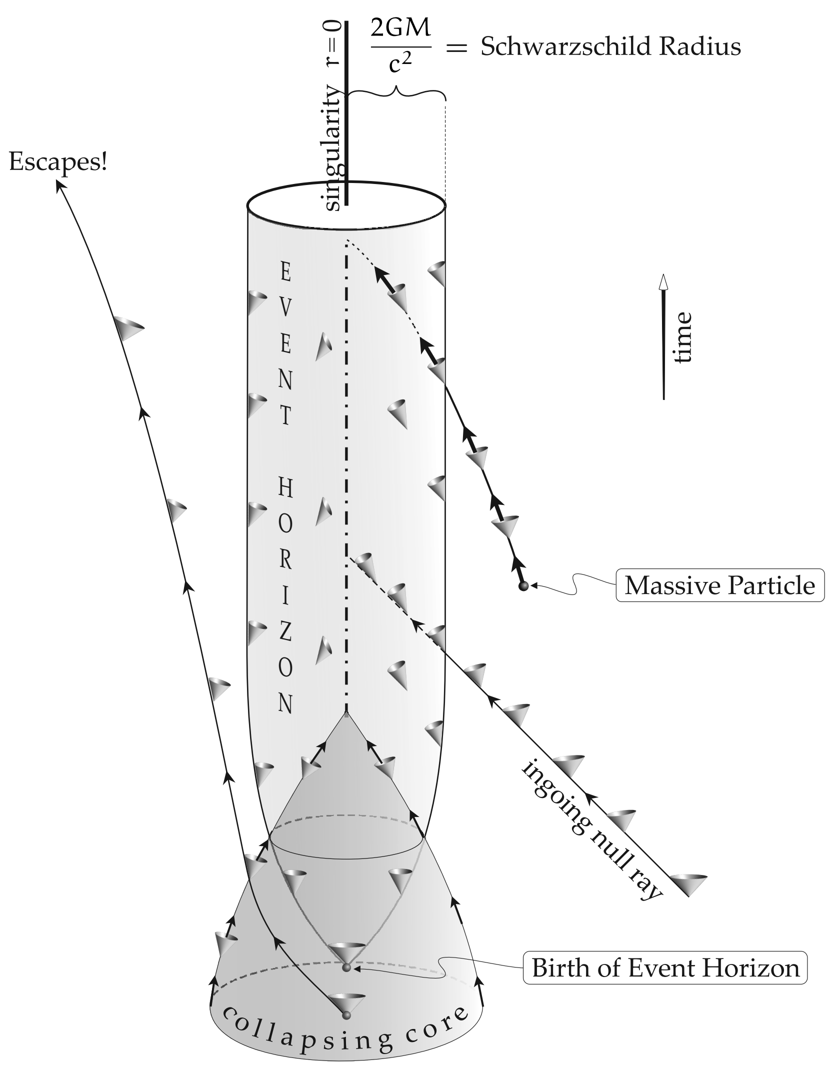
Imagine that flashes of light are emitted from the centre of the core during the collapse, and further imagine that these flashes can pass through through the matter of the core, as though they were neutrinos. The fate of an individual flash depends on *when* it is emitted. If the flash is emitted early enough, then, as illustrated, it is able to escape the gravitational field.

But there comes a critical moment when the flash from the centre expands at first but then slows and ultimately *hovers* at the Schwarzschild radius, $r_s = (2GM/c^2)$. This hovering sphere of light is the ***event horizon***, and its interior is a ***black hole***: no matter or information can escape this

[17]For the fascinating details of this dance, see Schutz (2003, §12).

[18]The precise value of the TOV limit is a work in progress; $2.2\ M_{\odot}$ was the best estimate when this was written, in 2019.

[19]***Birkhoff's Theorem*** tells us that as the spherical mass M shrinks and remains spherical, the geometry of the vacuum outside it *must* be the Schwarzschild solution, and that the M-value in the metric does not change as the object's radius R shrinks; the only change is the size of the vacuum region $r > R$ to which the solution applies.



[30.9] The Birth of a Black Hole: the Gravitational Collapse of a Postsupernova Core. *Einstein's field equation tells us that the collapse of a core of sufficient mass will reach a point of no return: gravity will relentlessly crush the entire core down to a point of infinite density and infinite tidal forces at* $r = 0$*, called a* **spacetime singularity***; what remains is a pure, vacuum gravitational field. If a flash is emitted from the centre of the collapsing core early enough, it can escape the gravitational field. But there comes a critical moment when the flash from the centre expands at first but then slows and ultimately* hovers *at the Schwarzschild radius,* $r_s = (2GM/c^2)$*. This hovering sphere of light is the* **event horizon***, and its interior is a* **black hole***: no matter or information can escape this region once it has been formed. The null cones are tangent to the horizon and so they allow matter and light to pass inwards, but never outwards, for matter must always travel* inside *the null cones.*

region once it has been formed. The null cones are tangent to the horizon and so they allow matter and light to pass inwards, but never outwards, for matter must always travel *inside* the null cones.

Therefore, once an object has fallen into the black hole, it is relentlessly dragged towards the central singularity, and the tidal forces exerted on it grow in proportion to $(1/r^3)$. Any solid object must ultimately succumb to these rapidly growing tidal forces, being stretched out in the radial direction, and compressed in the perpendicular directions. As the tidal forces tends to infinity, this process will stretch/compress the object into a long thin shape, like a piece of spaghetti. For this reason, physicists actually call this process ***spaghettification***!

Black holes can also form in other ways. We have mentioned that a gravitational wave (GW170817) detected in 2017 originated from two neutron stars spiralling around and into each other, finally colliding and merging. After the merger, the resulting object's mass exceeded the TOV limit. Remarkably, the gravitational wave scientists were able to spatially *localize* the source to a sufficiently small patch of the southern sky that traditional X-ray astronomers were then able to look there, successfully identifying and studying the aftermath of the neutron-star collision. This X-ray data now seems to confirm that the merged neutron stars immediately underwent full collapse to form a black hole!

We now know that the Universe is littered with black holes of these types, a few solar masses each. However, in addition, astronomers have now confirmed that almost every galaxy (including our own Milky Way) has, at its centre, a totally different kind of black hole—a ***supermassive black hole*** that is *millions* or even *thousands of millions* times more massive than our Sun! Although astrophysicists have hypothesized different mechanisms by which they might have been formed, the true origin of the supermassive black holes remains a mystery.

As we noted earlier, if a body of linear dimensions l passes through the event horizon of a black hole of mass M, the tidal forces exerted on it there are of order $(c^6 l/G^2)(1/M^2)$, which gets *smaller* as M gets bigger. For a supermassive black hole with $M = 10^8\ M_\odot$, this force is much less than the tidal force we experience on the surface of the Earth! Thus, if you were to fly your spaceship into such black hole, you would not even realize that you had done so—*you would feel nothing* as you crossed the event horizon.

Nevertheless, regardless of the power of your spaceship's engines (and how you might attempt to deploy them), after you cross the horizon, the *maximum* time you have till you hit the singularity is proportional to the mass of the black hole, and is given by[20]

$$\boxed{t_{\max} = \frac{M}{M_\odot} \times 15.5 \times 10^{-6} \text{ seconds.}}$$

If you were "lucky enough" to fly your spaceship into a black hole of sufficient mass—a few galaxies' worth of mass will do!—then you could even *live* inside the black hole for *years* thereafter; remember, your friends outside the hole can still send care packages *into* the hole.

However, for more realistic supermassive black holes, the outlook is more grim. For example, the black hole at the centre of our own Milky Way galaxy (called Sagittarius A*) has an estimated mass[21] $M = 3.6 \times 10^6\ M_\odot$, so if you were to cross its horizon, your maximum survival time would be *less than one minute*!

Regardless of $t_{\max}$, once you have entered a black hole, there can be no escape—sadly, your ultimate and completely inevitable demise is spaghettification!

30.11 The Cosmological Constant: *"The Greatest Blunder of My Life."*

We began this chapter with Einstein's happiest thought, and we shall end it with his *least* happy thought.

But this seemingly sad movie has a sudden, bizarre, *superhappy* plot twist at the very end! Indeed, after the curtain falls, and you finally walk through the exit door of Act IV, you will probably be muttering to yourself, "That could *never* happen in real life!"

[20] See Taylor and Wheeler (2000, §3–21).

[21] Measurements in 2018 revised upward the earlier estimate of $M = 2.6 \times 10^6\ M_\odot$.

In 1916 the widespread belief (based more on philosophy than science) was that the Universe was *static*, unchanging. But Einstein's field equation was quickly discovered to be in conflict with this idea—*it automatically led to a Universe that was expanding or contracting.*

Mind you, at this time, the "Universe" was thought to comprise *only* of our Milky Way galaxy—the existence of galaxies beyond our own was first discovered by Edwin Hubble in 1924.

In order to salvage a static Universe,[22] Einstein employed a daring and desperate gambit—despite the fact that it was the fruit of ten years of intense labour, he *changed* his field equation, (30.22)!

We have noted that the Differential Bianchi Identity, (29.17), *almost* uniquely determines the change that must be made to Einstein's original equation (30.18) in order to restore conservation of energy. However, Einstein realized that there remained one, and *only* one, additional freedom in his equation: he could add to the left-hand side a constant multiple Λ of the metric tensor, $\mathbf{g}$.

Einstein made this change to his equation in 1917, and to distinguish this new equation from the original 1915 equation, we shall call it the

Cosmological Einstein Field Equation: $$\boxed{\mathbf{G} + \Lambda \mathbf{g} = 8\pi G\, \mathbf{T}.} \qquad (30.25)$$

If true, this constant Λ—which is called the ***cosmological constant***—would need to be extraordinarily small to avoid conflict with observational evidence. At the scale of the solar system, or even the galaxy, the new Cosmological Field Equation would then make the same predictions as the original field equation. Only at the cosmological scale would its predictions be starkly different.

In 1929, twelve years after Einstein had made this change to his equation, Edwin Hubble made a *second* extraordinary discovery. Building on his earlier discovery of galaxies beyond our own, he sought to measure their distances from us and then to correlate those distances with the radial speed of the galaxies (measured using their redshifts). He found that the galaxies all had velocities directed *away* from us, and that these *velocities of the galaxies were proportional to their distances from us*. This is now known as ***Hubble's Law***.

Thus the Universe was *not* static, after all—it was *expanding!* Einstein realized that if he had only remained faithful to his original equation of 1915, he could have made the most extraordinary scientific prediction in human history! He told George Gamow that his introduction of the cosmological constant had been, "*The greatest blunder of my life.*" Thereafter, Einstein retracted the Cosmological Field Equation in favour of the original.

In subsequent decades, many (most?) experts tended to side with Einstein's retraction, believing that Λ was identically zero, and that the more elegant, original Einstein equation was correct. Nevertheless, once the Λ-genie was out of the bottle, it was impossible to put it back in. Thus, research papers always tended to hedge their bets, at least examining how their conclusions might be affected if $\Lambda \neq 0$.

We now leap forward in time to 1998, and to a cosmological discovery[23] as great as Hubble's discovery, 70 years earlier. Although Hubble had found that the Universe was expanding, the expectation had been that the positive, attracting gravitational pull of the matter within would *slow* the expansion of the Universe over time. However, by observing ***type Ia supernovas***—which are exploding white dwarfs that have exceeded their stability limit, and which therefore have similar masses and therefore standardized intrinsic luminosity—the scientists made the astonishing discovery that *the expansion of the Universe is accelerating!*

But how can we *explain* this gobsmacking discovery?! Einstein's answer had been sitting there, patiently waiting, since 1917!

---

[22] The full story of Einstein's motivations is far more complex and subtle. See Pais (1982, §15c).

[23] The discovery was made by two independent research teams, and the 2011 Nobel Prize in Physics was divided, one-half awarded to Saul Perlmutter, the other half jointly to Brian P. Schmidt and Adam G. Riess.

If $\Lambda > 0$, then (moving $\Lambda\mathbf{g}$ to the right-hand side of the equation) its effect is *mathematically equivalent* to a fictitious[24] *negative energy density* $\rho = -(\Lambda/8\pi G)$, *resulting in* ***gravitational repulsion***. While [30.8] illustrates the attractive nature of ordinary matter, and positive sectional curvatures, the *opposite* behaviour in the presence of *negative* energy and *negative* sectional curvatures is illustrated in [29.11a], page 301. When viewed as a spacetime picture of the expansion of the Universe, this shows the *repulsive* effect of *negative* energy arising from a *positive* value of Λ, evidently resulting in *accelerating* expansion of the Universe.

In short, the observed accelerating expansion of the Universe agrees (in detail) with the Cosmological Einstein Field equation, with $\Lambda > 0$. This, then, is the happy, scarcely believable, final plot twist: Einstein's "greatest blunder" turned out to be, instead, one of his greatest strokes of genius!

30.12 The End

We hope that these final two chapters of Act IV have inspired you to grapple in greater depth with the connected wonders of Riemann's curvature and Einstein's curved spacetime. However, *we* must stop here—a *full* explanation would require Volumes 2 and 3!

If we have done our job, you cannot wait to learn more—let the *Further Reading* section at the end of this book be your guide. There we recommend *many* excellent works, each with its own, distinctive *spécialité de la maison*.

That said, let us impatiently assert that we can think of no smoother or faster road to enlightenment than Penrose's *Road to Reality* (2005), for it contains his countless original and revelatory insights into both the mathematics and the physics, all expressed with astonishing clarity by means of his singularly beautiful hand drawings.

For even greater technical detail, we also *strongly* recommend—see Prologue—*Gravitation* by Misner, Thorne, and Wheeler (1973). And for more up-to-date developments, especially on gravitational waves and modern cosmology, we highly recommend the massive and authoritative work, *Modern Classical Physics*, by Thorne and Blandford (2017).

Last, after reading Act V, we recommend Dray (2015), which uses Forms to clarify the geometry of General Relativity. Indeed, it is to *Forms* that we now turn, for they are the subject of our final Act.

But, as for our treatment of purely *Visual* Differential Geometry, this is, finally, the end.

[24] While I am personally content to accept the Cosmological Field Equation with $\Lambda > 0$, many famous scientists *instead* believe that there exists an actual, *nonfictitious thing*, permeating the entire Universe, that causes this repulsion—they call it ***dark energy***.

Chapter 31

Exercises for Act IV

Extrinsic Constructions

1. **Geodesic Curvature on the Sphere.** On the sphere of radius R, suppose a particle travels at unit speed along the circle at latitude ϕ.

   (i) What is the magnitude and direction of the acceleration of the particle?

   (ii) Sketch the projection of this acceleration onto the tangent plane of the sphere, and deduce that the geodesic curvature is $\kappa_g = \frac{\cot\phi}{R}$.

   (iii) Verify that this formula for κ_g yields the geometrically correct answer as $\phi \to 0$ and as $\phi \to (\pi/2)$.

   (iv) Show that the component of the acceleration directed towards the centre of the sphere has magnitude $(1/R)$, *independent* of ϕ. Explain this geometrically.

2. **Geodesic Curvature on the Cone.** Reconsider the cone of semivertical angle α shown in [14.3], page 145. Suppose a particle travels at unit speed along a horizontal circle of radius r on this cone.

   (i) What is the magnitude and direction of the acceleration of the particle?

   (ii) Sketch the projection of this acceleration onto the tangent plane of the cone, and deduce that the geodesic curvature is $\kappa_g = \frac{\sin\alpha}{r}$.

   (iii) Let s be the distance within the cone (along a geodesic generator) from the vertex to the circle. Cut the cone along a generator and press it flat onto the plane, as in [14.3], page 145, so that the circle becomes an arc of a circle in the plane of radius s and therefore of curvature $(1/s)$ within the plane. Verify that this curvature is the *same* as κ_g. (NOTE: What is the radius of this circle from the *intrinsic* point of view? Well, if two ants at neighbouring points of the circle start walking inward along radii—along geodesic generators orthogonal to the circle—they will meet at the vertex, so the intrinsic radius is simply the distance s from the vertex to the circle. The natural *intrinsic* definition of the curvature of the circle is therefore, once again, $(1/s)$.)

   (iv) Let ρ be the distance along the normal from the particle to the axis of symmetry. Show that the component of the acceleration along the normal is $(1/\rho)$, and explain this geometrically.

3. **Geodesic Curvature along Touching Surfaces.** Suppose that two surfaces touch along a common curve $\mathcal{C}$.

   (i) If $\mathcal{C}$ is a geodesic of one surface, explain geometrically why it must also be a geodesic of the other surface.

   (ii) If $\mathcal{C}$ has geodesic curvature κ_g for one surface, explain geometrically why it must also have geodesic curvature κ_g for the other surface.

(iii) Deduce (i) as a special case of (ii).

(iv) Use the *formulas* for κ_g in the previous two exercises to confirm (ii) in the case that one surface is a sphere, and the other surface is a cone, and $\mathcal{C}$ is a circle of latitude on the sphere.

Intrinsic Constructions

4. Geodesic Curvature via Intrinsic Differentiation. Let $\mathbf{v}$ be the velocity of a particle travelling at unit speed along a trajectory K over the surface $\mathcal{S}$ of some peelable fruit. If o is the launch point, let $(\mathbf{e}_1, \mathbf{e}_2)$ be an orthonormal basis for T_o. Now parallel transport $\mathbf{e}_1$ along K, and let $\theta_{||}$ denote the angle between this parallel-transported $\mathbf{e}_1$ and $\mathbf{v}$.

(i) Explain why $\mathbf{e}_2$ is automatically parallel-transported along K, too:

$$D_{\mathbf{v}}\mathbf{e}_1 = 0 \qquad \Longrightarrow \qquad D_{\mathbf{v}}\mathbf{e}_2 = 0.$$

(ii) Let $\mathbf{v}^{\perp}$ be the unit vector obtained by rotating $\mathbf{v}$ by $(\pi/2)$ within T_p. Write $\mathbf{v}$ and $\mathbf{v}^{\perp}$ in terms of $(\mathbf{e}_1, \mathbf{e}_2)$. Now use *calculation* to prove that

$$\boldsymbol{\kappa}_g = D_{\mathbf{v}}\mathbf{v} = \left[D_{\mathbf{v}}\theta_{||}\right]\mathbf{v}^{\perp},$$

thereby proving (23.4): $|\boldsymbol{\kappa}_g| = |D_{\mathbf{v}}\theta_{||}|$.

(iii) Draw a sketch illustrating $\mathbf{v}$, $(\mathbf{e}_1, \mathbf{e}_2)$, and $\theta_{||}$ at several points along K, *after* the narrow strip of peel surrounding K has been removed from $\mathcal{S}$ and pressed flat onto the plane.

(iv) Deduce that the geodesic curvature of K is indeed the ordinary curvature of the plane curve obtained by flattening the narrow strip surrounding K.

Holonomy

5. Holonomy on Cone and Sphere.

(i) Reconsider the cone of the semivertical angle α shown in [14.3], page 145. There we showed that if the tip of the cone is blunted, its spherical image allows us assign a definite curvature $\mathcal{K}$ to the tip of the otherwise intrinsically flat cone, given by (14.2):

$$\mathcal{K}(\text{spike}) = \beta = \text{split angle of flattened spike} = 2\pi(1 - \sin\alpha).$$

By carrying out parallel transport within the flattened cone in [14.3], show that holonomy assigns the *same* total curvature to the spike.

(ii) On $\mathbb{S}^2$, let us use extrinsic, potato-peeler parallel transport to find the holonomy of a circle of latitude of fixed angle ϕ. Imagine a cone of semivertical angle α resting on the sphere so that it touches the sphere along this circle. The strip of peel of the sphere along this circle is therefore (ultimately) the same as the strip of the touching cone. Show that $\alpha = \frac{\pi}{2} - \phi$, and use (i) to confirm that the total curvature within the polar cap is indeed the holonomy of the circle of latitude that bounds it.

(iii) On the equator of $\mathbb{S}^2$, imagine a vector pointing due east. Now parallel transport it due east along the geodesic equator, so that it returns home seemingly unchanged, i.e., with *vanishing* holonomy. But this loop encloses half the sphere, with total curvature 2π! Use (ii) to reconcile these facts, by gradually increasing ϕ from 0 to $(\pi/2)$.

6. General Local Gauss–Bonnet Theorem.

(i) By approximating a smooth, closed loop L with a geodesic m-gon, then letting $m \to \infty$, deduce that the holonomy formula (24.6) becomes,

$$\mathcal{R}(L) = 2\pi - \oint_L \kappa_g \, ds,$$

where κ_g is the geodesic curvature along L, and s is distance along L.

(ii) If P is a closed "polygon" with external angles φ_i, but with edges that are *not* geodesic (i.e., $\kappa_g \neq 0$), then show that the generalization of (24.6) is,

$$\boxed{\mathcal{R}(P) = 2\pi - \left[\oint_P \kappa_g \, ds + \sum_i \varphi_i\right].}$$

(iii) If R denotes the interior of P, deduce the ***General Local Gauss–Bonnet Theorem***:

$$\boxed{\iint_R \mathcal{K} \, d\mathcal{A} = 2\pi - \left[\oint_P \kappa_g \, ds + \sum_i \varphi_i\right].}$$

Curvature as a Force between Neighbouring Geodesics

7. Minding's Theorem. Using the same geodesic polar coordinates as in Section 28.3, the metric takes the form

$$ds^2 = dr^2 + g^2(r) \, d\theta^2,$$

so that

$$g'' = -\mathcal{K} g.$$

Using the fact that $g(r) \asymp r$ (as r vanishes), solve this differential equation in each of the three cases of *constant* $\mathcal{K}$:

(i) If $\mathcal{K} = 0$ everywhere, deduce that the surface is locally isometric to the Euclidean plane.

(ii) If $\mathcal{K} = (1/R^2)$ is constant throughout the surface, deduce that the space is locally isometric to the sphere of radius R.

(iii) If $\mathcal{K} = -(1/R^2)$ is constant throughout the surface, deduce that the space is locally isometric to the pseudosphere of radius R.

8. Jacobi Equation on a General Surface of Revolution. As in Exercise 22, page 89, imagine a particle travelling along a curve in the (x, y)-plane at unit speed, and let its position at time t be $[x(t), y(t)]$. Now imagine rotating this plane through angle θ about the x-axis. As θ varies from 0 to 2π, the curve sweeps out a surface of revolution.

(i) Explain why $\dot{x}^2 + \dot{y}^2 = 1$, where the dot represents the time derivative.

(ii) Show geometrically that the metric of the surface is $ds^2 = dt^2 + y^2 \, d\theta^2$.

(iii) By considering the relative acceleration of two neighbouring meridian geodesics, deduce from the Jacobi Equation (28.2) that $\mathcal{K} = -\ddot{y}/y$. (Recall that this is the same formula (13.4) that we previously obtained in the course of proving the *Theorema Egregium* for surfaces of revolution.)

9. Gauss's Lemma via Computation. As in [28.6], consider two neighbouring, unit-speed geodesics launched from o with angular separation $\delta\theta$, and let $\mathbf{v}$ denote their unit velocity vectors. Let $\boldsymbol{\xi}$ be the connecting vector between the geodesics, connecting two points at the *same distance* $\sigma = t$ from o, so that both lie on the geodesic circle $K(\sigma)$. Then, as $\delta\theta \to 0$, $\boldsymbol{\xi}$ is tangent to $K(\sigma)$. To establish Gauss's Lemma, (28.4), we must therefore show that $\mathbf{v} \bullet \boldsymbol{\xi} = 0$. (NOTE: In the following, $\nabla_{\mathbf{v}}$ is the ordinary $\mathbb{R}^3$ derivative, not the intrinsic surface derivative, $D_{\mathbf{v}} = \boldsymbol{\nabla}_{\mathbf{v}}$.)

(i) Explain why $\lim_{\sigma\to 0} \mathbf{v} \bullet \boldsymbol{\xi} = 0$.

(ii) Deduce that to prove $\mathbf{v} \bullet \boldsymbol{\xi} = 0$, it suffices to show that $\nabla_{\mathbf{v}} [\mathbf{v} \bullet \boldsymbol{\xi}] = 0$.

(iii) Explain why $\mathbf{v} \bullet \nabla_{\boldsymbol{\xi}} \mathbf{v} = 0$.

(iv) Explain why $[\mathbf{v}, \boldsymbol{\xi}] = 0$.

(v) Using the fact that $\mathbf{v}$ is the velocity of a *geodesic*, deduce that $\boldsymbol{\xi} \bullet \nabla_{\mathbf{v}} \mathbf{v} = 0$.

(vi) Combine the three previous results to prove that

$$\nabla_{\mathbf{v}} [\mathbf{v} \bullet \boldsymbol{\xi}] = 0,$$

thereby completing the computational proof of Gauss's Lemma.

Riemann's Curvature

10. Two Symmetries of the Riemann Tensor.

(i) Let us confirm the First (Algebraic) Bianchi Identity, (29.15), which is also called the ***Bianchi Symmetry***:

$$\mathcal{R}(\mathbf{u}, \mathbf{v})\, \mathbf{w} + \mathcal{R}(\mathbf{v}, \mathbf{w})\, \mathbf{u} + \mathcal{R}(\mathbf{w}, \mathbf{u})\, \mathbf{v} = \mathbf{0} \quad \Longleftrightarrow \quad R_{ijkm} + R_{jkim} + R_{kijm} = 0.$$

To prove the general result it actually suffices (by virtue of linearity) to prove it in the case that all three vector fields are ***coordinate vector fields***, in which case their commutators all vanish. Prove that in this case,

$$\mathcal{R}(\mathbf{u}, \mathbf{v})\, \mathbf{w} + \mathcal{R}(\mathbf{v}, \mathbf{w})\, \mathbf{u} + \mathcal{R}(\mathbf{w}, \mathbf{u})\, \mathbf{v} = \boldsymbol{\nabla}_{\mathbf{u}}[\mathbf{v}, \mathbf{w}] + \boldsymbol{\nabla}_{\mathbf{v}}[\mathbf{w}, \mathbf{u}] + \boldsymbol{\nabla}_{\mathbf{w}}[\mathbf{u}, \mathbf{v}] = 0.$$

A more elegant proof using curvature 2-forms can be found in Section 38.12.4.

(ii) With the same commuting vector fields as in (i), we can give a *geometrical* explanation of the Bianchi Symmetry, as follows. (NOTE: I certainly do not believe I am the first to discover this proof, but I have not been able to find it in print, and therefore do not know to whom credit should be assigned.) Starting at some point, draw $\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w}$, where ϵ is small and ultimately vanishing. Begin to construct a polyhedral "box" by completing each pair of these edges, creating three parallelogram faces, which are *closed* by virtue of their vanishing commutators. Now construct (and sketch) two new vector edges $\mathbf{A}$ and $\mathbf{B}$ of this box, obtained by parallel transporting $\epsilon\mathbf{w}$ first along $\epsilon\mathbf{v}$ and then along $\epsilon\mathbf{u}$ (yielding $\mathbf{A}$), and, in the reverse order, first along $\epsilon\mathbf{u}$ and then along $\epsilon\mathbf{v}$ (yielding $\mathbf{B}$). Next connect the end of $\mathbf{B}$ to the end of $\mathbf{A}$, thereby creating a new vector edge of a new triangular face of the box. Deduce that this new edge is given by,

$$\mathbf{A} - \mathbf{B} \asymp \epsilon^3\, \mathcal{R}(\mathbf{u}, \mathbf{v})\, \mathbf{w}.$$

Repeat this construction using the other two faces, thereby constructing two more triangular faces of the box. Your figure should now reveal that $\epsilon^3\,\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w}$, $\epsilon^3\,\mathcal{R}(\mathbf{v},\mathbf{w})\,\mathbf{u}$, and $\epsilon^3\,\mathcal{R}(\mathbf{w},\mathbf{u})\,\mathbf{v}$ form the vector edges of a fourth *triangular face* that closes the box, thereby proving the Bianchi Symmetry.

(iii) Let us confirm that *the Riemann tensor is symmetric under interchange of the first and second pairs of vectors*, (29.16):

$$[\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}]\bullet\mathbf{y}=[\mathcal{R}(\mathbf{x},\mathbf{y})\,\mathbf{u}]\bullet\mathbf{v}\qquad\Longleftrightarrow\qquad R_{ijkl}=R_{klij}.$$

As of this writing, I have failed to make geometrical sense of this result, so we must now resort to proving it by means of Satanic (see Prologue) Algebra. If we define

$$\mathbf{B}(\mathbf{u},\mathbf{v},\mathbf{x},\mathbf{y})\equiv[\mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{x}]\bullet\mathbf{y}+[\mathcal{R}(\mathbf{v},\mathbf{x})\,\mathbf{u}]\bullet\mathbf{y}+[\mathcal{R}(\mathbf{x},\mathbf{u})\,\mathbf{v}]\bullet\mathbf{y},$$

then

$$\mathbf{B}(\mathbf{u},\mathbf{v},\mathbf{x},\mathbf{y})=0,$$

by virtue of the Bianchi Symmetry in (i). Recalling that the Riemann tensor is antisymmetric in both the first two slots and in the second two slots, prove the result by cancelling like terms from the following (manifestly trivial!) identity:

$$\mathbf{B}(\mathbf{u},\mathbf{v},\mathbf{x},\mathbf{y})+\mathbf{B}(\mathbf{v},\mathbf{x},\mathbf{y},\mathbf{u})=0=\mathbf{B}(\mathbf{x},\mathbf{y},\mathbf{u},\mathbf{v})+\mathbf{B}(\mathbf{y},\mathbf{u},\mathbf{v},\mathbf{x}).$$

(Can you smell the sulphur?)

11. Counting the Components of the Riemann Tensor. (The following proof, which we found in Lightman et al. (1975), is shorter and simpler than other standard proofs we have seen.)

(i) Given that the components R_{ijkl} of the Riemann tensor in an n-manifold are antisymmetric in ij and kl, deduce that there are $P=\frac{1}{2}n(n-1)$ nontrivial ways of choosing *pairs* ij, and likewise P ways of choosing pairs kl.

(ii) Given the symmetry (29.16) under interchange of the first and second pair of indices, $R_{ijkl}=R_{klij}$ (see previous exercise), deduce that if we only take these pair symmetries into account, there are $\frac{1}{2}P(P+1)$ independent ways of choosing $ijkl$.

(iii) Defining $B_{ijkl}\equiv R_{ijkl}+R_{iklj}+R_{iljk}$, as in the previous exercise, the *Bianchi Symmetry* (29.15) states that $B_{ijkl}=0$ (see previous exercise). Verify that the pair symmetries now ensure that B_{ijkl} is *totally* antisymmetric on all four indices, and that the constraint $B_{ijkl}=0$ is therefore trivially satisfied *unless all four indices are distinct*.

(iv) If $n<4$, the Bianchi Symmetry therefore does not impose any new constraints. Use (ii) to deduce that if $n=2$ the Riemann tensor has one component (the Gaussian curvature), and if $n=3$ it has six components.

(v) If $n\geqslant 4$, deduce that the number of additional constraints resulting from the Bianchi symmetry is equal to the number of ways of choosing four objects from n objects.

(vi) Deduce that the number of independent components of the Riemann tensor is

$$\frac{1}{2}P(P+1)-\frac{n!}{(n-4)!\,4!},$$

in which the second term correctly disappears if $n<4$.

(vii) Verify that this does indeed yield formula (29.1): *The number of independent components of the Riemann tensor is*

$$\frac{1}{12}n^2(n^2-1).$$

12. The Exponential Operator and Curvature. Recall the exponential series:

$$\exp(x) = e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \cdots .$$

Now compare this to the Taylor series of a general function $f(x)$:

$$\begin{aligned}
f(a + \delta x) &= f + \delta x \frac{df}{dx} + \frac{1}{2!}(\delta x)^2 \frac{d^2 f}{dx^2} + \frac{1}{3!}(\delta x)^3 \frac{d^3 f}{dx^3} + \cdots \bigg|_a \\
&= \left[1 + \delta x \frac{d}{dx} + \frac{1}{2!}(\delta x)^2 \left(\frac{d}{dx}\right)^2 + \frac{1}{3!}(\delta x)^3 \left(\frac{d}{dx}\right)^3 + \cdots \right] f \bigg|_a \\
&\equiv \exp\left[\delta x \frac{d}{dx}\right] f \bigg|_a ,
\end{aligned}$$

in which the last line serves as the definition of the ***exponential operator***. The extension to vector fields follows naturally. If $\mathbf{w}$ is vector field defined in the vicinity of a, then, as we move distance δu in the direction of the unit vector $\mathbf{u}$,

$$\mathbf{w}(a + \mathbf{u}\,\delta u) = \exp\left[\delta u \nabla_{\mathbf{u}}\right] \mathbf{w}\big|_a .$$

(i) Reconsider the derivation of the Riemann tensor in [29.8], but suppose for simplicity's sake that $[\mathbf{u}, \mathbf{v}] = 0$, so that the parallelogram closes up, in which case the curvature operator simplifies to

$$\mathcal{R}(\mathbf{u}, \mathbf{v}) = [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}].$$

Explain why the vector holonomy of the parallelogram is given by

$$-\delta \mathbf{w}_{||} = [\exp(\delta u \nabla_{\mathbf{u}}), \exp(\delta v \nabla_{\mathbf{v}})]\,\mathbf{w}.$$

(ii) Deduce that

$$-\delta \mathbf{w}_{||} = (\delta u\, \delta v)\, \mathcal{R}(\mathbf{u}, \mathbf{v})\, \mathbf{w} + (\text{third-order error}),$$

where the "third-order error" is made up of terms involving $(\delta u)^p (\delta v)^q$, where $(p + q) \geqslant 3$.

Einstein's Curved Spacetime

13. Eclipses and the Tides. The tidal influence of the Moon on the oceans is more than twice as great as the Sun's, despite the fact the Sun's gravitational pull on the Earth is about 200 times more powerful than the Moon's. Let us try to understand this seemingly paradoxical reversal.

(i) It is a remarkable fact that total solar eclipses occur. This is only possible because of an empirical coincidence: the Moon and Sun look almost exactly the same (angular) size as seen from Earth. If the radii of the Moon and Sun are r_m and r_s, and their distances from Earth are R_m and R_s, respectively, deduce that $(R_m/R_s) \approx (r_m/r_s)$.

(ii) We have seen that the tidal force exerted on the oceans by a body of mass M at distance R from Earth is proportional to $\frac{M}{R^3}$. By appealing to (i), deduce that *the ratio of the lunar to solar tidal forces is the ratio of the* densities *of the Moon to the Sun.*

(iii) The average density of the Moon is approximately 3300 kilograms per cubic meter, while that of the Sun is approximately 1400 kilograms per cubic meter. Use (ii) to *explain* the strange opening fact!

14. Conservation of the Einstein Tensor. (NOTE: Unless you are already familiar with tensor ***contractions***, and with raising/lowering tensor indices, it would be best to defer this exercise until you have studied the relevant sections of Act V, namely, 33.7 and 33.8.) Recall from (30.21) that the *Einstein tensor* **G** is

$$\mathbf{G} \equiv \mathbf{R}\text{icci} - \frac{1}{2} R\, \mathbf{g}.$$

This exercise demonstrates that the Einstein tensor is "conserved," in the same way that energy–momentum is ***conserved***:

$$\boxed{\nabla^a G_{ab} = 0.}$$

This crucial, *purely mathematical* fact was unknown to Einstein when he wrote down his field equation, (30.19), on the 25th of November, 1915 (Einstein's original form being mathematically equivalent of the modern form, $\mathbf{G} = 8\pi\mathbf{T}$). *Subsequently*, it was recognized that the link that Einstein had discovered between geometry and matter actually *implies* that energy–momentum *must* be conserved: $\nabla^a T_{ab} = 0$!

(i) Check that the *Second (Differential) Bianchi Identity*, (29.17), can be written

$$\nabla_a R_{bcd}{}^e + \nabla_b R_{cad}{}^e + \nabla_c R_{abd}{}^e = 0.$$

(ii) Now perform two *contractions* of this equation, by (1) raising a and renaming d to a, so that the two a's are summed over; (2) renaming e to c, so that the two c's are summed over. Confirm that this yields

$$\nabla^a R_{bca}{}^c + \nabla_b R_c{}^a{}_a{}^c + \nabla_c R^a{}_{ba}{}^c = 0.$$

(iii) Show that the previous equation can be rewritten as

$$-\nabla^a R_{ba} + \nabla_b R - \nabla_c R_b{}^c = 0.$$

(iv) Show that this may in turn be rewritten as

$$\nabla^a R_{ba} - \frac{1}{2}\nabla_b R = 0.$$

(v) Deduce that $\nabla^a G_{ab} = 0$, as was to be shown.

15. Weyl Curvature. The 20 components of the Riemann tensor in spacetime can be split between the 10 components of the Ricci tensor (generated by matter and energy, via the Einstein field equation) and the 10 components that represent pure gravitational degrees of freedom, present in vacuum. These gravitational degrees of freedom are completely encoded by the ***Weyl curvature tensor***:

$$\boxed{C_{ij}{}^{kl} \equiv R_{ij}{}^{kl} - 2R_{[i}{}^{[k}\, g_{j]}{}^{l]} + \tfrac{1}{3}\, R\, g_{[i}{}^{k}\, g_{j]}{}^{l}.}$$

Here the square brackets enclosing pairs of indices denote the operation of ***antisymmetrization***, as defined later, in (33.9), on page 369.

(i) Explicitly perform these antisymmetrizations, and thereby deduce the following unwieldy (and hard-to-remember) formula:

$$C_{ijkl} = R_{ijkl} + \frac{1}{2}(R_{il}\, g_{jk} + R_{jk}\, g_{il} - R_{ik}\, g_{jl} - R_{jl}\, g_{ik}) + \frac{1}{6}R\,(g_{ik}\, g_{jl} - g_{jk}\, g_{il}).$$

(ii) Deduce from (i) that the Weyl tensor has the same symmetries as the full Riemann tensor.

(iii) Verify that the Einstein Field Equation immediately implies that, in vacuum, the Weyl tensor *is* the Riemann tensor.

(iv) Show that all the *traces* of the Weyl tensor vanish: in particular,

$$\boxed{C_{ij}{}^{ki} = 0.}$$

Thus the equivalent of the Ricci tensor—the matter–energy part of the Riemann tensor—vanishes for the Weyl tensor.

(v) (NOTE: The remainder of this exercise is more advanced, and really requires more than we have explained thus far. The solution can be found in the references cited at the conclusion of this exercise.) Suppose we subject the spacetime metric to a ***conformal transformation***:

$$\mathbf{g} \longrightarrow \Omega^2\, \mathbf{g},$$

where Ω is a function that *varies* from point to point throughout spacetime. But, at each point, all local distances are stretched by the *same* factor Ω, evaluated at that point. Such a transformation *preserves angles*, and also preserves the shapes of small objects. The Riemann tensor transforms in an extremely complex manner under such a conformal transformation. *However*, prove that the Weyl tensor—the purely gravitational/vacuum part—enjoys the remarkable and vitally important property that it merely *scales*:

$$\boxed{\boxed{C_{ijkl} \longrightarrow \Omega^2\, C_{ijkl}.}}$$

(vi) If we raise one index, show that we may rephrase this as saying that $C_{ijk}{}^{l}$ is ***conformally invariant***:

$$\boxed{\boxed{C_{ijk}{}^{l} \longrightarrow C_{ijk}{}^{l}.}}$$

NOTE: The most natural and elegant expression of the Weyl tensor is in terms of Penrose's 2-spinor formalism, where it takes the form of the ***Weyl conformal spinor*** Ψ_{ABCD}, which is *totally symmetric* and *conformally invariant*:

$$\boxed{\boxed{\Psi_{ABCD} = \Psi_{(ABCD)} \longrightarrow \Psi_{ABCD}.}}$$

Ψ_{ABCD} is also called, simply, the ***gravitational spinor***. See Penrose and Rindler (1984, §4.6, §6.8) or Wald (1984, §13.2). For an intuitive discussion of Weyl curvature, and Penrose's conjecture on its vital importance in characterizing the extraordinarily special nature of the Big Bang, see Penrose (2005, §19.7, §28.8).

ACT V

Forms

Chapter 32

1-Forms

32.1 Introduction

Act V represents cathartic release from four successive Acts of relentlessly strict geometrical rule.

As the Prologue foretold, our purpose now shall be to build the "Devil's machine," and to make it available to *undergraduate* students. We are speaking of a seductively powerful and elegant method of computation whose full name is ***The Exterior Calculus of Differential Forms***, which we have abbreviated here to ***Forms***.

Forms were discovered by the extraordinarily profound, original, and wide-ranging Élie Cartan (pictured in [32.1]) around 1900, more than a century ago, but it took even Cartan another 40 years to harness the *full* power of his discovery.

[32.1] *Élie Cartan (1869–1951).*

Our aim in Act V is to confront—succinctly, and in the plainest of Anglo-Saxon—a century-long scandal, namely, that the vast majority of undergraduates (in both mathematics and physics) will obtain their degrees without ever having glimpsed Cartan's Forms.

Although our principal purpose is to develop a novel method of *computation*, we desire to have our cake and eat it, too. That is, we shall endeavour to shed as much *geometrical* light as possible upon Cartan's Forms—much more so than in standard[1] treatments, which frequently careen into a perversely formal, shockingly abstract parallel Universe, in which Forms have been stripped of all vestiges of *meaning!*

Instead, our approach will be brutally concrete and vivid—occasionally downright lurid. We shall hold up Cartan's (Platonic!) Forms directly before your eyes, light gleaming upon their facets; we shall place them in your hands, so that you may feel their intricate shape and satisfying heft; finally, you will bear witness to their fearsome power.

[1]Misner, Thorne, and Wheeler (1973) and Schutz (1980) are two superb exceptions, and we *strongly* recommend both of them as companion (or follow-up) reading for Act V. Other noteworthy efforts can found in the *Further Reading* section at the end of this book.

But heed the warning that heralded the Prologue: do not be blinded by this power! Put this diabolical machine to work in the *service* of geometry, not in place of it!

—Here endeth the lesson. Amen!—

While the applications of Cartan's Forms are legion, *our* primary application (in the final chapter of Act V) will naturally be to Differential Geometry, enabling us to *reprove by symbolic means results that were proved geometrically in the first four Acts.*

First, however, *we shall fully develop Cartan's ideas in their own right, providing a self-contained introduction to Forms that is completely independent of the first four Acts.* We have done this because Forms find fruitful applications across diverse areas of mathematics, physics, and other disciplines. In short, *our aim is to make Forms accessible to the widest possible range of readers, even if their primary interest is* not *Differential Geometry.*

32.2 Definition of a 1-Form

The starting point, and the fundamental building block of Cartan's diabolically ingenious machine, is the concept of a ***1-form***. Let us waste no time in stating its definition:

A **1-form** *is a linear, real-valued function of a single vector input.*

NOTES: The "1-" refers to the *single* vector input; later we will meet ***2-forms*** that take *two* vectors as their input, ***3-forms*** that take *three* vectors as their input, and so on. A 1-form is therefore an especially simple kind of tensor. Older works instead call this concept a ***covariant vector***, or a ***covector***.[2] We will denote 1-forms by lowercase bold Greek letters, while continuing to denote vectors by lowercase bold Roman letters.

More explicitly, if k_1 and k_2 are arbitrary constants, and $\mathbf{v}_1$ and $\mathbf{v}_2$ are arbitrary vectors, then

$$\boxed{\boldsymbol{\omega} \text{ is a 1-form} \qquad \Longleftrightarrow \qquad \boldsymbol{\omega}(k_1\mathbf{v}_1 + k_2\mathbf{v}_2) = k_1\,\boldsymbol{\omega}(\mathbf{v}_1) + k_2\,\boldsymbol{\omega}(\mathbf{v}_2).} \tag{32.1}$$

When checking that a particular $\boldsymbol{\omega}$ is a 1-form, it can be conceptually helpful to break down this single requirement into two simpler ones:

$$\boxed{\boldsymbol{\omega}(\mathbf{v}_1 + \mathbf{v}_2) = \boldsymbol{\omega}(\mathbf{v}_1) + \boldsymbol{\omega}(\mathbf{v}_2),} \tag{32.2}$$

and

$$\boxed{\boldsymbol{\omega}(k\mathbf{v}) = k\,\boldsymbol{\omega}(\mathbf{v}).} \tag{32.3}$$

Check for yourself that (32.2) and (32.3) together imply (32.1), and vice versa: the two definitions of linearity are equivalent.

[2] This terminology is explained in Exercise 3.

A 1-form is *defined* by its action on vectors: two 1-forms are equal if and only if they have the same effect on all vectors. Given two distinct 1-forms, $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$, there is therefore a natural way to define their sum, $(\boldsymbol{\omega}+\boldsymbol{\varphi})$, by its action on a general vector $\mathbf{v}$:

$$(\boldsymbol{\omega}+\boldsymbol{\varphi})(\mathbf{v}) \equiv \boldsymbol{\omega}(\mathbf{v})+\boldsymbol{\varphi}(\mathbf{v}),$$

and it is easy to check [exercise] using (32.2) and (32.3) that this sum is itself a 1-form. Likewise, we may multiply a 1-form $\boldsymbol{\omega}$ by a constant k to obtain a new 1-form, $[k\boldsymbol{\omega}]$, defined by

$$[k\boldsymbol{\omega}](\mathbf{v}) \equiv k[\boldsymbol{\omega}(\mathbf{v})].$$

Thus the set of 1-forms is closed under addition and multiplication by constants, and therefore constitutes what is called a ***vector space***. This vector space of 1-forms is said to be ***dual*** to the space of vectors upon which it acts. The reason for this terminology is that there is a symmetrical relationship between these two spaces: we can also think of the space of vectors as "dual" to the space of 1-forms.

To see this symmetry, let us *think of a vector* $\mathbf{v}$ *as a function that acts on 1-forms* $\boldsymbol{\omega}$, this action being defined by

$$\boxed{\mathbf{v}(\boldsymbol{\omega}) \equiv \boldsymbol{\omega}(\mathbf{v}).}$$

This symmetrical action of a vector and a 1-form upon each other is often called the ***contraction*** of one with the other, and is sometimes denoted[3] $\langle \boldsymbol{\omega}, \mathbf{v} \rangle$, to emphasize the equal footing of the two kinds of object.

It follows from this that a vector $\mathbf{v}$ is a *linear* function of 1-forms:

$$\begin{aligned} \mathbf{v}(\boldsymbol{\omega}+\boldsymbol{\varphi}) &= (\boldsymbol{\omega}+\boldsymbol{\varphi})(\mathbf{v}) \\ &= \boldsymbol{\omega}(\mathbf{v})+\boldsymbol{\varphi}(\mathbf{v}) \\ &= \mathbf{v}(\boldsymbol{\omega})+\mathbf{v}(\boldsymbol{\varphi}), \end{aligned}$$

and,

$$\mathbf{v}(k\,\boldsymbol{\omega}) = k\,\boldsymbol{\omega}(\mathbf{v}) = k\,\mathbf{v}(\boldsymbol{\omega}).$$

Just as T_p denotes the space of vectors at the point p, so T_p^* denotes the dual vector space of 1-forms at p. And just as a vector field assigns a vector $\mathbf{v}_p$ to each point p, so a ***field of 1-forms*** assigns a 1-form $\boldsymbol{\omega}_p$ to each point p.

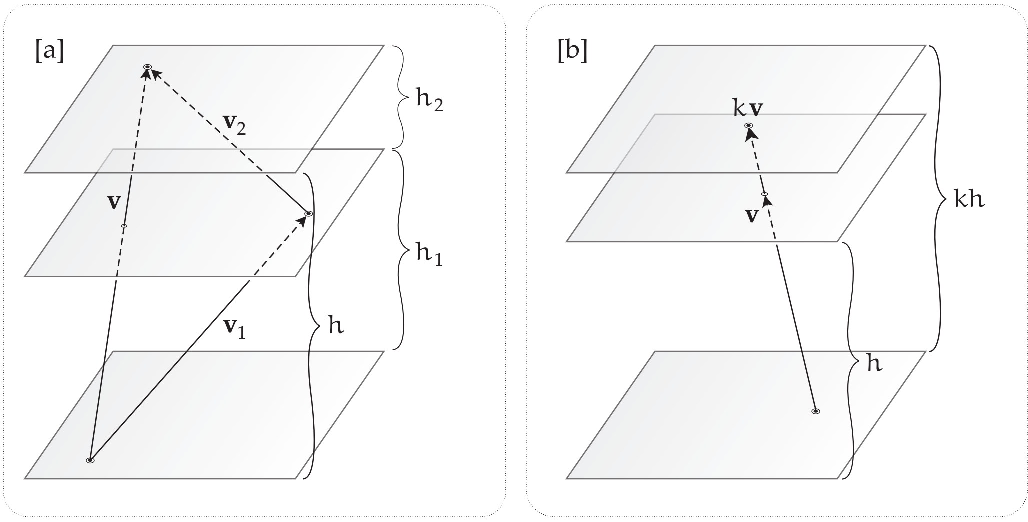
The next section justifies the introduction of 1-forms (albeit after the fact!) by revealing a secret: you have spent your adult life surrounded by 1-forms—you just didn't know it.

32.3 Examples of 1-Forms

32.3.1 Gravitational Work

Let F be the magnitude of the gravitational force exerted on a unit mass near the surface of the Earth. We shall make the approximation that F is constant if we only move the mass short distances. If we lift the mass through a vertical distance h, then the ***work*** done—i.e., the energy we must expend to make this happen—is $\omega = Fh$.

[3]The similarity to the Dirac bra–ket notation is not coincidental, as we shall see in a moment.



[32.2] **[a]** *When the mass is moved along* $\mathbf{v} = (\mathbf{v}_1 + \mathbf{v}_2)$, *the increase* h *in its height is simply the sum* $(h_1 + h_2)$ *of the separate increases in height resulting from moving along* $\mathbf{v}_1$ *and* $\mathbf{v}_2$, *separately.* **[b]** *Stretching the movement* $\mathbf{v}$ *by a factor of* k *also stretches the vertical movement by the same factor.*

Now suppose that we move the mass along the vector $\mathbf{v}$, and define

$$\boxed{\boldsymbol{\omega}(\mathbf{v}) \equiv \text{work done moving the mass along } \mathbf{v}.}$$

Then $\boldsymbol{\omega}$ is a 1-form!

To verify this, we must check (32.2) and (32.3). As illustrated in [32.2a], when the mass is moved along $\mathbf{v} = (\mathbf{v}_1 + \mathbf{v}_2)$, the increase h in its height is simply the sum $(h_1 + h_2)$ of the separate increases in height resulting from moving along $\mathbf{v}_1$ and $\mathbf{v}_2$, separately. Likewise, as illustrated in [32.2b], stretching the movement $\mathbf{v}$ by a factor of k also stretches the vertical movement (and hence the work) by the same factor. Done!

32.3.2 Visualizing the Gravitational Work 1-Form

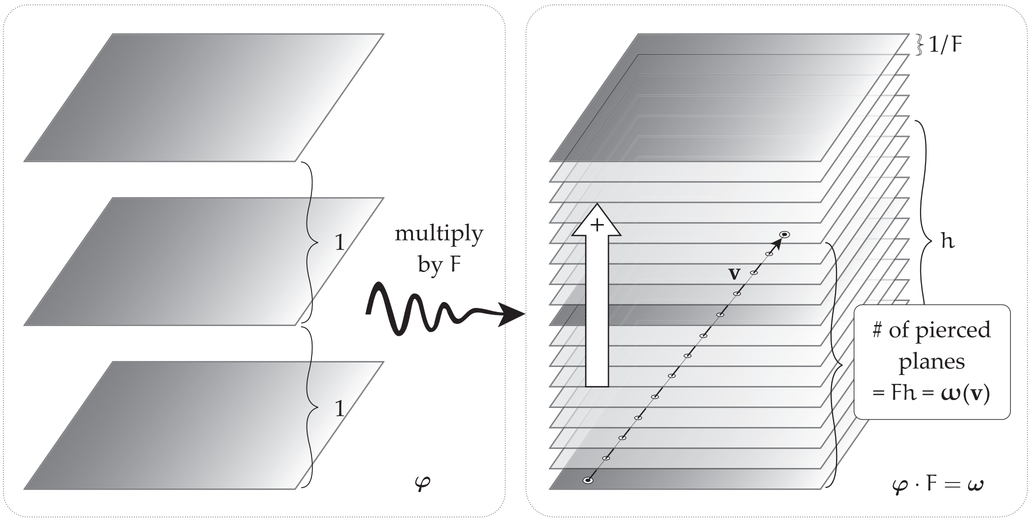
This example naturally suggests a means of *visualizing* a 1-form. Imagine a family of equally spaced surfaces of constant height. These are spheres centred at the centre of the Earth, but, in the local vicinity of the mass, these spheres will look like a stack of equally spaced horizontal planes. See the right-hand side of [32.3].

As we move the mass along $\mathbf{v}$, the change in height, and hence the work, will be proportional to the number of planes pierced by $\mathbf{v}$. By adjusting the spacing of the surfaces/planes to be $(1/F)$, we can therefore make

$$\boxed{\boldsymbol{\omega}(\mathbf{v}) = \text{number of surfaces of } \boldsymbol{\omega} \text{ pierced by } \mathbf{v}.}$$

Of course, in general, the tip of $\mathbf{v}$ will lie between two planes, but we can easily imagine filling the space between two planes with interpolating planes, so that if, for example, the tip of $\mathbf{v}$ lies half way between plane 17 and plane 18, we can say that it pierces 17.5 planes.

Thus the 1-form $\boldsymbol{\omega}$ itself may be visualized as this stack of equally spaced surfaces of constant gravitational potential energy, and *the greater the force* F *represented by* $\boldsymbol{\omega}$, *the more densely packed the surfaces of* $\boldsymbol{\omega}$.



[32.3] Visualization of a 1-Form. *On the right is the stack representing the gravitational work 1-form $\boldsymbol{\omega}$, its planes being $(1/F)$ apart, and its direction upward. Thus $\boldsymbol{\omega}(\mathbf{v})$ can be visualized as the (signed) number of planes pierced by $\mathbf{v}$. On the left is the* unit-*spaced 1-form $\boldsymbol{\varphi}$. When this is multiplied by F, the density of its planes is increased by F,* shrinking *their spacing to $(1/F)$, yielding $\boldsymbol{\varphi} \cdot F = \boldsymbol{\omega}$.*

To drive home this point, [32.3] shows a 1-form $\boldsymbol{\varphi}$ (on the left) with *unit* spacing being multiplied by F to yield our work 1-form, $\boldsymbol{\varphi} \cdot F = \boldsymbol{\omega}$, on the right.

Note that this visualization of the contraction of a vector with a 1-form has the virtue of very clearly putting the two kinds of objects on an equal footing. We call the set of equally spaced parallel planes a ***stack***. Note that in order to complete interpretation in [32.3], we must attach a ***direction*** (variously known as a *sense* or an *orientation*) to the stack, upward in our case, as illustrated. If the direction of $\mathbf{v}$ *agrees* with the direction of $\boldsymbol{\omega}$, then the piercings are counted as *positive*, but if the vector goes *against* the direction of the stack, then the piercings are counted as *negative*.

This visualization can be applied more generally to an arbitrary 1-form. At the point p, the representative surface $\mathcal{S}$ of $\boldsymbol{\omega}$ is defined by the fact that vectors emanating from p tangent to $\mathcal{S}$ satisfy $\boldsymbol{\omega}(\mathbf{v}) = 0$, corresponding to zero surfaces being pierced. To put this into the language of Linear Algebra, the set of tangent vectors to $\mathcal{S}$ satisfy $\boldsymbol{\omega}(\mathbf{v}) = 0$, so they comprise the ***kernel*** of $\boldsymbol{\omega}$.

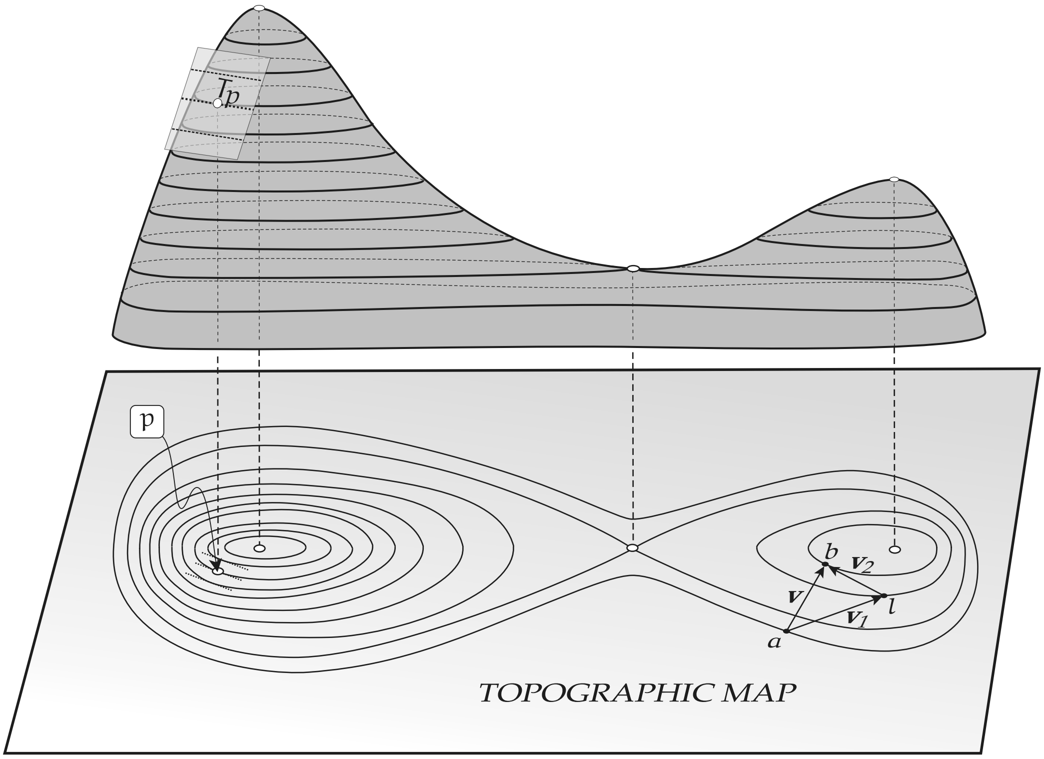
In general, it is only possible to represent a 1-form field in this way as a stack of planes at a point—typically the small pieces of planes at *neighbouring* points will fail to mesh together to create a smooth family of space-filling surfaces, called a ***foliation***, although they *do* happen to in our gravitational example. For a discussion of this point, see Penrose (2005, §12.3), Bachman (2012, §5.7), or Dray (2015, §13.8).

In $\mathbb{R}^3$, the set of vectors satisfying $\boldsymbol{\omega}(\mathbf{v}) = 0$ spans a 2-dimensional plane, but in n-dimensional space $\boldsymbol{\omega}$ will instead be represented by $(n-1)$-dimensional spaces.

32.3.3 *Topographic Maps and the Gradient 1-Form*

If you are planning a long hike through rugged, mountainous terrain, you will be well-advised to plan your route using a *topographic map*, so as to avoid having to climb too hard, or to scamble down a slope that is dangerously steep. Recall from the example, [19.6], on page 203 that such a map shows contours of constant height $h(x, y)$: as the point $p = (x, y)$ in the map travels along such a contour, $h(p) = \text{constant}$.

Consider the two hills and the pass depicted in [32.4], and its topographic map below. If you hike along one of these contours, the ground is flat: you neither climb nor descend. If, on the other hand, you set off in the direction orthogonal to the contour, you will rise (or descend) as rapidly as possible: *the direction orthogonal to a contour is the direction of steepest ascent/descent*. The contours



[32.4] Topographic Map and the Gradient 1-Form. *As we zoom in on a small region surrounding a point* p *in the topographic map of a surface, the contours look straighter and straighter, and more and more evenly spaced. Ultimately they become the representatives of the topographic map of the tangent plane* T_p *to the surface at the point directly above* p*, the latter map being a depiction of the* gradient 1-form $\boldsymbol{\zeta}$.

are drawn with equal increments δh in the height h, so if you walk in the direction orthogonal to a contour, *the more crowded together the contours in the map, the steeper the terrain*.

In greater detail, suppose we draw lots of contours with a fixed, small difference δh between one and the next. In the immediate vicinity of the point p, the density of the contours is roughly constant; let's call this roughly constant horizontal distance between neighbouring contours in the map δr. Then the slope of terrain in that small region is

$$\text{slope} \asymp \frac{\delta h}{\delta r}.$$

In other words, *the steepness of the terrain is inversely proportional to the gap between neighbouring contours in the map*.

This is very reminiscent of our visualization of the gravitational work 1-form: the stronger the gravitational force, the denser the packing of the surfaces representing the corresponding work 1-form. With this in mind, let us attempt to explicitly identify a steepness 1-form associated with our topographic map.

In the map plane, let $\mathbf{v} = \overrightarrow{ab}$ be a vector emanating from a, with its tip at b, as illustrated. Suppose we *attempt* to define a steepness 1-form as follows:

$$\begin{aligned} \boldsymbol{\eta}(\mathbf{v}) &\equiv (\text{number of contours pierced by } \mathbf{v}) \cdot (\delta h) \\ &= h(b) - h(a) \\ &= \text{change in elevation from tail to tip of } \mathbf{v}. \end{aligned}$$

Does this satisfy the linearity conditions, (32.2) and (32.3)?

Let $\mathbf{v}_1 = \overrightarrow{al}$, $\mathbf{v}_2 = \overrightarrow{lb}$, and $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, as illustrated; then, at first glance, it might appear that all goes well:

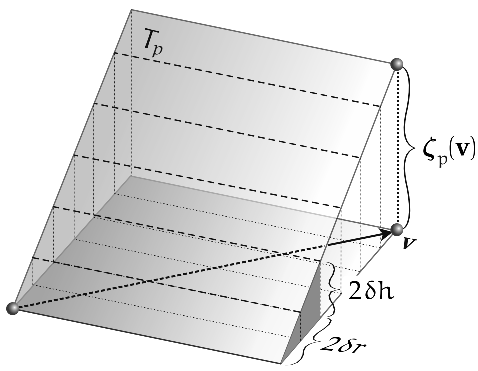
$$\boldsymbol{\eta}(\mathbf{v}_1) + \boldsymbol{\eta}(\mathbf{v}_2) = [h(l) - h(a)] + [h(b) - h(l)] = h(b) - h(a) = \boldsymbol{\eta}(\mathbf{v}).$$

However, unlike a true 1-form, the value of $\boldsymbol{\eta}(\mathbf{v})$ depends on *where* the vector $\mathbf{v}$ is drawn: if we imagine $\mathbf{v}_2$ drawn at a instead of at l, then clearly (32.2) is *no longer* satisfied! Thus our putative 1-form $\boldsymbol{\eta}$ at p has *not* actually passed the first test. Furthermore, even if we restrict ourselves to vectors emanating from a, it fails the second test, too, for it is clear that if we double the length of the illustrated vector $\mathbf{v}$ then we end up on the far side of the hill, and the altitude actually *decreases*, instead of doubling, as it should. Thus $\boldsymbol{\eta}$ does *not* satisfy (32.3).

Nevertheless, this discussion points the way to a genuine 1-form. The smaller the region surrounding a point p that we examine, the more precisely uniform the pattern of contours: *they look like evenly spaced, parallel lines.* Thus if we only apply $\boldsymbol{\eta}$ to very short (ultimately vanishing) vectors emanating from p, then $\boldsymbol{\eta}$ *does* (ultimately) satisfy the linearity requirements!

Now let us try to extend this local action of $\boldsymbol{\eta}$ at p from tiny vectors emanating from p to large vectors, drawn anywhere in the map plane. All we need to do is to take this local pattern of parallel, equally spaced lines in the immediate vicinity of p (all parallel to the tangent line to the contour through p) and extend it out to cover the entire plane. If we look thoughtfully at [32.4], we see that what we have just constructed is in fact *the topographic map of the tangent plane* T_p *to the surface at the point directly above* p!

In order to obtain a genuine 1-form field $\boldsymbol{\zeta}$, *we need only apply our old definition of* $\boldsymbol{\eta}$ *to this topographic map of* T_p. See [32.5]. We repeat that this topographic map consists of *equally spaced parallel lines*, all running in the same direction as the original contour through p, and *the gap between one line and the next is* $\delta r \asymp (\delta h)/(\textit{slope of } T_p)$: the bigger the slope of T_p, the more densely packed the parallel lines in its topographic map. The direction orthogonal to these lines is the direction of steepest ascent up the surface at the point p.



[32.5] The Gradient 1-Form $\boldsymbol{\zeta}_p$ *of the height of a surface, when applied to a vector* $\mathbf{v}$ *in the plane, yields the change in the height of* T_p *as we travel along* $\mathbf{v}$*, where* T_p *is the tangent plane to the surface at the point directly above the point* p.

The ***value*** $\boldsymbol{\zeta}_p$ of the 1-form field $\boldsymbol{\zeta}$ at p is the 1-form that is defined by the tangent plane T_p *at* the point of the surface directly above p. As illustrated in [32.5],

$$\begin{aligned}\boldsymbol{\zeta}_p(\mathbf{v}) &\equiv (\textit{number of lines pierced by } \mathbf{v}) \cdot (\delta h) \\ &= \textit{change in height of } T_p \textit{ from tail to tip of } \mathbf{v}.\end{aligned} \tag{32.4}$$

As p varies, the direction and spacing of the original contour lines around p both vary, reflecting the fact that the tangent plane T_p to the surface is varying, and so we have a *different* topographic map of T_p at each point, and hence a different value $\boldsymbol{\zeta}_p$ of the 1-form field $\boldsymbol{\zeta}$.

If we apply $\boldsymbol{\zeta}_p$ to very short (ultimately vanishing) vectors emanating from p, then its effect is ultimately equal to that of $\boldsymbol{\eta}$; that is, $\boldsymbol{\zeta}_p(\mathbf{v}) \asymp \boldsymbol{\eta}_p(\mathbf{v})$. But as the vector gets bigger, $\boldsymbol{\eta}(\mathbf{v})$ tells us the nonlinear variation in the height of the actual terrain/surface, while the 1-form $\boldsymbol{\zeta}_p(\mathbf{v})$ tells us the perfectly linear variation in the height of the tangent plane to the surface at that point.

This 1-form field $\boldsymbol{\zeta}$ is called the ***gradient*** of h. You are probably asking yourself if this is connected with the familiar, eponymous concept from Vector Calculus. Indeed it is! The explanation of the connection is the subject of Section 32.6. This idea of the gradient 1-form field $\boldsymbol{\zeta}$ is extremely important, and we shall see how it can be generalized to functions h that depend on more than two variables, which therefore cannot be so simply visualized as the height $z = h(x, y)$ of a surface about the (x, y)-plane.

32.3.4 Row Vectors

Consider a specific 2-dimensional row vector, such as $\boldsymbol{\omega} = [-3, 2]$. We can define the action of $\boldsymbol{\omega}$ on 2-dimensional column vectors $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ by means of standard matrix multiplication on the left:

$$\boldsymbol{\omega}(\mathbf{v}) = [-3, 2] \begin{bmatrix} x \\ y \end{bmatrix} = -3x + 2y.$$

We leave it to you to check that this does indeed satisfy (32.2) and (32.3). Thus, viewed in this way, *row vectors are 1-forms*.

We can relate this example to the gravitational work example. To do so, let us choose Cartesian coordinates (x, y, z) with the z-axis vertical, in the direction of the Earth's gravitational field. Then we can represent the gravitational work as the 1-form (i.e., row vector),

$$\boldsymbol{\omega} = [0, 0, F], \qquad \text{because then} \qquad \boldsymbol{\omega}(\mathbf{v}) = [0, 0, F] \begin{bmatrix} x \\ y \\ h \end{bmatrix} = Fh = \text{work}.$$

32.3.5 Dirac's Bras

NOTE: This example assumes that you are already familiar with quantum mechanics; feel free to skip it if you are not.

Although we have defined 1-forms to be *real*-valued, and shall continue to do so—with one exception—for the remainder of the book, it is often useful to broaden the definition to include *complex*-valued 1-forms. Indeed, this generalization arises naturally in quantum mechanics, and it is *essential* in that context.

If we think of ***Dirac kets*** $|\mathbf{v}\rangle$ (i.e., the quantum states) as the "vectors," then *a **Dirac bra** $\langle\boldsymbol{\omega}|$ is a 1-form*, provided we define the contraction of a 1-form and a vector to be the standard (complex) bra–ket inner product:

$$\boldsymbol{\omega}(\mathbf{v}) \equiv \langle\boldsymbol{\omega}|\mathbf{v}\rangle.$$

To confirm that a bra $\langle\boldsymbol{\omega}|$ is a 1-form, we simply check the definition, (32.1):

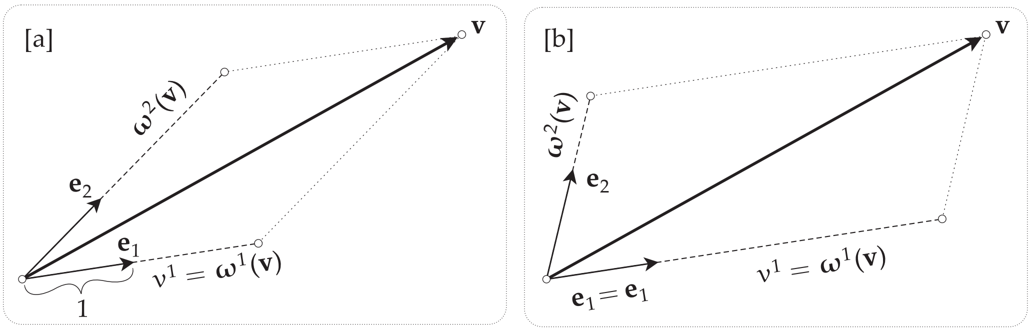
$$\begin{aligned} \boldsymbol{\omega}(k_1\mathbf{v}_1 + k_2\mathbf{v}_2) &= \langle\boldsymbol{\omega}|k_1\mathbf{v}_1 + k_2\mathbf{v}_2\rangle \\ &= k_1 \langle\boldsymbol{\omega}|\mathbf{v}_1\rangle + k_2 \langle\boldsymbol{\omega}|\mathbf{v}_2\rangle \\ &= k_1\, \boldsymbol{\omega}(\mathbf{v}_1) + k_2\, \boldsymbol{\omega}(\mathbf{v}_2). \end{aligned}$$

32.4 Basis 1-Forms

At a point p in an n-manifold, suppose that we have chosen a basis $\{\mathbf{e}_j\}$ for T_p, so that a general vector can be written, using the Einstein summation convention, as

$$\mathbf{v} = v^j\, \mathbf{e}_j.$$

We shall *not* assume that this basis is orthonormal. Given this basis, there exists a natural way to associate with it a ***dual basis*** $\{\boldsymbol{\omega}^i\}$ for the space T_p^* of 1-forms at p:



[32.6] **[a]** *The first basis 1-form $\boldsymbol{\omega}^1$ picks out the 1st component v^1 of* $\mathbf{v}$*; likewise, the second basis 1-form $\boldsymbol{\omega}^2$ picks out the second component.* **[b]** *With the same* $\mathbf{e}_1$ *as before, merely changing* $\mathbf{e}_2$ *causes* both *of the basis 1-forms to change.*

$$\boxed{\boldsymbol{\omega}^i \text{ picks out the } i\text{-th component of } \mathbf{v} \qquad \Longleftrightarrow \qquad \boldsymbol{\omega}^i(\mathbf{v}) = v^i.} \tag{32.5}$$

These $\{\boldsymbol{\omega}^i\}$ are indeed 1-forms, which can be confirmed by checking (32.2) and (32.3): the i-th component of the sum of two vectors is the sum of their i-th components, and the i-th component of $k\mathbf{v}$ is just kv^i. Done!

Figure [32.6a] illustrates this definition in the case of a 2-manifold (i.e., a surface) for which the tangent plane $T_p = \mathbb{R}^2$. Here, as illustrated, we have chosen a basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ of nonorthogonal but ***unit vectors***, and

$$\mathbf{v} = v^1\,\mathbf{e}_1 + v^2\,\mathbf{e}_2 = \boldsymbol{\omega}^1(\mathbf{v})\,\mathbf{e}_1 + \boldsymbol{\omega}^2(\mathbf{v})\,\mathbf{e}_2.$$

Thus, in this case, $\boldsymbol{\omega}^1(\mathbf{v})$ is simply the length of the projection (parallel to $\mathbf{e}_2$) of $\mathbf{v}$ onto $\mathbf{e}_1$, and correspondingly for $\boldsymbol{\omega}^2(\mathbf{v})$, as illustrated.

Let us nip the following common misconception in the bud. While the *set* of basis 1-forms $\{\boldsymbol{\omega}^1, \boldsymbol{\omega}^2\}$ is dual to the *set* of basis vectors $\{\mathbf{e}_1, \mathbf{e}_2\}$, it is simply wrong to think that $\boldsymbol{\omega}^1$ is dual to $\mathbf{e}_1$, and that $\boldsymbol{\omega}^2$ is dual to $\mathbf{e}_2$.

Figure [32.6b] reveals the error of this thinking. Here we have illustrated the same vector $\mathbf{v}$ broken down into components relative to a new basis $\{\widetilde{\mathbf{e}}_1, \widetilde{\mathbf{e}}_2\}$, in which $\widetilde{\mathbf{e}}_1 = \mathbf{e}_1$, but $\widetilde{\mathbf{e}}_2 \neq \mathbf{e}_2$. As you can see, even though we have only changed one basis vector, *both* basis 1-forms have changed! Likewise, in the general n-dimensional case, changing a single basis vector can change the *entire* dual basis of 1-forms.

We should note that our preferred definition (32.6) of the dual basis is not the conventional one, though it is the definition used by Schutz (1980), and it is mathematically equivalent to the standard one, as we now explain.

First recall the definition of the

$$\textbf{\textit{Kronecker delta}}: \qquad \delta^i_j \equiv \left\{\begin{array}{ll} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{array}\right\}.$$

It is easy to prove that our definition (32.6) is equivalent to the following, standard definition:

$$\text{Equivalent definitions of a } \textbf{\textit{basis}}\ \{\boldsymbol{\omega}^i\}\text{:} \qquad \boxed{\boldsymbol{\omega}^i(\mathbf{e}_j) = \delta^i_j \qquad \Longleftrightarrow \qquad \boldsymbol{\omega}^i(\mathbf{v}) = v^i.}$$

The implication from right to left is immediate, and the implication from left to right goes like this:

$$\boldsymbol{\omega}^i(\mathbf{v}) = \boldsymbol{\omega}^i(v^j\,\mathbf{e}_j) = v^j\,\boldsymbol{\omega}^i(\mathbf{e}_j) = v^j\,\delta^i_j = v^i.$$

32.5 Components of a 1-Form

The ever-vigilant reader will have noticed that we have not actually proved that the set $\{\boldsymbol{\omega}^i\}$ is a *basis* for the 1-forms. Rather than proving directly that they are linearly independent, we shall follow the more elegant and enlightening approach employed by Schutz (1980, §2.20).

Let the general 1-form $\boldsymbol{\varphi}$ act on the general vector $\mathbf{v}$:

$$\begin{aligned}\boldsymbol{\varphi}(\mathbf{v}) &= \boldsymbol{\varphi}(v^j\,\mathbf{e}_j)\\ &= v^j\,\boldsymbol{\varphi}(\mathbf{e}_j)\\ &= \boldsymbol{\omega}^j(\mathbf{v})\,\boldsymbol{\varphi}(\mathbf{e}_j).\end{aligned}$$

We now define the

$$\boxed{\textbf{\textit{Components}}\ \varphi_j\ \text{of}\ \boldsymbol{\varphi}:\qquad \varphi_j \equiv \boldsymbol{\varphi}(\mathbf{e}_j).}$$

Thus,

$$\boldsymbol{\varphi}(\mathbf{v}) = \varphi_j\,\boldsymbol{\omega}^j(\mathbf{v}).$$

But since a 1-form is *defined* by its action on a general vector, we may "abstract away" the vector $\mathbf{v}$ on both sides of this equation, and thereby equate the 1-forms *themselves*, and, in so doing, decompose the arbitrary 1-form $\boldsymbol{\varphi}$ into its unique components in the 1-form basis $\{\boldsymbol{\omega}^i\}$ dual to the vector basis $\{\mathbf{e}_j\}$:

$$\boxed{\boxed{\boldsymbol{\varphi} = \varphi_j\,\boldsymbol{\omega}^j = \boldsymbol{\varphi}(\mathbf{e}_j)\,\boldsymbol{\omega}^j.}} \tag{32.6}$$

32.6 The Gradient as a 1-Form: $\mathbf{d}f$

32.6.1 Review of the Gradient as a Vector: $\boldsymbol{\nabla} f$

Recall from Vector Calculus that in $\mathbb{R}^2$ the ***gradient*** of a function f is defined to be the *vector*

$$\boldsymbol{\nabla} f \equiv \begin{bmatrix}\partial_x f\\ \partial_y f\end{bmatrix}.$$

The *significance* of this vector is that,

$$\boxed{\text{$\boldsymbol{\nabla} f$ points in the direction of most rapid increase of f, and its magnitude $|\boldsymbol{\nabla} f|$ equals that maximum rate of increase of f with distance as we move in that direction.}} \tag{32.7}$$

This interpretation springs from a more primitive fact, which we now derive. Let $\{\mathbf{e}_1, \mathbf{e}_2\}$ be the standard orthonormal basis along the $(x^1, x^2) = (x, y)$ axes, and let δf be the small change in f resulting from moving along the short (ultimately vanishing) vector

$$\mathbf{v} = \delta x^1\,\mathbf{e}_1 + \delta x^2\,\mathbf{e}_2 = \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}.$$

By definition, $\partial_x f$ is the rate of change of f as we move in the x-direction, so the change in f resulting from moving δx is ultimately equal to $(\partial_x f)\,\delta x$; and likewise for y. Therefore,

$$\boxed{\delta f \asymp (\partial_x f)\,\delta x + (\partial_y f)\,\delta y = (\nabla f) \cdot \mathbf{v}.} \tag{32.8}$$

In the classical notation of the eighteenth century, this was instead expressed as

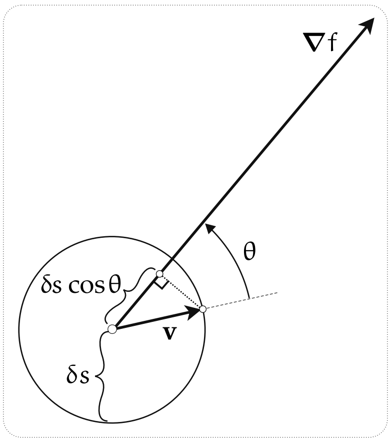
$$df = (\partial_x f)\,dx + (\partial_y f)\,dy, \tag{32.9}$$

in which df, dx, and dy were understood to be *infinitesimals*—a concept we have eschewed in this work. *However*, as we shall see shortly, this ancient formula takes on a new life, with a precise and rigorous meaning, when viewed through the modern prism of 1-forms.

In order to derive (32.8), consider [32.7], which interprets (32.9) geometrically. If we keep the length $|\mathbf{v}| \equiv \delta s$ fixed, then as $\mathbf{v}$ rotates around the circle,

$$\delta f \asymp |\nabla f|\,(\delta s\ \cos\theta),$$

where θ is the angle that $\mathbf{v}$ makes with the gradient ∇f, so that $(\delta s\ \cos\theta)$ is simply the projection of $\mathbf{v}$ onto the direction of ∇f, as illustated.



[32.7] The Gradient Vector ∇f. *The small change δf resulting from moving along the short vector $\mathbf{v}$ is ultimately equal to $|\nabla f|\,(\delta s\ \cos\theta)$.*

Let $\hat{\mathbf{v}} = \mathbf{v}/|\mathbf{v}|$ denote the ***unit vector*** in the direction of $\mathbf{v}$. Then the above result can instead be expressed as

$$\boxed{\left|\nabla_{\hat{\mathbf{v}}} f\right| = \frac{df}{ds} = |\nabla f|\ \cos\theta.}$$

This immediately confirms the interpretation of ∇f given in (32.8). It also shows that if we move in the direction *orthogonal* to ∇f, then the rate of change of f vanishes. This is the direction of the tangent to the curve defined by $f = constant$.

In $\mathbb{R}^3$ the analysis is exactly as before, but now there is an entire *plane* of directions orthogonal to ∇f: this is the tangent plane to the *surface*, $f = constant$.

32.6.2 The Gradient as a 1-Form: $\mathbf{d}f$

Although we are accustomed to thinking of the gradient as a vector, it is in fact more naturally a 1-form. Let us immediately define this 1-form, and then proceed to make sense of it.

The ***gradient 1-form*** $\mathbf{d}f$ of a function f is defined by its action on vectors:

$$\boxed{\mathbf{d}f(\mathbf{v}) \equiv \nabla_{\mathbf{v}} f.} \tag{32.10}$$

The bold $\mathbf{d}$ operator is called the ***exterior derivative***; it will play a central role in what follows.

Let us check that $\mathbf{d}f$ really is a 1-form, by verifying (32.2) and (32.3):

$$(\mathbf{d}f)\,(\mathbf{v}_1+\mathbf{v}_2)=\boldsymbol{\nabla}_{\mathbf{v_1}+\mathbf{v_2}}f=\boldsymbol{\nabla}_{\mathbf{v}_1}f+\boldsymbol{\nabla}_{\mathbf{v}_2}f=\mathbf{d}f(\mathbf{v}_1)+\mathbf{d}f(\mathbf{v}_2)$$

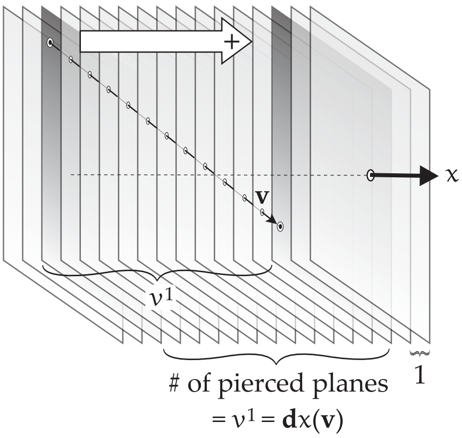
and,

$$(\mathbf{d}f)\,(k\mathbf{v})=\boldsymbol{\nabla}_{k\mathbf{v}}f=k\,\boldsymbol{\nabla}_{\mathbf{v}}f=k\,\mathbf{d}f(\mathbf{v}).$$

Observe that since $\boldsymbol{\nabla}_{\mathbf{v}}$ obeys the Leibniz Rule (aka Product Rule), so does the exterior derivative:

$$\boxed{\mathbf{d}(fg)=f\,\mathbf{d}g+g\,\mathbf{d}f.} \tag{32.11}$$

32.6.3 The Cartesian 1-Form Basis: $\{\mathbf{d}x^j\}$



[32.8] The Cartesian Basis 1-form $\mathbf{d}x$. *When applied to a vector $\mathbf{v}$, the 1-form $\mathbf{d}x$ picks out the x-component:* $\mathbf{d}x(\mathbf{v})=v^1$.

What is the meaning of 1-forms $\mathbf{d}x=\mathbf{d}x^1$ and $\mathbf{d}y=\mathbf{d}x^2$ in $\mathbb{R}^2$? To answer this question, we must determine their effect on a general vector, $\mathbf{v}=\begin{bmatrix}v^1\\ v^2\end{bmatrix}$. The definition (32.11) tells us that

$$\begin{aligned}(\mathbf{d}x)\begin{bmatrix}v^1\\ v^2\end{bmatrix} &= (\mathbf{d}x)\mathbf{v}\\ &= \boldsymbol{\nabla}_{\mathbf{v}}\,x\\ &= \boldsymbol{\nabla}_{v^1\mathbf{e}_1+v^2\mathbf{e}_2}\,x\\ &= (v^1\,\boldsymbol{\nabla}_{\mathbf{e}_1}+v^2\boldsymbol{\nabla}_{\mathbf{e}_2})\,x\\ &= (v^1\,\partial_x+v^2\,\partial_y)\,x\\ &= v^1.\end{aligned}$$

In other words, $\mathbf{d}x$ *is the 1-form that picks out the* x*-component of a vector*. Likewise, we see that $(\mathbf{d}y)\mathbf{v}=v^2$. But according to the definition (32.6), this means that $\{\mathbf{d}x,\mathbf{d}y\}$ is the *dual basis* of $\{\mathbf{e}_1,\mathbf{e}_2\}$. Equivalently, $(\mathbf{d}x^i)\mathbf{e}_j=\delta^i_j$.

Geometrically, $\mathbf{d}x$ may be pictured as the family of unit-spaced lines perpendicular to the x-axis, and $(\mathbf{d}x)\mathbf{v}$ is the number of these lines that $\mathbf{v}$ pierces. The above reasoning applies equally well to $\mathbb{R}^3$, in which case $\mathbf{d}x$ is the family of unit-spaced *planes* perpendicular to the x-axis, as illustrated in [32.8].

Clearly, all of this immediately generalizes to any number of dimensions. Let $\{\mathbf{e}_i\}$ be an orthonormal basis for $\mathbb{R}^n$, with Cartesian coordinates $\{x^j\}$, so that $\mathbf{v}=v^i\,\mathbf{e}_i$ is a general vector. Then $\mathbf{d}x^j$ picks out the j-th component of the vector, i.e., $(\mathbf{d}x^j)\,\mathbf{v}=v^j$. In particular,

$$\boxed{(\mathbf{d}x^i)\mathbf{e}_j=\delta^i_j.} \tag{32.12}$$

We call this the ***Cartesian basis***:

$$\boxed{\{\mathbf{d}x^j\}=\{\boldsymbol{\omega}^j\}\text{is the Cartesian basis of 1-forms dual to }\{\mathbf{e}_j\}.}$$

It follows from (32.7) that a general 1-form φ can be decomposed into components in this dual basis as

$$\boxed{\varphi = \varphi_j\, \boldsymbol{\omega}^j = \varphi(\mathbf{e}_j)\, \mathbf{d}x^j.} \tag{32.13}$$

32.6.4 The 1-Form Interpretation of $df = (\partial_x f)\, dx + (\partial_y f)\, dy$

Taking $\varphi = \mathbf{d}f$ in (32.14), we have come full circle, decomposing the gradient 1-form $\mathbf{d}f$ of a general function f into its Cartesian basis 1-form components, as follows:

$$\boxed{\mathbf{d}f = \left[(\mathbf{d}f)\, \mathbf{e}_j\right] \mathbf{d}x^j = \left[\partial_{x^j} f\right] \mathbf{d}x^j.} \tag{32.14}$$

On a formal level, this is *identical* to the classical formula (32.10), but now it has a precise, rigorous meaning that does not call upon infinitesimals. And yet it is very directly and intuitively connected to the geometrically meaningful *ultimate equality*, (32.9), illustrated in [32.7].

To see this, once again think of $\mathbf{v}$ as a short, ultimately vanishing vector,

$$\mathbf{v} = \delta x^1\, \mathbf{e}_1 + \delta x^2\, \mathbf{e}_2 = \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}.$$

Then,

$$\mathbf{d}f(\mathbf{v}) = \left\{ \left[\partial_x f\right] \mathbf{d}x + \partial_y f\right] \mathbf{d}y \right\} \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = (\partial_x f)\, \delta x + (\partial_y f)\, \delta y \asymp \delta f,$$

which is *exactly* (32.9). Thus the 1-form $\mathbf{d}f$ gives us the best of both worlds!

The above analysis sought to connect the new 1-form $\mathbf{d}f$ to your prior knowledge of the vector gradient $\boldsymbol{\nabla} f$. However, if you look back at our discussion of topographic maps, you will see that we had *already* arrived there at this new 1-form by geometrical reasoning alone.

Indeed, if you look at our definition (32.5) (illustrated in [32.5]) of the 1-form $\boldsymbol{\zeta}$ associated with a topographic map, which describes how the height of the tangent plane varies as we move off in an arbitrary direction, you will see that it is indeed the gradient of the height function h:

$$\boxed{\boldsymbol{\zeta} = \mathbf{d}h.}$$

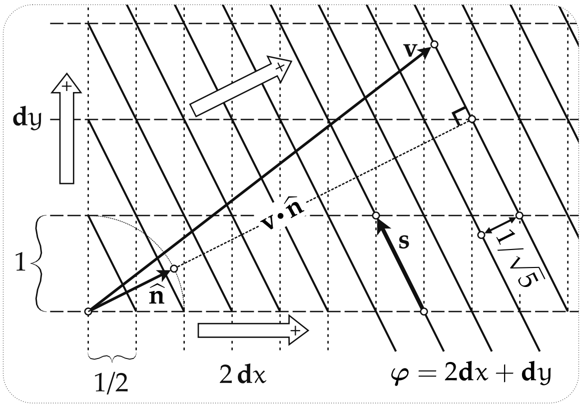
In particular, if $\mathbf{v}$ is the direction at p of the contour of constant height passing through that point, then $\boldsymbol{\zeta}(\mathbf{v}) = \mathbf{d}h(\mathbf{v}) = 0$, as it should.

32.7 Adding 1-Forms Geometrically

We already know the geometric meaning of multiplying a 1-form by a constant k: it *compresses* the stack by k. But what does *addition* of 1-forms mean?

Figure [32.9] superimposes the stacks representing the 1-forms $2\mathbf{d}x$ and $\mathbf{d}y$, then joins the resulting intersection points to create a new stack. *This* is the geometrical construction of their sum, $\varphi = 2\mathbf{d}x + \mathbf{d}y$!

If $\widetilde{\varphi}$ denotes the 1-form corresponding to the constructed stack, then we must prove that $\widetilde{\varphi}=\varphi$.



[32.9] Geometric Addition of 1-Forms. *To add the 1-forms* $2\,\mathbf{d}x$ *and* $\mathbf{d}y$*, superimpose their stacks, then join their intersection points to create the stack of* $\varphi=2\mathbf{d}x+\mathbf{d}y$.

We will offer *three* proofs, not because the result is especially important, but simply as valuable practice with thinking about and manipulating 1-forms.

First note [exercise] that the unit vector normal to $\widetilde{\varphi}$ is

$$\widehat{\mathbf{n}}=\frac{1}{\sqrt{5}}\begin{bmatrix}2\\1\end{bmatrix}.$$

Also note that this serves to specify the *direction* of the stack, which [exercise] must have positive components along both of the stacks that are being added.

If $\mathbf{v}=\begin{bmatrix}a\\b\end{bmatrix}$ denotes a general vector, then its projection onto the direction of $\widehat{\mathbf{n}}$ is $\mathbf{v}\boldsymbol{\cdot}\widehat{\mathbf{n}}$, as illustrated. But [exercise] the spacing of $\widetilde{\varphi}$ is $1/\sqrt{5}$, and so the number of lines of the stack $\widetilde{\varphi}$ pierced by $\mathbf{v}$ is

$$\widetilde{\varphi}(\mathbf{v})=\text{number of lines pierced}=\frac{\mathbf{v}\boldsymbol{\cdot}\widehat{\mathbf{n}}}{(1/\sqrt{5})}=2a+b=(2\mathbf{d}x+\mathbf{d}y)\begin{bmatrix}a\\b\end{bmatrix}=\varphi(\mathbf{v}).$$

Since $\mathbf{v}$ is a general vector, we may abstract it away, confirming our claim that $\widetilde{\varphi}=\varphi$.

Here is a second proof, which is simpler, but less direct. We know that for any 1-form, such as our φ, the direction(s) $\mathbf{s}$ lying within the stack are the kernel of the 1-form: $\mathbf{s}$ does not pierce any lines (or planes, in $\mathbb{R}^3$) of the stack, so

$$0=\varphi(\mathbf{s})=(2\mathbf{d}x+\mathbf{d}y)\begin{bmatrix}p\\q\end{bmatrix}=2p+q \qquad\Longrightarrow\qquad \mathbf{s}\propto\begin{bmatrix}-1\\2\end{bmatrix},$$

but, as illustrated, this is *also* the direction of the $\widetilde{\varphi}$ stack, connecting grid points. Thus the two stacks are parallel and can only differ as regards their spacing. Put algebraically, the two 1-forms must be *proportional*: $\widetilde{\varphi}=K\,\varphi$, for some constant K. But if we apply this equation to a specific vector for which the number of pierced lines is evident—try $\begin{bmatrix}0\\1\end{bmatrix}$—then we instantly deduce that $K=1$. Done!

The third proof is perhaps the simplest of all. In [32.9], think of $\mathbf{v}=\begin{bmatrix}x\\y\end{bmatrix}$ as being a *position vector*, emanating from a fixed origin. What is the equation of a line in the φ-stack? Well, as the tip of $\mathbf{v}$ moves along such a line, $\mathbf{v}$ pierces a *fixed* number of stack lines, so

$$\textit{constant}=\varphi(\mathbf{v})=(2\mathbf{d}x+\mathbf{d}y)\begin{bmatrix}x\\y\end{bmatrix}=2x+y \qquad\Longrightarrow\qquad y=-2x+\textit{constant}.$$

But these lines, with slope -2, are *parallel* to the $\widetilde{\varphi}$-stack. The rest of the proof is the same as before.

It is easy [exercise] to generalize these arguments to confirm that this geometrical construction works for $p\,\mathbf{d}x + q\,\mathbf{d}y$, yielding a stack with

$$\text{normal vector} \begin{bmatrix} p \\ q \end{bmatrix} \text{ and spacing } \frac{1}{\sqrt{p^2+q^2}}.$$

Indeed, in $\mathbb{R}^3$, we may [exercise] construct $p\,\mathbf{d}x + q\,\mathbf{d}y + r\,\mathbf{d}z$ by superimposing the three orthogonal stacks of planes, constructing a new stack of planes passing through the intersection points, with

$$\text{normal vector} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \text{ and spacing } \frac{1}{\sqrt{p^2+q^2+r^2}}.$$

Chapter 33

Tensors

33.1 Definition of a Tensor: Valence

In our discussion of the Riemann tensor, we tentatively defined a tensor as a *multilinear function of multiple vector inputs*. This was the best we could do at the time, because we lacked the concept of a 1-form. But a fully general tensor is, in fact, a multilinear function of vectors *and* 1-forms, and the ***valence*** tells us how many of each:

> *A* **tensor H of valence** $\left\{ \begin{smallmatrix} \mathfrak{f} \\ \mathfrak{v} \end{smallmatrix} \right\}$ *at a point* $\mathfrak{p}$ *is a multilinear, real-valued function of* $\mathfrak{f}$ *1-forms, and* $\mathfrak{v}$ *vectors, such that the value of* $\mathbf{H}$ *at* $\mathfrak{p}$ *only depends on the values of the 1-forms and vectors* at $\mathfrak{p}$.

Thus a 1-form is a tensor of valence $\left\{ \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right\}$, because it has one "slot" into which a single vector can be fed, the output being the contraction of that vector with the 1-form. Likewise, a vector is a tensor of valence $\left\{ \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right\}$, because it has one slot into which a single 1-form can be fed, the output being the contraction of that 1-form with the vector.

In the general case, we may divide the input slots of $\mathbf{H}$ into two groups: $\mathfrak{f}$ slots to input the $\mathfrak{f}$ 1-forms, $\varphi_1, \ldots, \varphi_{\mathfrak{f}}$, and a second group of $\mathfrak{v}$ slots to input the $\mathfrak{v}$ vectors, $\mathbf{v}_1, \ldots, \mathbf{v}_{\mathfrak{v}}$. Using $\|$ to denote the boundary between these two groups, we have

$$\mathbf{H}(\varphi_1, \ldots, \varphi_{\mathfrak{f}} \,\|\, \mathbf{v}_1, \ldots, \mathbf{v}_{\mathfrak{v}}).$$

In general, the *order* in which we feed 1-forms and vectors into these slots *matters*: if we swap a pair of 1-forms or a pair of vectors in these slots, the output will in general be completely unrelated to the original value.

Let us re-examine the Riemann tensor $\mathbf{R}$ from this new, general point of view. In (29.8) we defined this as a *vector*-valued, multilinear function of three vector inputs: $\mathbf{w}$ is the vector that is parallel-transported around the parallelogram with edges defined by $\mathbf{u}$ and $\mathbf{v}$, which we now imagine to be very short, and the output is the vector holonomy $\delta\mathbf{w}$:

$$\delta\mathbf{w} = \mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w}) = \left\{ [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]} \right\} \mathbf{w}.$$

In order to make this into a tensor in the new sense, it must output a real *number*, instead of the vector $\delta\mathbf{w}$. To extract a real number from $\delta\mathbf{w}$, we must contract it with a 1-form. For this reason, in place of our original definition, the *standard* definition of the Riemann tensor is instead a tensor of valence $\left\{ \begin{smallmatrix} 1 \\ 3 \end{smallmatrix} \right\}$:

$$\mathbf{R}(\varphi \,\|\, \mathbf{u}, \mathbf{v}, \mathbf{w}) \equiv \langle \varphi, \delta\mathbf{w} \rangle = \varphi(\delta\mathbf{w}).$$

NOTE ON NOTATION: We shall reserve uppercase bold Roman letters for tensors with multiple inputs, such as the Riemann tensor, **R**, the energy-momentum tensor, **T**, and the Einstein tensor, **G**. On the other hand, some such higher-valence tensors are traditionally represented by *lower-case* bold Roman letters, the most notable example being the metric tensor, **g**, which must not be confused with a vector!

33.2 Example: Linear Algebra

It should already be clear that an enormous range of mathematical and physical objects fall into our new, generalized category of tensors. To drive home this point, consider the vast and vastly important subject of *Linear Algebra*. From our new vantage point, this is "merely" the study of tensors $\mathbf{L}(\boldsymbol{\varphi} \,\|\, \mathbf{v})$ of valence $\left\{ \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right\}$!

To see why this is, consider $\mathbf{L}(\ldots \,\|\, \mathbf{v})$, in which we have left the first slot empty. In order to yield a number, we must complete this by re-inserting the 1-form φ. Therefore, $\mathrm{L}(\mathbf{v}) \equiv \mathbf{L}(\ldots \,\|\, \mathbf{v})$ must be a *vector*, ready to be contracted with a 1-form φ to yield a number. This is perfectly analogous to our original definition of the Riemann tensor: there we had a vector-valued function of three vectors; now we have a vector-valued function L of a single vector, which only yields a number once we contract it with a 1-form.

But, by the definition of a tensor, $\mathbf{L}(\varphi \,\|\, \mathbf{v})$ is a *linear* function of both φ and $\mathbf{v}$. Thus,

$$\begin{aligned} \mathrm{L}(k_1\mathbf{v}_1 + k_2\mathbf{v}_2) &= \mathbf{L}(\ldots \,\|\, k_1\mathbf{v}_1 + k_2\mathbf{v}_2) \\ &= k_1\,\mathbf{L}(\ldots \,\|\, \mathbf{v}_1) + k_2\,\mathbf{L}(\ldots \,\|\, \mathbf{v}_2) \\ &= k_1\,\mathrm{L}(\mathbf{v}_1) + k_2\,\mathrm{L}(\mathbf{v}_2). \end{aligned}$$

Therefore L is a *linear transformation*, sending vectors to vectors—the fundamental object of study in Linear Algebra! Of course, by the same token, if we leave the *vector* slot open, we may instead interpret $\mathbf{L}(\varphi \,\|\, \ldots)$ as a linear, 1-form-valued function of 1-forms.

33.3 New Tensors from Old

33.3.1 Addition

Clearly it makes no sense to try to add tensors with different valences, which require entirely different kinds of inputs. However, if the two tensors have the *same* valence, the definition of addition is obvious. For example, if **H** and **J** are both of valence $\left\{ \begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \right\}$, then

$$(\mathbf{H}+\mathbf{J})(\varphi, \psi \,\|\, \mathbf{v}) \equiv \mathbf{H}(\varphi, \psi \,\|\, \mathbf{v}) + \mathbf{J}(\varphi, \psi \,\|\, \mathbf{v}).$$

It is easy to check [exercise] that $(\mathbf{H}+\mathbf{J})$ is indeed another *tensor* of valence $\left\{ \begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \right\}$.

33.3.2 Multiplication: The Tensor Product

Given two 1-forms, φ and ψ, each of which acts on a single vector, it is natural to define their ***tensor product***, acting on *two* vectors, and therefore of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$, as follows:

$$\boxed{(\boldsymbol{\varphi} \otimes \boldsymbol{\psi})(\mathbf{v}, \mathbf{w}) \equiv \boldsymbol{\varphi}(\mathbf{v})\, \boldsymbol{\psi}(\mathbf{w}).} \tag{33.1}$$

Note that *order matters*: $\boldsymbol{\varphi} \otimes \boldsymbol{\psi} \neq \boldsymbol{\psi} \otimes \boldsymbol{\varphi}$.

TERMINOLOGY: The tensor product is also commonly called the ***direct product*** or the ***outer product***.

The tensor product of higher valence tensors goes the same way. For example, we may multiply a tensor $\mathbf{J}(\boldsymbol{\varphi}, \boldsymbol{\psi} \,\|\, \mathbf{u})$ of valence $\left\{ \begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \right\}$ and a tensor $\mathbf{T}(\mathbf{v}, \mathbf{w})$ of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$ to obtain the tensor $\mathbf{J} \otimes \mathbf{T}$ of valence $\left\{ \begin{smallmatrix} 2 \\ 3 \end{smallmatrix} \right\}$:

$$(\mathbf{J} \otimes \mathbf{T})(\boldsymbol{\varphi}, \boldsymbol{\psi} \,\|\, \mathbf{u}, \mathbf{v}, \mathbf{w}) \equiv \mathbf{J}(\boldsymbol{\varphi}, \boldsymbol{\psi} \,\|\, \mathbf{u}) \cdot \mathbf{T}(\mathbf{v}, \mathbf{w}).$$

Note that under tensor multiplication, valences add like vectors: in our example, $\left\{ \begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \right\} + \left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\} = \left\{ \begin{smallmatrix} 2 \\ 3 \end{smallmatrix} \right\}$. Since multiplication by a scalar does not change the valence, the valence of a scalar must be taken to be $\left\{ \begin{smallmatrix} 0 \\ 0 \end{smallmatrix} \right\}$.

33.4 Components

As usual, let $\{\mathbf{e}_i\}$ be an orthonormal vector basis, and let $\{\mathbf{d}x^j\}$ be the dual Cartesian basis of 1-forms. Just as we did to obtain the components of vectors and 1-forms, we obtain the ***components*** of a more general tensor by filling its slots with basis 1-forms and vectors. For example, the components of $\mathbf{T}(\mathbf{v}, \mathbf{w})$ are

$$T_{ij} = \mathbf{T}(\mathbf{e}_i, \mathbf{e}_j)$$

We can decompose the full tensor $\mathbf{T}$ into tensor components, as follows:

$$\begin{aligned}
\mathbf{T}(\mathbf{v}, \mathbf{w}) &= \mathbf{T}(v^i \mathbf{e}_i, w^j \mathbf{e}_j) \\
&= \mathbf{T}(\mathbf{e}_i, \mathbf{e}_j)\, v^i w^j \\
&= T_{ij}\, v^i w^j \\
&= T_{ij} \left[\mathbf{d}x^i(\mathbf{v})\right] \left[\mathbf{d}x^j(\mathbf{w})\right] \\
&= T_{ij}\, (\mathbf{d}x^i \otimes \mathbf{d}x^j)(\mathbf{v}, \mathbf{w})
\end{aligned}$$

But since both $\mathbf{v}$ and $\mathbf{w}$ are general vectors, we may abstract them away, to find that the tensor itself is

$$\boxed{\mathbf{T} = T_{ij}\, (\mathbf{d}x^i \otimes \mathbf{d}x^j).} \tag{33.2}$$

Thus we see that

The set of tensors $\{\mathbf{d}x^i \otimes \mathbf{d}x^j\}$ forms a **basis** *for tensors of valence* $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$.

Note that if we apply this idea to (33.1), then we find that

$$\boldsymbol{\varphi} \otimes \boldsymbol{\psi} = \varphi_i \psi_j\, (\mathbf{d}x^i \otimes \mathbf{d}x^j).$$

However, such tensors are special: a general $\mathbf{T}$ *cannot* be factorized in this way.

We can see this from the fact that while $\mathbf{T}$ has n^2 components T_{ij} in n dimensions, $(\boldsymbol{\varphi}\otimes\boldsymbol{\psi})$ is uniquely determined by just $2n$ numbers: the n components φ_i of $\boldsymbol{\varphi}$, and the n components ψ_i of $\boldsymbol{\psi}$.

In exactly the same way, we can take the tensor product of two vectors:

$$(\mathbf{v}\otimes\mathbf{w})(\boldsymbol{\varphi},\boldsymbol{\psi}) = \mathbf{v}(\boldsymbol{\varphi})\,\mathbf{w}(\boldsymbol{\psi}) = v^i\,w^j\,(\mathbf{e}_i\otimes\mathbf{e}_j)(\boldsymbol{\varphi},\boldsymbol{\psi}).$$

And, likewise, a general tensor $\mathbf{K}(\boldsymbol{\varphi},\boldsymbol{\psi})$ of valence $\left\{\begin{smallmatrix}2\\0\end{smallmatrix}\right\}$ can be decomposed into basis tensors as

$$\boxed{\mathbf{K} = K^{ij}\,(\mathbf{e}_i\otimes\mathbf{e}_j).}$$

Clearly the above technique can be applied to tensors of *arbitrary* valence $\left\{\begin{smallmatrix}\mathfrak{f}\\\mathfrak{v}\end{smallmatrix}\right\}$, decomposing them into components using the basis tensors,

$$(\mathbf{e}_{i_1}\otimes\mathbf{e}_{i_1}\otimes\cdots\otimes\mathbf{e}_{i_\mathfrak{f}})\otimes(\mathbf{d}x^{j_1}\otimes\mathbf{d}x^{j_2}\otimes\cdots\otimes\mathbf{d}x^{j_\mathfrak{v}}).$$

Note that the components of a tensor of valence $\left\{\begin{smallmatrix}\mathfrak{f}\\\mathfrak{v}\end{smallmatrix}\right\}$ have $\mathfrak{f}$ *upper* indices, and $\mathfrak{v}$ *lower* indices.[1]

33.5 Relation of the Metric Tensor to the Classical Line Element

Earlier we discussed the fact that the modern gradient 1-form $\mathbf{d}f$ is closely connected to the classical, infinitesimal-based formula (33.9) for the differential df. Similarly, let us now consider how the modern metric tensor is related to the classical, infinitesimal-based ***line element*** ds of Gauss (1827).

Gauss considered infinitesimal changes du and dv in the coordinates of a point on the surface, resulting in an infinitesimal line element ds within the surface. Then, in Gauss's original[2] notation,

$$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2.$$

In place of infinitesimals, we instead employ ultimate equalities, expressing this as

$$\delta s^2 \asymp E\,\delta u^2 + 2F\,\delta u\,\delta v + G\,\delta v^2.$$

Now let

$$x^1 = u,\quad x^2 = v \qquad\text{and}\qquad g_{11} = E,\quad g_{12} = g_{21} = F,\quad g_{22} = G.$$

Then, using (33.2), the metric tensor can be expressed as

$$\mathbf{g} = g_{ij}\,(\mathbf{d}x^i\otimes\mathbf{d}x^j) = E\,(\mathbf{d}u\otimes\mathbf{d}u) + F\,(\mathbf{d}u\otimes\mathbf{d}v) + F\,(\mathbf{d}v\otimes\mathbf{d}u) + G\,(\mathbf{d}v\otimes\mathbf{d}v),$$

which looks tantalizingly similar to Gauss's formula for ds^2.

[1] In older literature, upper indices were called ***contravariant***, and lower indices were called ***covariant***.

[2] Actually, as we remarked earlier, Gauss (1827) used p and q instead of u and v.

To make the connection explicit, let $\delta\mathbf{r} = \begin{bmatrix} \delta u \\ \delta v \end{bmatrix}$ be a short (ultimately vanishing) vector in the coordinate map plane, resulting in a small movement δs within the surface. Then, since

$$\mathbf{d}u \begin{bmatrix} \delta u \\ \delta v \end{bmatrix} = \delta u, \qquad \text{and} \qquad \mathbf{d}v \begin{bmatrix} \delta u \\ \delta v \end{bmatrix} = \delta v,$$

we do indeed recover the line-element formula from the metric tensor:

$$\begin{aligned} \delta s^2 &\asymp \mathbf{g}(\delta\mathbf{r}, \delta\mathbf{r}) \\ &= [\mathsf{E}\,(\mathbf{d}u \otimes \mathbf{d}u) + \mathsf{F}\,(\mathbf{d}u \otimes \mathbf{d}v) + \mathsf{F}\,(\mathbf{d}v \otimes \mathbf{d}u) + \mathsf{G}\,(\mathbf{d}v \otimes \mathbf{d}v)] \left(\begin{bmatrix} \delta u \\ \delta v \end{bmatrix}, \begin{bmatrix} \delta u \\ \delta v \end{bmatrix} \right) \\ &= \mathsf{E}\,\delta u^2 + \mathsf{F}\,\delta u\,\delta v + \mathsf{F}\,\delta v\,\delta u + \mathsf{G}\,\delta v^2. \end{aligned}$$

33.6 Example: Linear Algebra (Again)

What is the connection between the familiar matrices of a traditional Linear Algebra course, and the seemingly abstract tensor view of the subject, as the study of tensors of valence $\left\{ \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right\}$? This section uses the concepts of components, and of tensor bases, to explain the connection.

First, our tensor $\mathbf{L}(\boldsymbol{\varphi} \,\|\, \mathbf{v})$ has components

$$L^i{}_j = \mathbf{L}(\mathbf{d}x^i \,\|\, \mathbf{e}_j).$$

In order to keep things as simple as possible and as concrete as possible, we restrict ourselves to two dimensions.

If the vector basis is represented by the column vectors

$$\{\mathbf{e}_1, \mathbf{e}_2\} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\},$$

then, as discussed in Section 32.3.4, the dual 1-form basis corresponds to row vectors:

$$\left\{ \mathbf{d}x^1, \mathbf{d}x^2 \right\} = \left\{ \begin{bmatrix} 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \end{bmatrix} \right\}.$$

Thus,

$$\mathbf{d}x^1(\mathbf{v}) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} = v^1,$$

as it should. Likewise for $\mathbf{d}x^2$.

It follows that $\mathbf{L}$ is indeed represented by a famiiar 2×2 matrix, whose entries are the components of the tensor:

$$\begin{aligned} \mathbf{L} &= L^i{}_j\, \mathbf{e}_i \otimes \mathbf{d}x^j \\ &= L^1{}_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + L^1{}_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} + L^2{}_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + L^2{}_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} \end{aligned}$$

$$= L^1{}_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + L^1{}_2 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + L^2{}_1 \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + L^2{}_2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} L^1{}_1 & L^1{}_2 \\ L^2{}_1 & L^2{}_2 \end{bmatrix}.$$

In Linear Algebra, this matrix describes a linear transformation, by means of standard matrix multiplication:

$$\begin{bmatrix} v^1 \\ v^2 \end{bmatrix} \to \begin{bmatrix} L^1{}_1 & L^1{}_2 \\ L^2{}_1 & L^2{}_2 \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} = \begin{bmatrix} L^1{}_1 v^1 + L^1{}_2 v^2 \\ L^2{}_1 v^1 + L^2{}_2 v^2 \end{bmatrix} = \begin{bmatrix} L^1{}_j v^j \\ L^2{}_j v^j \end{bmatrix}.$$

In other words,

$$v^i \longrightarrow L^i{}_j v^j \qquad \Longleftrightarrow \qquad \mathbf{v} = v^i \mathbf{e}_i \longrightarrow (L^i{}_j v^j)\, \mathbf{e}_i.$$

Expressed in this form, we can now see how this is equivalent to our discussion in Section 33.2, in which we used the tensor to define the linear transformation $L(\mathbf{v})$:

$$\mathbf{v} \longrightarrow L(\mathbf{v}) = \mathbf{L}(\ldots \,\|\, \mathbf{v}) = L^i{}_j\, (\mathbf{e}_i \otimes \mathbf{d}x^j)\,(\ldots, \mathbf{v}) = L^i{}_j\, \mathbf{e}_i(\ldots)\, \mathbf{d}x^j(\mathbf{v}) = (L^i{}_j v^j)\, \mathbf{e}_i(\ldots).$$

33.7 Contraction

Consider the contraction of a 1-form and a vector from the point of view of components:

$$\boldsymbol{\varphi}(\mathbf{v}) = (\varphi_i\, \mathbf{d}x^i)(v^j\, \mathbf{e}_j) = (\varphi_i\, v^j)\, \mathbf{d}x^i(\mathbf{e}_j) = (\varphi_i\, v^j)\, \delta^i_j = \varphi_j\, v^j.$$

As we know, this contraction is a *geometrical* operation, and the answer is independent of the specific components of $\boldsymbol{\varphi}$ and $\mathbf{v}$, *despite* the fact that these components *do* depend on the choice of the dual bases $\{\mathbf{e}_j\}$ and $\{\mathbf{d}x^i\}$.

Next, consider the components $L^i{}_j$ of the tensor in the previous section. We now define, analogously, the ***contraction*** of this tensor to be

$$L^j{}_j = \mathbf{L}(\mathbf{d}x^j \,\|\, \mathbf{e}_j) = L^1{}_1 + L^2{}_2 = \mathrm{Tr} \begin{bmatrix} L^1{}_1 & L^1{}_2 \\ L^2{}_1 & L^2{}_2 \end{bmatrix}.$$

As we know from Linear Algebra, this trace is *also* basis-independent.[3]

We will prove a greatly generalized version of this basis-independence result—using purely tensor methods, and without restricting ourselves to two dimensions—but first we must explain what *contraction* means in the general case.

The idea of contraction can be applied to *any* tensor whose inputs include at least one 1-form and at least one vector: *we sum over one upper index and one lower index*. We have already encountered one very important and natural example of such a contraction, namely, the contraction of the Riemann tensor that yields the Ricci tensor:

$$R_{mij}{}^m = R_{ij}.$$

Note that this contraction process changes the valence, in this example, from $\left\{ {1 \atop 3} \right\}$ to $\left\{ {0 \atop 2} \right\}$. In general, the input slots of the new, contracted tensor accept one fewer 1-form and one fewer vector, thereby eliminating one upper and one lower index.

[3] There is a *geometrical* reason for this, although it is hard to find in Linear Algebra texts. See Arnol'd (1973, §16.3).

In fact contraction has a still broader meaning in the context of tensor products. Suppose we form $\mathbf{A} \otimes \mathbf{B}$, and then perform a summation over an upper index of $\mathbf{A}$ and a lower index of $\mathbf{B}$. Let us prove that the result of this contraction is another *tensor*, independent of the specific $\{\mathbf{e}_j\}$ and $\{\mathbf{d}x^i\}$ that gave rise to the components being summed.

Following the example of Schutz (1980, §2.25), we shall communicate the idea of the proof by means of a concrete example. Let $\mathbf{A}$ be of valence $\left\{ \begin{smallmatrix} 2 \\ 0 \end{smallmatrix} \right\}$, and let $\mathbf{B}$ be of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$. Let us prove that the contraction

$$A^{ij} B_{jk} \equiv C^i{}_k$$

yields the components of a new *tensor* $\mathbf{C}$ of valence $\left\{ \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right\}$, such that

$$\mathbf{C}(\boldsymbol{\varphi} \| \mathbf{v}) = C^i{}_k \varphi_i v^k.$$

To prove this, first observe that

$$\begin{aligned} C^i{}_k \varphi_i v^k &= A^{ij} B_{jk} \varphi_i v^k \\ &= [\varphi_i \, \mathbf{A}(\mathbf{d}x^i, \mathbf{d}x^j)] \, [\mathbf{B}(\mathbf{e}_j, \mathbf{e}_k) \, v^k] \\ &= \mathbf{A}(\varphi_i \, \mathbf{d}x^i, \mathbf{d}x^j) \, \mathbf{B}(\mathbf{e}_j, v^k \, \mathbf{e}_k) \\ &= \mathbf{A}(\boldsymbol{\varphi}, \mathbf{d}x^j) \, \mathbf{B}(\mathbf{e}_j, \mathbf{v}). \end{aligned}$$

But the $\mathbf{B}(\mathbf{e}_j, \mathbf{v})$ are numbers, and therefore the linearity of $\mathbf{A}$ in the second slot implies that

$$C^i{}_k \varphi_i v^k = \mathbf{A}(\boldsymbol{\varphi}, \mathbf{B}(\mathbf{e}_j, \mathbf{v}) \, \mathbf{d}x^j).$$

But, for fixed $\mathbf{v}$,

$$\mathbf{B}(\mathbf{e}_j, \mathbf{v}) \, \mathbf{d}x^j = \mathbf{B}(\ldots, \mathbf{v})$$

is a 1-form, requiring the insertion of a vector to produce a number. Therefore,

$$\mathbf{C}(\boldsymbol{\varphi} \| \mathbf{v}) = \mathbf{A}(\boldsymbol{\varphi}, \mathbf{B}(\ldots, \mathbf{v}))$$

is indeed a tensor.

33.8 Changing Valence with the Metric Tensor

We already know that the metric tensor $\mathbf{g}$ of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$ is *the* fundamental structure of a manifold: it gives us the geodesics, parallel transport, and the curvature. But $\mathbf{g}$ also plays another crucial role: *it allows us to change the valence of a tensor.* The first step is to see how the metric allows us to associate a particular 1-form with a particular vector, and vice versa.

If we leave one slot of $\mathbf{g}$ open, and insert the vector $\mathbf{n}$ into the other, we obtain a unique 1-form $\boldsymbol{\nu}$ corresponding to $\mathbf{n}$:

$$\boxed{\text{vector } \mathbf{n} \longrightarrow \text{1-form } \boldsymbol{\nu}, \quad \text{where } \boldsymbol{\nu}(\mathbf{w}) \equiv \mathbf{g}(\mathbf{w}, \mathbf{n}).} \tag{33.3}$$

How are the *components* of the 1-form $\boldsymbol{\nu}$ related to the components of the original vector $\mathbf{n}$? We need only apply $\boldsymbol{\nu}$ to the basis vectors. But first ...

NOTATIONAL WARNING: *the universal convention is that the components of the 1-form corresponding to the vector* **n** *are denoted* n_i, violating our Greek/Roman dichotomy between 1-forms and vectors. Henceforth, the reader must therefore be more vigilant than before: if the index is up, then it belongs to a vector, and if the index is down, it belongs to a 1-form.

With this convention in place,

$$\boldsymbol{\nu} = n_i\,\mathbf{d}x^i.$$

Thus,

$$n_i = \boldsymbol{\nu}(\mathbf{e}_i) = \mathbf{g}(\mathbf{e}_i, \mathbf{n}) = \mathbf{g}(\mathbf{e}_i, n^j\,\mathbf{e}_j) = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j)n^j,$$

so,

$$\boxed{n_i = g_{ij}\,n^j.} \tag{33.4}$$

For example, if $\mathbf{n} = n^j\,\mathbf{e}_j$ is a vector in Minkowski spacetime, with metric (30.6), then [exercise] the corresponding 1-form is

$$\boldsymbol{\nu} = n_i\,\mathbf{d}x^i = n^0\,\mathbf{d}t - n^1\,\mathbf{d}x - n^2\,\mathbf{d}y - n^3\,\mathbf{d}z.$$

The metric's ability to transform a vector into a 1-form allows us to change the valence of any tensor. For example, consider the Riemann tensor $\mathbf{R}(\boldsymbol{\psi} \,\|\, \mathbf{u}, \mathbf{v}, \mathbf{w})$, of valence $\left\{ {1 \atop 3} \right\}$, with components $R_{ijk}{}^{m}$. Let us demonstrate how we may change this into a tensor of valence $\left\{ {0 \atop 4} \right\}$, by changing the single 1-form input into a vector input. By convention, this new tensor is *still* denoted **R**, and its components are written R_{ijkl}.

To evaluate the new tensor with four vector inputs, we simply replace the extra vector input **n** with its corresponding 1-form $\boldsymbol{\nu}$ in the *original* definition of the Riemann tensor:

$$\mathbf{R}(\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{n}) \equiv \mathbf{R}(\boldsymbol{\nu} \,\|\, \mathbf{u}, \mathbf{v}, \mathbf{w}).$$

In component form, this equation becomes

$$R_{ijkl}u^i v^j w^k n^l = R_{ijk}{}^{m}u^i v^j w^k n_m = R_{ijk}{}^{m}u^i v^j w^k (g_{ml}\,n^l).$$

Therefore,

$$\boxed{R_{ijkl} = R_{ijk}{}^{m}\,g_{ml}.}$$

This process is called (logically enough) ***index lowering***.

It is also possible to go in the opposite direction, changing a vector input into a 1-form input; this results in ***index raising***. In order to follow the same path a before, we need a mapping from 1-forms $\boldsymbol{\nu}$ to vectors **n**. What is needed is an analogue $\widetilde{\mathbf{g}}$ of the metric tensor, now of valence $\left\{ {2 \atop 0} \right\}$, taking two 1-forms as input.

The mapping is then obtained exactly as before, by inserting $\boldsymbol{\nu}$ into one slot, leaving the other slot open, resulting in a vector:

$$\boxed{\text{1-form } \boldsymbol{\nu} \longrightarrow \text{vector } \mathbf{n}, \qquad \text{where } \ \mathbf{n}(\boldsymbol{\varphi}) \equiv \widetilde{\mathbf{g}}(\boldsymbol{\varphi}, \boldsymbol{\nu}).} \tag{33.5}$$

In terms of components, we will then have,

$$\boxed{n^i = \widetilde{g}^{ik}\, n_k.} \tag{33.6}$$

And, by the same token, we can now raise any tensor index we wish, e.g.,

$$\boxed{R_{ijkl}\, \widetilde{g}^{km} = R_{ij\ \,l}^{\ \ m}}$$

The final step is to realize that $\widetilde{\mathbf{g}}$ is uniquely determined by the metric $\mathbf{g}$, as follows. Lowering and raising of indices must be *inverse* operations: if we lower an index and then raise it again, we should obtain the *same* tensor that we started with. Put differently, changing a vector into a 1-form with (33.3), then changing it back again with (33.5), should result in the *identity*. In other words, $\widetilde{\mathbf{g}}$ is the *inverse* of $\mathbf{g}$:

$$\mathbf{n} \xrightarrow{\mathbf{g}} \boldsymbol{\nu} \xrightarrow{\widetilde{\mathbf{g}}} \mathbf{n}.$$

Inserting (33.4) into (33.6), this implies that

$$n^i = \widetilde{g}^{ik}\, n_k = \widetilde{g}^{ik}\, g_{kj}\, n^j.$$

Therefore, the components of $\widetilde{\mathbf{g}}$ must be related to the components of the metric by

$$\boxed{\boxed{\widetilde{g}^{ik}\, g_{kj} = \delta^i_j.}} \tag{33.7}$$

Just as we always use the same symbol $\mathbf{R}$ to denote the Riemann tensor, regardless of the number of 1-forms and vectors it accepts as inputs, so it is traditional (if confusing) to simply write $\mathbf{g}$ instead of $\widetilde{\mathbf{g}}$, and to correspondingly write its components as g^{ik}. Thus, (33.7) is conventionally written, $g^{ik}\, g_{kj} = \delta^i_j$.

33.9 Symmetry and Antisymmetry

Recall the defining properties of an ***even function***, $f^+(x)$, and of an ***odd function***, $f^-(x)$:

$$f^+(-x) = +f^+(x) \qquad \text{and} \qquad f^-(-x) = -f^-(x).$$

We may instead describe this by saying that $f^+(x)$ is ***symmetric***, and $f^-(x)$ is ***antisymmetric***. Geometrically, the graph $y = f^+(x)$ has mirror symmetry in the y-axis, e.g., $y = x^2$ or $y = \cos x$, whereas $y = f^-(x)$ does the *opposite* at x and its mirror image $-x$, e.g., $y = x^3$ or $y = \sin x$.

Suppose we try to split an *arbitrary* function $f(x)$ (lacking any special symmetry) into the sum of a symmetric function and an antisymmetric function:

$$f(x) = f^+(x) + f^-(x).$$

If this is possible, then, by definition,

$$f(-x) = f^+(x) - f^-(x).$$

By adding and then subtracting the previous two equations, we deduce that the split *is* always possible, and that it is explicitly given by

$$f^+(x) = \left[\frac{f(x)+f(-x)}{2}\right] \qquad \text{and} \qquad f^-(x) = \left[\frac{f(x)-f(-x)}{2}\right].$$

For example, if we take $f(x) = e^x$ then this split naturally leads us to the hyperbolic functions: the symmetric part of e^x is $f^+(x) = \cosh x$, and the antisymmetric part is $f^-(x) = \sinh x$.

Now let us try to do something analogous with a general tensor $\mathbf{E}(\mathbf{v}, \mathbf{w})$ of valence $\left\{ {0 \atop 2} \right\}$. By analogy, let the defining properties of a ***symmetric tensor***, $\mathbf{E}^+$, and of an ***antisymmetric***[4] ***tensor***, $\mathbf{E}^-$, be

$$\mathbf{E}^+(\mathbf{w}, \mathbf{v}) = +\mathbf{E}^+(\mathbf{v}, \mathbf{w}) \qquad \text{and} \qquad \mathbf{E}^-(\mathbf{w}, \mathbf{v}) = -\mathbf{E}^-(\mathbf{v}, \mathbf{w}).$$

By following the same line of reasoning as before, we discover that it is always possible to split a general tensor of valence $\left\{ {0 \atop 2} \right\}$ into a symmetric part and an antisymmetric part:

$$\mathbf{E}(\mathbf{v}, \mathbf{w}) = \mathbf{E}^+(\mathbf{v}, \mathbf{w}) + \mathbf{E}^-(\mathbf{v}, \mathbf{w}), \tag{33.8}$$

where

$$\mathbf{E}^+(\mathbf{v}, \mathbf{w}) = \left[\frac{\mathbf{E}(\mathbf{v}, \mathbf{w}) + \mathbf{E}(\mathbf{w}, \mathbf{v})}{2}\right] \qquad \text{and} \qquad \mathbf{E}^-(\mathbf{v}, \mathbf{w}) = \left[\frac{\mathbf{E}(\mathbf{v}, \mathbf{w}) - \mathbf{E}(\mathbf{w}, \mathbf{v})}{2}\right].$$

If we insert basis vectors into the slots, we obtain the component forms of these equations. While we are at it, let us also introduce the following standard notation: ***round brackets denote symmetrization*** and ***square brackets denote antisymmetrization***:

$$\boxed{E_{(ij)} \equiv E^+_{ij} = \tfrac{1}{2}\left[E_{ij} + E_{ji}\right] \qquad \text{and} \qquad E_{[ij]} \equiv E^-_{ij} = \tfrac{1}{2}\left[E_{ij} - E_{ji}\right]} \tag{33.9}$$

Therefore,

$$\boxed{E_{ij} = E_{(ij)} + E_{[ij]}.}$$

[4] Also often called ***skew symmetric***.

Chapter 34

2-Forms

34.1 Definition of a 2-Form and of a p-Form

We have already encountered several vitally important tensors of valence $\left\{ \begin{matrix} 0 \\ 2 \end{matrix} \right\}$ that are *symmetric*: the metric tensor, the Ricci tensor, the energy–momentum tensor, and the Einstein tensor.

It may therefore come as a surprise to learn that the *secret power source* of Cartan's Forms, which will drive our drama to its conclusion, is *anti*symmetry!

Let us waste no time in introducing you to the second-born of Cartan's Forms, the first to appear after the 1-forms:

> *A* **2-form $\boldsymbol{\Psi}$** *is an antisymmetric tensor of valence* $\left\{ \begin{matrix} 0 \\ 2 \end{matrix} \right\}$. *That is,*
> $$\boldsymbol{\Psi}(\mathbf{v}, \mathbf{u}) = -\boldsymbol{\Psi}(\mathbf{u}, \mathbf{v}). \qquad (34.1)$$

As a matter of simple housekeeping, let us immediately state the general definition of a p-form, knowing full well that this will be essentially *meaningless* to you in the absence of concrete examples:

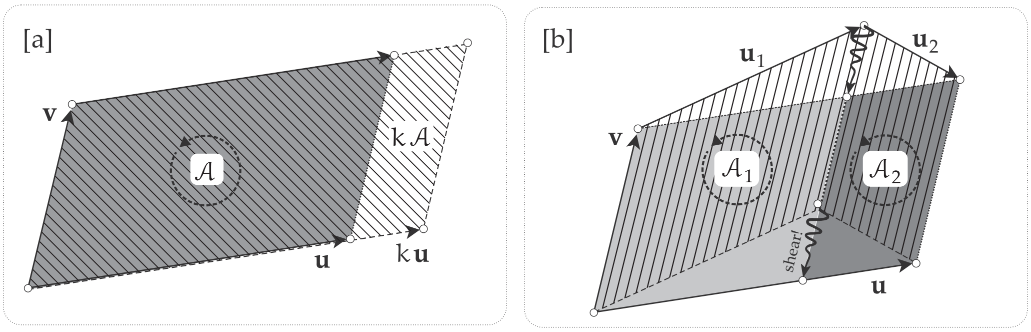
> *A* **p-form** *(also called a differential form of* **degree** p*) is a* **completely antisymmetric** *tensor of valence* $\left\{ \begin{matrix} 0 \\ \mathrm{p} \end{matrix} \right\}$, *meaning that swapping any two of the vector inputs reverses its sign.*

Naturally, if we take $\mathrm{p} = 2$ then we recover our definition (34.1) of a 2-form.

Our plan now is to build up an intuitive, *geometrical* understanding of p-forms by gradually increasing p, one step at a time. Act V opened with an entire chapter devoted to 1-forms, and this chapter is likewise completely devoted to understanding 2-forms. After that, we shall turn to 3-forms, ...

This sounds like a recipe for a book of infinite length! Fortunately, by the time we arrive at 3-forms, all the essential ideas needed to understand general p-forms will be in play. Also, as we shall explain in the next chapter, the highest p-form that can exist in an n-manifold is an n-form, and since spacetime has only four dimensions, 4-forms will therefore suffice for the purposes of this book. (NOTE: Higher values of p are, however, required in some important applications of Forms, such as the ***symplectic manifolds*** that naturally arise in Hamiltonian mechanics; see Arnol'd (1989, Ch. 8).)

But why devote precious time and energy to such *antisymmetric* tensors in the first place? After all, with the one notable exception of the Riemann tensor, *every* important tensor we have encountered thus far has been *symmetric*!



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

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

34.2 Example: The Area 2-Form

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

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

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

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

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

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

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

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

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

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

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

Then (32.2) requires that

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

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

Note that if $\boldsymbol{\Psi}$ is a general 2-form, then the antisymmetry under swapping of the two vector inputs implies that $\boldsymbol{\Psi}(\mathbf{u},\mathbf{u}) = -\boldsymbol{\Psi}(\mathbf{u},\mathbf{u})$, and so for arbitrary $\mathbf{u}$,

$$\boxed{\boldsymbol{\Psi}(\mathbf{u},\mathbf{u}) = 0.} \tag{34.2}$$

In the case of the area 2-form, this makes direct geometrical sense: if in $\mathcal{A}(\mathbf{u},\mathbf{v})$ we let $\mathbf{v}$ become $\mathbf{u}$, then the parallelogram collapses into a line segment of vanishing area, so $\mathcal{A}(\mathbf{u},\mathbf{u}) = 0$.

In fact, (34.2) is *equivalent* to our original antisymmetry-based definition of a 2-form, (34.1). To see this, we recycle an argument we previously employed in the context of the Riemann tensor. Let $\mathbf{x}$ and $\mathbf{y}$ be arbitrary vectors, and let $\mathbf{u} = \mathbf{x} + \mathbf{y}$. Then,

$$\begin{aligned} 0 &= \boldsymbol{\Psi}(\mathbf{u},\mathbf{u}) \\ &= \boldsymbol{\Psi}([\mathbf{x}+\mathbf{y}]\,,\,[\mathbf{x}+\mathbf{y}]) \\ &= \boldsymbol{\Psi}(\mathbf{x},\mathbf{x}) + \boldsymbol{\Psi}(\mathbf{x},\mathbf{y}) + \boldsymbol{\Psi}(\mathbf{y},\mathbf{x}) + \boldsymbol{\Psi}(\mathbf{y},\mathbf{y}) \\ &= 0 + \boldsymbol{\Psi}(\mathbf{x},\mathbf{y}) + \boldsymbol{\Psi}(\mathbf{y},\mathbf{x}) + 0. \end{aligned}$$

Therefore, $\boldsymbol{\Psi}(\mathbf{y},\mathbf{x}) = -\boldsymbol{\Psi}(\mathbf{x},\mathbf{y})$. Done.

34.3 The Wedge Product of Two 1-Forms

Recall from (33.8) that any tensor of valence $\left\{ {0 \atop 2} \right\}$ can be split into the sum of a symmetric tensor and an antisymmetric tensor. If we apply this procedure to the tensor product of two arbitrary 1-forms, then by swapping the 1-forms in the tensor products (instead of swapping the vectors) we find [exercise] that,

$$\boldsymbol{\varphi} \otimes \boldsymbol{\psi} = \frac{1}{2}\left[\boldsymbol{\varphi} \otimes \boldsymbol{\psi} + \boldsymbol{\psi} \otimes \boldsymbol{\varphi}\right] + \frac{1}{2}\left[\boldsymbol{\varphi} \otimes \boldsymbol{\psi} - \boldsymbol{\psi} \otimes \boldsymbol{\varphi}\right].$$

The antisymmetric part of this is, by definition, a *2-form* created by multiplying two 1-forms. This is a new kind of multiplication, called the ***wedge product***,[1] denoted $\wedge$:

$$\boxed{\boxed{\boldsymbol{\varphi} \wedge \boldsymbol{\psi} \equiv \boldsymbol{\varphi} \otimes \boldsymbol{\psi} - \boldsymbol{\psi} \otimes \boldsymbol{\varphi}.}} \tag{34.3}$$

Just as the tensor product allows us to systematically build tensors of higher valence from those of lower valence, so the wedge product allows us to build forms of higher degree from those of lower degree.

Let us stress that what makes $(\boldsymbol{\varphi} \wedge \boldsymbol{\psi})$ into a *2-form* is the fact that when it acts on a pair of vectors, it is linear in both slots, and *swapping the two vector inputs simply reverses the sign of the output, while keeping its magnitude fixed:*

$$(\boldsymbol{\varphi} \wedge \boldsymbol{\psi})(\mathbf{v}_1,\mathbf{v}_2) = \boldsymbol{\varphi}(\mathbf{v}_1)\boldsymbol{\psi}(\mathbf{v}_2) - \boldsymbol{\psi}(\mathbf{v}_1)\boldsymbol{\varphi}(\mathbf{v}_2) = -(\boldsymbol{\varphi} \wedge \boldsymbol{\psi})(\mathbf{v}_2,\mathbf{v}_1).$$

But observe that there is a *second* antisymmetry present, namely, that of the wedge product itself. Keeping the order of the vector inputs the same, but *swapping the 1-forms of the wedge product*, we see that

$$(\boldsymbol{\varphi} \wedge \boldsymbol{\psi})(\mathbf{v}_1,\mathbf{v}_2) = -(\boldsymbol{\psi} \wedge \boldsymbol{\varphi})(\mathbf{v}_1,\mathbf{v}_2).$$

[1] Also called the ***exterior product***.

Abstracting the vectors away, we have the antisymmetry of the wedge product itself, clearly evident in the definition in (34.3):

$$\boxed{(\boldsymbol{\varphi} \wedge \boldsymbol{\psi}) = -(\boldsymbol{\psi} \wedge \boldsymbol{\varphi}).}$$

Note that this implies, for arbitrary $\boldsymbol{\psi}$,

$$\boxed{\boldsymbol{\psi} \wedge \boldsymbol{\psi} = 0.}$$

Also note that the wedge product distributes over addition:

$$\boxed{\boldsymbol{\varphi} \wedge (\boldsymbol{\psi} + \boldsymbol{\sigma}) = \boldsymbol{\varphi} \wedge \boldsymbol{\psi} + \boldsymbol{\varphi} \wedge \boldsymbol{\sigma}.}$$

In $\mathbb{R}^2$, consider all possible wedge products of the Cartesian basis 1-forms. Since $\mathbf{d}x \wedge \mathbf{d}x = 0$ and $\mathbf{d}y \wedge \mathbf{d}y = 0$, the only nonvanishing wedge product is $\mathbf{d}x \wedge \mathbf{d}y = -\mathbf{d}y \wedge \mathbf{d}x$. And, in fact, this is none other than the *area 2-form*!

$$\boxed{\boldsymbol{\mathcal{A}} = \mathbf{d}x \wedge \mathbf{d}y.} \tag{34.4}$$

Note that this again bears a striking resemblance to the classical expression for area: when we do a double integral, we write the element of area as $dx\, dy$. The explicit connection follows the same pattern as our earlier examples of the differential df and of the line element ds. If we take $\mathbf{u} = \delta x\, \mathbf{e}_1$ and $\mathbf{v} = \delta y\, \mathbf{e}_2$ to be the sides of a small rectangle, then

$$(\mathbf{d}x \wedge \mathbf{d}y)(\mathbf{u}, \mathbf{v}) = (\mathbf{d}x \wedge \mathbf{d}y)(\delta x\, \mathbf{e}_1, \delta y\, \mathbf{e}_2) = \delta x\, \delta y.$$

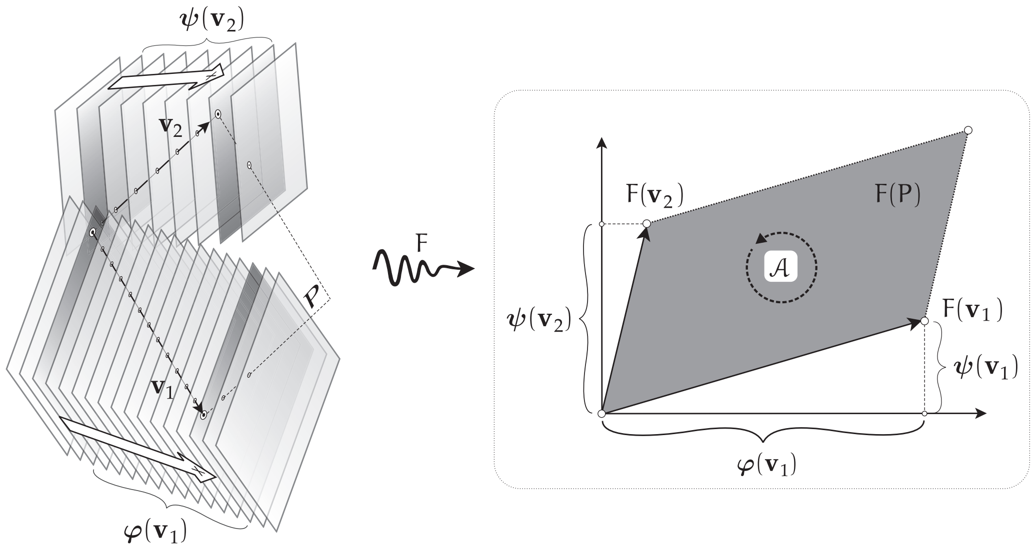
The general proof of (34.4) is easily achieved using components:

$$\begin{aligned}
(\mathbf{d}x \wedge \mathbf{d}y)(\mathbf{u}, \mathbf{v}) &= (\mathbf{d}x \otimes \mathbf{d}y - \mathbf{d}y \otimes \mathbf{d}x)\left(\begin{bmatrix} u^1 \\ u^2 \end{bmatrix}, \begin{bmatrix} v^1 \\ v^2 \end{bmatrix}\right) \\
&= u^1 v^2 - u^2 v^1 \\
&= \det \begin{bmatrix} u^1 & v^1 \\ u^2 & v^2 \end{bmatrix} \\
&= \boldsymbol{\mathcal{A}}(\mathbf{u}, \mathbf{v}).
\end{aligned}$$

Let us now describe the geometrical meaning of the wedge product $(\boldsymbol{\varphi} \wedge \boldsymbol{\psi})$ of two general 1-forms acting on vectors in $\mathbb{R}^n$. The argument will apply to any n, but to keep the idea vivid, let us illustrate it in the case of $\mathbb{R}^3$. In this case, both $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are represented by stacks of planes, and their action on a vector $\mathbf{v}$ in $\mathbb{R}^3$ is to count how many planes of each stack are pierced, as illustrated in [34.2].

Now comes the crucial idea: *let us use these two numbers $\boldsymbol{\varphi}(\mathbf{v})$ and $\boldsymbol{\psi}(\mathbf{v})$ as the x and y coordinates of a point in $\mathbb{R}^2$*. The two 1-forms thereby define a mapping F of vectors in $\mathbb{R}^3$ (or any $\mathbb{R}^n$) to vectors in $\mathbb{R}^2$:

$$\boxed{\mathbf{v} \longrightarrow \mathsf{F}(\mathbf{v}) \equiv \begin{bmatrix} \boldsymbol{\varphi}(\mathbf{v}) \\ \boldsymbol{\psi}(\mathbf{v}) \end{bmatrix}} \tag{34.5}$$



[34.2] Geometrical Meaning of the Wedge Product. *A parallelogram P in $\mathbb{R}^3$ has edges $\mathbf{v}_1$ and $\mathbf{v}_2$, and $\varphi(\mathbf{v}_k)$ and $\psi(\mathbf{v}_k)$ count how many planes of each stack are pierced by the edge $\mathbf{v}_k$. The mapping F fuses these pairs of piercing numbers $(\varphi(\mathbf{v}_k), \psi(\mathbf{v}_k))$ into Cartesian coordinates in $\mathbb{R}^2$, mapping P to the parallelogram $\mathsf{F}(\mathsf{P})$. Then $(\varphi \wedge \psi)(\mathbf{v}_1, \mathbf{v}_2)$ is the oriented area $\mathcal{A}$ of $\mathsf{F}(\mathsf{P})$.*

Under this mapping F, a parallelogram in $\mathbb{R}^n$ with edges $\mathbf{v}_1$ and $\mathbf{v}_2$ is mapped to a parallelogram in $\mathbb{R}^2$ with edges $\mathsf{F}(\mathbf{v}_1)$ and $\mathsf{F}(\mathbf{v}_2)$, as illustrated in [34.2]. We can now state the beautifully simple *meaning* of the wedge product:[2]

> *When the wedge product $(\varphi \wedge \psi)$ is applied to an arbitarary parallelogram in $\mathbb{R}^n$, it outputs the* oriented area of the image parallelogram in $\mathbb{R}^2$ *under the mapping* F*:*
> $$(\varphi \wedge \psi)(\mathbf{v}_1, \mathbf{v}_2) = \boldsymbol{\mathcal{A}}\big[\mathsf{F}(\mathbf{v}_1), \mathsf{F}(\mathbf{v}_2)\big]. \tag{34.6}$$

The proof is very short:

$$\begin{aligned}(\varphi \wedge \psi)(\mathbf{v}_1, \mathbf{v}_2) &= \varphi(\mathbf{v}_1)\psi(\mathbf{v}_2) - \psi(\mathbf{v}_1)\varphi(\mathbf{v}_2) \\ &= \det\begin{bmatrix} \varphi(\mathbf{v}_1) & \varphi(\mathbf{v}_2) \\ \psi(\mathbf{v}_1) & \psi(\mathbf{v}_2) \end{bmatrix} \\ &= \boldsymbol{\mathcal{A}}\big[\mathsf{F}(\mathbf{v}_1), \mathsf{F}(\mathbf{v}_2)\big].\end{aligned}$$

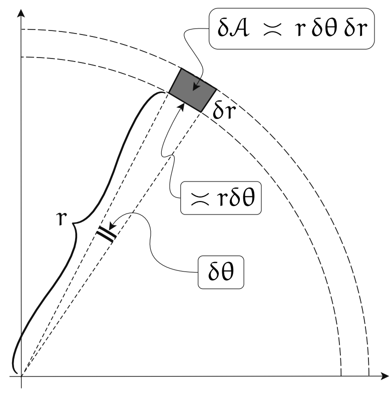
34.4 The Area 2-Form in Polar Coordinates

When we do a double integral in $\mathbb{R}^2$, we write the element of area as $d\mathcal{A} = dx\,dy$, and we have previously explained how this is connected to the area 2-form, (34.4): $\boldsymbol{\mathcal{A}} = \mathbf{d}x \wedge \mathbf{d}y$. But when we switch to polar coordinates we instead write the element of area as $d\mathcal{A} = r\,dr\,d\theta$. In this case,

[2]Arnol'd (1989, §32) goes a step further, taking (34.6) to be the *definition* of the wedge product.

there is a simple geometric derivation of this formula, shown in [34.3]. The corresponding 2-form is therefore $\mathcal{A} = r\,\mathbf{d}r \wedge \mathbf{d}\theta$.

However, when we do a more complicated change of coordinates, the traditional method (which you undoubtedly learned in multivariable calculus) requires us to use the *Jacobian* to find the expression for the element of area in the new coordinate system. As we now demonstrate, using polar coordinates as our example, the use of the area 2-form allows us to jettison the Jacobian, and turns the entire affair into (mindless) algebraic child's play!



[34.3] *Geometric proof that* $\delta\mathcal{A} \asymp r\,\delta\theta\,\delta r$.

Applying the Leibnitz Rule (32.12), and remembering that $\mathbf{d}r \wedge \mathbf{d}r = 0$, and $\mathbf{d}\theta \wedge \mathbf{d}\theta = 0$, and $\mathbf{d}\theta \wedge \mathbf{d}r = -\mathbf{d}r \wedge \mathbf{d}\theta$, we find that

$$\begin{aligned} \mathbf{d}x \wedge \mathbf{d}y &= \mathbf{d}(r\cos\theta) \wedge \mathbf{d}(r\sin\theta) \\ &= [(\mathbf{d}r)\cos\theta - r\,\sin\theta\,\mathbf{d}\theta] \wedge [(\mathbf{d}r)\sin\theta + r\,\cos\theta\,\mathbf{d}\theta] \\ &= \cos^2\theta\,r\,\mathbf{d}r \wedge \mathbf{d}\theta - \sin^2\theta\,r\,\mathbf{d}\theta \wedge \mathbf{d}r \\ &= r\,\mathbf{d}r \wedge \mathbf{d}\theta. \end{aligned}$$

34.5 Basis 2-Forms and Projections

It is clear that the sum of two 2-forms is another 2-form, and likewise multiplication by a constant also yields another 2-form. Therefore, the 2-forms constitute a vector space, and so it is natural to seek a *basis* for this space.

Recall from (33.2) that the set of tensors $\{\mathbf{d}x^i \otimes \mathbf{d}x^j\}$ forms a basis for tensors of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$. This applies to *all* such tensors, including the 2-forms, but in the case of 2-forms we go a step further:

$$\boxed{\textit{The set of 2-forms } \{\mathbf{d}x^i \wedge \mathbf{d}x^j\}\textit{, with } i<j\textit{, is a } \textbf{basis} \textit{ for the 2-forms.}} \tag{34.7}$$

The condition $i<j$ is imposed simply to avoid listing duplicate 2-forms: e.g., $(\mathbf{d}x^3 \wedge \mathbf{d}x^2) = -(\mathbf{d}x^2 \wedge \mathbf{d}x^3)$.

Since $\mathbf{d}x^i \wedge \mathbf{d}x^i = 0$, the nonvanishing (nonredundant) basis 2-forms result from picking an unordered pair of distinct $\mathbf{d}x^i$'s, (out of n) then forming their wedge product. So, if we assume (34.7) is true, then,

$$\boxed{\textit{In } \mathbb{R}^n\textit{, the set of 2-forms is a vector space of dimension } \tfrac{1}{2}n(n-1).} \tag{34.8}$$

To confirm (34.7), let us begin in $\mathbb{R}^2$. If $\boldsymbol{\Psi}$ were merely a generic tensor of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$, then we know from (33.2) that it can be expanded as

$$\boldsymbol{\Psi} = \Psi_{11}(\mathbf{d}x \otimes \mathbf{d}x) + \Psi_{12}(\mathbf{d}x \otimes \mathbf{d}y) + \Psi_{21}(\mathbf{d}y \otimes \mathbf{d}x) + \Psi_{22}(\mathbf{d}y \otimes \mathbf{d}y).$$

But, because $\boldsymbol{\Psi}$ is a *2-form*

$$\Psi_{11} = 0 = \Psi_{22} \qquad \text{and} \qquad \Psi_{21} = -\Psi_{12}.$$

Therefore,

$$\boxed{\boldsymbol{\Psi} = \Psi_{12}\,(\mathbf{d}x \wedge \mathbf{d}y) = \Psi_{12}\,\boldsymbol{\mathcal{A}}.}$$

In other words, $(\mathbf{d}x \wedge \mathbf{d}y)$ is a *basis* for 2-forms in $\mathbb{R}^2$, and every 2-form is proportional to the area 2-form.

If we now go up one dimension to $\mathbb{R}^3$, then [exercise] exactly the same reasoning yields

$$\boxed{\boldsymbol{\Psi} = \Psi_{23}\,(\mathbf{d}y \wedge \mathbf{d}z) + \Psi_{31}\,(\mathbf{d}z \wedge \mathbf{d}x) + \Psi_{12}\,(\mathbf{d}x \wedge \mathbf{d}y),} \tag{34.9}$$

confirming (34.7) for this case. (NOTE: The reason for writing these three terms in the way that we have, and in the strange order that we have, will be explained in the next section.)

The proof in $\mathbb{R}^n$ is a straightforward extension of these examples. The nonvanishing terms in the expansion (33.2) always come in opposite pairs:

$$\Psi_{ij}\,(\mathbf{d}x^i \otimes \mathbf{d}x^j) + \Psi_{ji}\,(\mathbf{d}x^j \otimes \mathbf{d}x^i) = \Psi_{ij}\,(\mathbf{d}x^i \otimes \mathbf{d}x^j) - \Psi_{ij}\,(\mathbf{d}x^j \otimes \mathbf{d}x^i) = \Psi_{ij}\,(\mathbf{d}x^i \wedge \mathbf{d}x^j).$$

Let us return to $\mathbb{R}^3$ and seek out the *meaning* of the basis 2-forms that arise in (34.9), beginning with $(\mathbf{d}x \wedge \mathbf{d}y)$. In $\mathbb{R}^2$ this was simply the area 2-form in the (x, y)-plane, and the meaning in $\mathbb{R}^3$ is closely related, as illustrated in [34.4]:

> *If* $(\mathbf{d}x \wedge \mathbf{d}y)$ *is applied to a parallelogram* P *in* $\mathbb{R}^3$*, the output is the* oriented area $\mathcal{A}_z$ of the orthogonal projection *(in the z-direction)* of P onto the (x, y)-plane.

This is a direct consequence of the geometric interpretation (34.6) of the wedge product of two arbitrary 1-forms. The key is to realize that in this case the mapping F simply reduces to orthogonal projection in the z-direction onto the (x, y)-plane. Simply compare with the general construction of the wedge product in [34.2]: here $\varphi = \mathbf{d}x$ is a unit-spaced stack perpendicular to the x-axis, and likewise $\psi = \mathbf{d}y$ is a unit-spaced stack perpendicular to the y-axis. (Please take a moment to make sure that you can actually *see* all this.)

As illustrated in [34.4], the meaning of the other two basis 2-forms is completely analogous: $(\mathbf{d}y \wedge \mathbf{d}z)$ is the area $\mathcal{A}_x$ of the projection (in the x-direction) onto the (y, z)-plane; and $(\mathbf{d}z \wedge \mathbf{d}x)$ is the area $\mathcal{A}_y$ of the projection (in the y-direction) onto the (z, x)-plane.

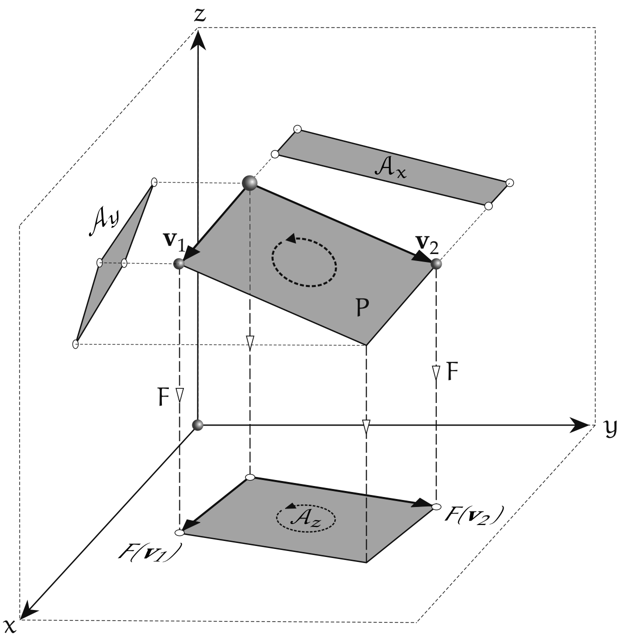
In $\mathbb{R}^4$, the generalization of (34.9) instead contains *six* basis 2-forms, $(\mathbf{d}x^i \wedge \mathbf{d}x^j)$, but each one has the same meaning as above: for example, $(\mathbf{d}x^1 \wedge \mathbf{d}x^3)$ yields the oriented area in the (x^1, x^3)-plane after projection along the x^2 and x^4 axes.

34.6 Associating 2-Forms with Vectors in $\mathbb{R}^3$: Flux

As we shall see, it is possible to do calculus with Forms in any number of dimensions. But the *Vector Calculus* we learn as undergraduates in college is *only* possible in *three* dimensions, and in this section and the next we begin to understand why.

Geometrically, there is certainly no confusion possible between 2-forms and vectors. Even algebraically the distinction is clear: in n dimensions, a vector has n components, whereas a 2-form has $\frac{1}{2}n(n-1)$ components, as we saw in (34.8).

But this implies that something remarkable happens when $n=3$: *In three dimensions, and* only *in three dimensions, 2-forms have the same number of components as vectors*.



[34.4] Geometric Meaning of Basis 2-Forms. *Each basis 2-form yields the area of the projection of* P *onto the relevant coordinate plane. For example,* $(\mathbf{d}x \wedge \mathbf{d}y)(\mathbf{v}_1, \mathbf{v}_2) = \mathcal{A}_z$*, the area of the projection of* P *in the* z*-direction onto the* (x, y)*-plane.*

Vector Calculus, so successfully pioneered by physicists in the 1880s, and still considered an indispensable tool of modern science in the twenty-first century, can now be seen to rest upon this singular *numerical coincidence*. Though the pioneers of Vector Calculus could not have had any inkling as to why their mathematical engine purred with such power, the *reason* was that it actually had *2-forms* hidden under the bonnet, but *masquerading* as vectors!

To begin to look under the bonnet, let us rewrite the components (34.9) of the general 2-form $\boldsymbol{\Psi}$ with a single raised index, and immediately identity these with components of a corresponding *vector*, which we shall denote by the same symbol as the 2-form, but ***underlined***: $\underline{\boldsymbol{\Psi}}$, i.e.,

$$\boxed{\boldsymbol{\Psi} = \Psi^1\,(\mathbf{d}x^2 \wedge \mathbf{d}x^3) + \Psi^2\,(\mathbf{d}x^3 \wedge \mathbf{d}x^1) + \Psi^3\,(\mathbf{d}x^1 \wedge \mathbf{d}x^2) \quad \rightleftarrows \quad \underline{\boldsymbol{\Psi}} = \begin{bmatrix} \Psi^1 \\ \Psi^2 \\ \Psi^3 \end{bmatrix}.} \tag{34.10}$$

The most natural mathematical way to understand this correspondence is via the ***Hodge star duality operator*** ($\star$), which is the subject of Exercise 15. It is a purely mathematical operation that (in n dimensions) maps a p-form to an $(n-p)$-form, called its ***Hodge dual*** (usually abbreviated simply to ***dual***), named after the important British mathematician, Sir W.V.D. Hodge (1903–1975).[3]

Here, however, we shall instead pursue a logical and compelling *physical* reason for associating a 2-form with a vector in this way, and to understand it we must now introduce the physicist's powerful concept of ***flux***.[4]

---

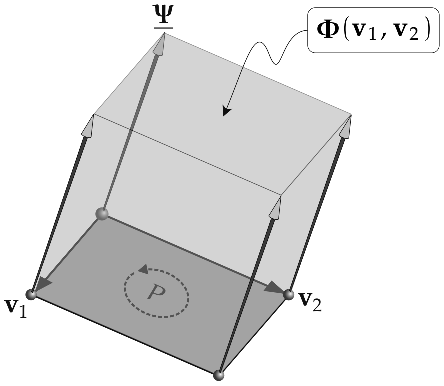
[3]Hodge was my mathematical grandfather: I studied under Penrose, who studied under Hodge.

[4]We shall explain the concept of flux using fluid flow, but it was originally introduced intuitively by Faraday to describe his ingenious experiments (beginning in the 1820s) on the flow of electric and magnetic fields through a surface in space; in the decades that followed, Maxwell gave the concept its current mathematical form. See Forbes and Mahon (2014, §10).

Reconsider [34.4], and imagine a uniform flow of fluid through space with velocity $\underline{\boldsymbol{\Psi}}$. Now define,

$$\boxed{\boldsymbol{\Phi}(\mathbf{v}_1,\mathbf{v}_2) \equiv \text{Amount of fluid crossing P per unit time} = \textbf{\textit{flux}} \text{ of } \underline{\boldsymbol{\Psi}} \text{ through P.}} \tag{34.11}$$

We count the flux as *positive* if the orientation of P is *counterclockwise* around $\underline{\boldsymbol{\Psi}}$. If we picture all the fluid that passes through P in unit time, it will fill a parallelepiped with edges $\mathbf{v}_1$, $\mathbf{v}_2$, $\underline{\boldsymbol{\Psi}}$, and the flux will be its volume, as illustrated in [34.5]. It follows [exercise] that the flux $\boldsymbol{\Phi}$ is actually a *2-form*.



[34.5] Flux *of* $\underline{\boldsymbol{\Psi}}$ *through* P = *volume* $\boldsymbol{\Phi}(\mathbf{v}_1,\mathbf{v}_2)$

It is natural to now ask the following question: *How is this flux 2-form* $\boldsymbol{\Phi}$ *related to the 2-form* $\boldsymbol{\Psi}$ *associated with the flow velocity* $\underline{\boldsymbol{\Psi}}$*?*

We begin with a simple example. Suppose in [34.4] that $\underline{\boldsymbol{\Psi}} = \Psi^3\,\mathbf{e}_3$ is purely upward, parallel to the z-axis. Clearly the flux through P is the same as the flux through its projection $\mathcal{A}_z = \mathcal{A}_3$. But if we picture the fluid flowing upward with speed Ψ^3 through $\mathcal{A}_3$ for one unit of time, then the fluid will fill a solid whose base area is $\mathcal{A}_3$ and whose height is Ψ^3, so the flux is $\Psi^3\mathcal{A}_3$.

In the general case, the same reasoning applies to each component of $\underline{\boldsymbol{\Psi}}$, so the total flux is simply the sum:

$$\begin{aligned}
\boldsymbol{\Phi}(\mathbf{v}_1,\mathbf{v}_2) &= \Psi^1\mathcal{A}_1 + \Psi^2\mathcal{A}_2 + \Psi^3\mathcal{A}_3 \\
&= \Psi^1\,(\mathbf{d}x^2\wedge\mathbf{d}x^3)(\mathbf{v}_1,\mathbf{v}_2) + \Psi^2\,(\mathbf{d}x^3\wedge\mathbf{d}x^1)(\mathbf{v}_1,\mathbf{v}_2) + \Psi^3\,(\mathbf{d}x^1\wedge\mathbf{d}x^2)(\mathbf{v}_1,\mathbf{v}_2) \\
&= \boldsymbol{\Psi}(\mathbf{v}_1,\mathbf{v}_2).
\end{aligned}$$

Therefore,

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}\,!$$

Thus we have obtained a satisfying physical justification for associating a 2-form with a vector via (34.10):

$$\boxed{\text{\textit{If fluid flows through 3-dimensional space with velocity } } \underline{\boldsymbol{\Psi}}\text{\textit{, its flux 2-form is }} \boldsymbol{\Psi}\text{\textit{, and vice versa.}}} \tag{34.12}$$

In addition to using components to connect the 2-form $\boldsymbol{\Psi}$ to its associated vector $\underline{\boldsymbol{\Psi}}$, as we initially did in (34.10), we may also identify the *direction* of the associated vector *geometrically*, instead. The flux of $\underline{\boldsymbol{\Psi}}$ across the parallelogram P *vanishes* if and only if it contains the direction $\underline{\boldsymbol{\Psi}}$, so *we can uniquely characterize the flow direction by means of vanishing flux:*

$$\boxed{\boldsymbol{\Psi}(\underline{\boldsymbol{\Psi}},\ldots) = 0.} \tag{34.13}$$

34.7 Relation of the Vector and Wedge Products in $\mathbb{R}^3$

The conjuring trick that pulls Vector Calculus out of the top hat labelled "Forms" has a second component: we also associate *1-forms* with vectors, just as we previously did in Section 33.8. Given a 1-form φ, we will again denote the corresponding vector with same symbol as the 1-form, but ***underlined***:

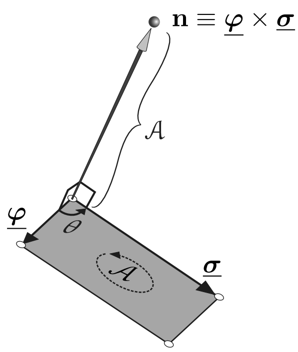
$$\boxed{\varphi = \varphi_1\, \mathbf{d}x^1 + \varphi_2\, \mathbf{d}x^2 + \varphi_3\, \mathbf{d}x^3 \qquad \rightleftarrows \qquad \underline{\varphi} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{bmatrix}.} \tag{34.14}$$

The two fundamental kinds of multiplication in Vector Calculus are the ***scalar product*** (aka ***dot product***) and the ***vector product*** (aka ***cross product***). The former readily generalizes to any number of dimensions, and can be expressed as

$$\boxed{\varphi(\mathbf{v}) = \underline{\varphi} \bullet \mathbf{v}.} \tag{34.15}$$

But the vector product, illustrated in [34.6], is peculiar to three dimensions. Given two vectors $\underline{\varphi}$ and $\underline{\sigma}$, with the angle from the first to the second being θ, recall that we define the ***vector product*** $\underline{\varphi} \times \underline{\sigma}$ to be the vector orthogonal to both, its direction given by the right-hand rule, and its length being the area of the parallelogram they span:

$$|\underline{\varphi} \times \underline{\sigma}| \equiv \mathcal{A}(\underline{\varphi}, \underline{\sigma}) = |\underline{\varphi}|\, |\underline{\sigma}| \sin\theta.$$



[34.6] The Vector Product.

We should immediately be suspicious: the *length* of a vector is equal to an *area*?! The reality is that we are dealing with a second sleight of hand: the vector product is not what it appears to be—it is actually the *wedge* product of the corresponding *1-forms*! The full explanation for this lies with the previously mentioned *Hodge duality*; see Exercise 15.

$$\boxed{\underline{\varphi} \times \underline{\sigma} \qquad \rightleftarrows \qquad \varphi \wedge \sigma.} \tag{34.16}$$

Before proving this correspondence, let us pause to recognize how special it is: in *any* number of dimensions, we are free to associate two vectors with two 1-forms, and we may then form their wedge product, but the resulting 2-form will have $\frac{1}{2}n(n-1)$ components, and so in general it *cannot* be reinterpreted as a vector. *Only* in three dimensions can a 2-form be identified as the flux 2-form $\mathbf{\Psi}$ of a vector flow $\underline{\mathbf{\Psi}}$.

In order to prove (34.16), let us define the 2-form $\boldsymbol{\Psi} \equiv \boldsymbol{\varphi} \wedge \boldsymbol{\sigma}$. Then we must show that $\underline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\varphi}} \times \underline{\boldsymbol{\sigma}}$. We will prove this in two stages. First we will show that the vectors $\underline{\boldsymbol{\Psi}}$ and $\mathbf{n} \equiv \underline{\boldsymbol{\varphi}} \times \underline{\boldsymbol{\sigma}}$ point in the same direction, and then we will show that they have the same length.

Since $\mathbf{n} \equiv \underline{\boldsymbol{\varphi}} \times \underline{\boldsymbol{\sigma}}$ is, by definition, orthogonal to $\underline{\boldsymbol{\varphi}}$ and $\underline{\boldsymbol{\sigma}}$,

$$\boldsymbol{\varphi}(\mathbf{n}) = \underline{\boldsymbol{\varphi}} \bullet \mathbf{n} = 0 \qquad \text{and} \qquad \boldsymbol{\sigma}(\mathbf{n}) = \underline{\boldsymbol{\sigma}} \bullet \mathbf{n} = 0.$$

This then implies that the flux vanishes for any parallelogram containing $\mathbf{n}$:

$$\boldsymbol{\Psi}(\mathbf{n}, \ldots) = (\boldsymbol{\varphi} \wedge \boldsymbol{\sigma})(\mathbf{n}, \ldots) = \boldsymbol{\varphi}(\mathbf{n})\, \boldsymbol{\sigma}(\ldots) - \boldsymbol{\sigma}(\mathbf{n})\, \boldsymbol{\varphi}(\ldots) = 0.$$

Thus, as discussed in (34.13), this implies that $\mathbf{n}$ is aligned with the flow: $\underline{\boldsymbol{\Psi}} \propto \mathbf{n}$.

It remains to show that the *speeds* of the two flows match: $|\underline{\boldsymbol{\Psi}}| = |\mathbf{n}| = \mathcal{A}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}}) = |\underline{\boldsymbol{\varphi}}|\, |\underline{\boldsymbol{\sigma}}| \sin\theta$. We begin by noting that since we now know that $\underline{\boldsymbol{\Psi}}$ is orthogonal to the parallelogram P with edges $\underline{\boldsymbol{\varphi}}$ and $\underline{\boldsymbol{\sigma}}$, its flux across P is given by

$$\text{flux} = |\underline{\boldsymbol{\Psi}}|\, \mathcal{A}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}}).$$

But,

$$\begin{aligned}
\text{flux} &= \boldsymbol{\Psi}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}}) \\
&= (\boldsymbol{\varphi} \wedge \boldsymbol{\sigma})(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}}) \\
&= \boldsymbol{\varphi}(\underline{\boldsymbol{\varphi}})\, \boldsymbol{\sigma}(\underline{\boldsymbol{\sigma}}) - \boldsymbol{\sigma}(\underline{\boldsymbol{\varphi}})\, \boldsymbol{\varphi}(\underline{\boldsymbol{\sigma}}) \\
&= |\underline{\boldsymbol{\varphi}}|^2\, |\underline{\boldsymbol{\sigma}}|^2 - |\underline{\boldsymbol{\varphi}}|^2\, |\underline{\boldsymbol{\sigma}}|^2 \cos^2\theta \\
&= [\mathcal{A}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}})]^2.
\end{aligned}$$

Combining the previous two results,

$$|\underline{\boldsymbol{\Psi}}|\, \mathcal{A}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}}) = \text{flux} = [\mathcal{A}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}})]^2.$$

Therefore,

$$|\underline{\boldsymbol{\Psi}}| = \mathcal{A}(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{\sigma}}) = |\underline{\boldsymbol{\varphi}} \times \underline{\boldsymbol{\sigma}}| \qquad \Longrightarrow \qquad \underline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\varphi}} \times \underline{\boldsymbol{\sigma}},$$

proving (34.16).

Having established (34.16) geometrically, we may now use it to derive the standard algebraic expression for the vector product from the underlying formula for the wedge product:

$$\begin{aligned}
\boldsymbol{\Psi} &= \boldsymbol{\varphi} \wedge \boldsymbol{\sigma} \\
&= (\varphi_1\, \mathbf{d}x^1 + \varphi_2\, \mathbf{d}x^2 + \varphi_3\, \mathbf{d}x^3) \wedge (\sigma_1\, \mathbf{d}x^1 + \sigma_2\, \mathbf{d}x^2 + \sigma_3\, \mathbf{d}x^3) \\
&= (\varphi_2\, \sigma_3 - \varphi_3\, \sigma_2)(\mathbf{d}x^2 \wedge \mathbf{d}x^3) + (\varphi_3\, \sigma_1 - \varphi_1\, \sigma_3)(\mathbf{d}x^3 \wedge \mathbf{d}x^1) + (\varphi_1\, \sigma_2 - \varphi_2\, \sigma_1)(\mathbf{d}x^1 \wedge \mathbf{d}x^2).
\end{aligned}$$

The associated vector $\underline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\varphi}} \times \underline{\boldsymbol{\sigma}}$ is therefore given by the familiar formula from Vector Calculus:

$$\begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{bmatrix} \times \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} = \begin{bmatrix} \varphi_2\, \sigma_3 - \varphi_3\, \sigma_2 \\ \varphi_3\, \sigma_1 - \varphi_1\, \sigma_3 \\ \varphi_1\, \sigma_2 - \varphi_2\, \sigma_1 \end{bmatrix}.$$

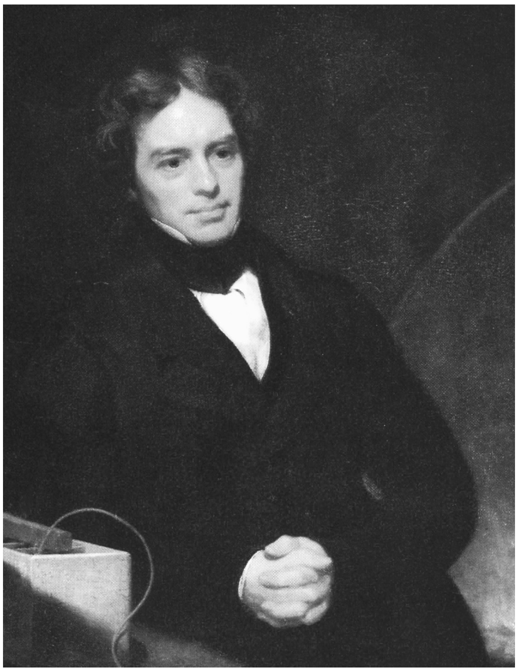
Last, referring back to [34.5] and (34.12), we also deduce the standard formula for the volume of a parallelepiped with edges **u**, **v**, and $\underline{\boldsymbol{\Omega}}$:

$$\begin{aligned}\text{Volume} = \text{flux} = \boldsymbol{\Omega}(\mathbf{u}, \mathbf{v}) &= \begin{bmatrix} \Omega^1 \\ \Omega^2 \\ \Omega^3 \end{bmatrix} \bullet \begin{bmatrix} u^2 v^3 - u^3 v^2 \\ u^3 v^1 - u^1 v^3 \\ u^1 v^2 - u^2 v^1 \end{bmatrix} \\ &= \det \begin{bmatrix} \Omega^1 & u^1 & v^1 \\ \Omega^2 & u^2 & v^2 \\ \Omega^3 & u^3 & v^3 \end{bmatrix}.\end{aligned} \tag{34.17}$$

34.8 The Faraday and Maxwell Electromagnetic 2-Forms

INTRODUCTORY NOTES: This section is entirely optional; it is not required to understand the mathematics that will follow (though it is a prerequisite for Section 36.6). That said, this section *is* very important, for it is our first demonstration that Forms are not only inhabitants of a Platonic, *mathematical* universe—they also reach out and weave their way into the very fabric of the laws of the *physical* Universe! One more thing: in order to ease the transition of the reader in passing from our treatment to more advanced works—such as Misner, Thorne, and Wheeler (1973), Schutz (1980), Baez and Muniain (1994), and Frankel (2012)—we have (reluctantly) adopted their convention, in which *the metric coefficients of spacetime are the negatives of those we have employed up to this point*. So, now and hereafter, the Minkowski metric is

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2.$$



[34.7] *Michael Faraday (1791–1867).*

There are several 2-forms in 4-dimensional spacetime that play fundamental roles in the laws of physics. In this section we shall describe two of these that together lead to an extraordinarily compact and elegant description of the laws of electromagnetism (detailed in Section 36.6). These 2-forms are named after Michael Faraday (1791–1867) and James Clerk Maxwell (1831–1879), who are pictured in [34.7] and [34.8], respectively.

The laws of electromagnetism were first discovered empirically in a long, symphonic sequence of utterly ingenious experiments conducted by the self-educated Michael Faraday in his basement laboratory within the Royal Institution of London.[5] These experiments began in the 1820s, and reached their climax in 1830s. See Forbes and Mahon (2014, §§4–5).

[5]We *strongly* encourage you to visit the remains of Faraday's laboratory, where you may kneel down (reverentially!) to study the *first* electric generator—carefully preserved, mere inches away from you behind glass—ever constructed by the human race!

[34.8] *James Clerk Maxwell (1831–1879).*

Maxwell was quick to recognize the explanatory value of Faraday's ideas of lines of force and of flux, and, unlike Faraday, Maxwell possessed the mathematical skills to give them precise form. Indeed, not since Newton had the world witnessed such an alchemical fusion of profound physical insight and raw mathematical power.

By 1873 the job was done: Maxwell announced four[6] equations that completely described electricity, magnetism, and the *interaction* between the two. In particular, Maxwell used his equations to calculate the speed with which purely theoretical *waves* of electromagnetic energy would travel through empty space. His numerical answer coincided with the experimentally measured speed of light—*Maxwell was the first human to understand what light* is—*an electromagnetic wave!*

It is moving to read the letters[7] that passed between these two giants of science. One is struck by Maxwell's deep admiration of the profundity of Faraday's physical insights and his experimental ingenuity, and, symmetrically, Faraday's awe of Maxwell's extraordinary prowess in crystalizing physical intuitions into precise mathematical laws.

Einstein's 1905 discovery of Special Relativity was directly connected to these electromagnetic revelations of Faraday and Maxwell. Indeed, the title of Einstein's epoch-making paper *makes no mention of either space or time*—it reads, "On the Electrodynamics of Moving Bodies."[8]

The impetus for Einstein's breakthrough was his realization that electric and magnetic fields do not *separately* have an absolute existence—they are merely two, observer-dependent aspects of a single *electromagnetic* field, and waves of this field (light!) travel at the same speed for all observers. This *led* him (via Minkowski) to the realization that space and time cannot separately have an absolute existence—they are observer-dependent aspects of the geometry of *spacetime*, which does have absolute existence.

Let us simply quote the opening paragraph of Einstein's paper of 1905:

> It is known that Maxwell's electrodynamics—as usually understood at the present time—when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena. Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighbourhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary

[6]It was actually Oliver Heaviside who in 1885 reduced Maxwell's original set of 20 equations to the four we now know as "Maxwell's Equations"! See Forbes and Mahon (2014, §16) and Mahon (2017).

[7]See Forbes and Mahon (2014), and Jones (1870).

[8]The entire paper (translated into English) is freely available on the Internet: simply search for its title.

and the conductor in motion, no electric field arises in the neighbourhood of the magnet. In the conductor, however, we find an electromotive force, to which in itself there is no corresponding energy, but which gives rise—assuming equality of relative motion in the two cases discussed—to electric currents of the same path and intensity as those produced by the electric forces in the former case.

The classical description of these fields is as two 3-dimensional vector fields, with six components in all, each one being a function of space and of time:

$$\text{Electric field} = \underline{\mathbf{E}} = \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \qquad \text{and} \qquad \text{Magnetic field} = \underline{\mathbf{B}} = \begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix}.$$

Each of these may be associated both with a flux 2-form—via (34.10)—*and* with a 1-form, via (34.14). For the electric field we shall call the flux 2-form $\mathbf{E}$, and the 1-form $\boldsymbol{\varepsilon}$:

$$\boxed{\mathbf{E} = E_x\,(\mathbf{d}y \wedge \mathbf{d}z) + E_y\,(\mathbf{d}z \wedge \mathbf{d}x) + E_z\,(\mathbf{d}x \wedge \mathbf{d}y),} \tag{34.18}$$

and

$$\boxed{\boldsymbol{\varepsilon} = E_x\,\mathbf{d}x + E_y\,\mathbf{d}y + E_z\,\mathbf{d}z.} \tag{34.19}$$

We do the same for the magnetic field, writing the flux 2-form as $\mathbf{B}$, and the 1-form as $\boldsymbol{\beta}$:

$$\boxed{\mathbf{B} = B_x\,(\mathbf{d}y \wedge \mathbf{d}z) + B_y\,(\mathbf{d}z \wedge \mathbf{d}x) + B_z\,(\mathbf{d}x \wedge \mathbf{d}y)} \tag{34.20}$$

and

$$\boxed{\boldsymbol{\beta} = B_x\,\mathbf{d}x + B_y\,\mathbf{d}y + B_z\,\mathbf{d}z.} \tag{34.21}$$

As Einstein explained (see above), *none* of these 2-forms and 1-forms has an absolute, observer-independent existence. That is, observers in relative motion will disagree about the electric and magnetic fields, but they will *also* disagree about space and time in precisely such a way as to make physical phenomena independent of the observer's motion: there can be only *one* physical reality!

Very remarkably—and very beautifully!—*Nature informs us that the electric and magnetic 2-forms and 1-forms combine into a single* electromagnetic *2-form that* does *have absolute meaning in spacetime!*

It is called[9] the ***Faraday 2-form***, and is denoted $\mathbf{F}$:

$$\boxed{\boxed{\mathbf{F} = \mathbf{Faraday} = \boldsymbol{\varepsilon} \wedge \mathbf{d}t + \mathbf{B}.}} \tag{34.22}$$

Let us immediately describe the physical (observer-independent) meaning of $\mathbf{F}$.

[9]This is the terminology employed by Misner, Thorne, and Wheeler (1973), but it is not universal: for example, Penrose (2005) instead calls $\mathbf{F}$ the "Maxwell 2-form." As we shall explain in a moment, we reserve the name ***Maxwell 2-form*** for a *different* 2-form.

If we insert a vector $\mathbf{u}$ into the second slot of $\mathbf{F}$ then we obtain $\mathbf{F}(\ldots, \mathbf{u})$. This will yield a number when a vector is inserted into the first slot; therefore, as it stands, it must be a 1-form.

Suppose we place a charge q in the electromagnetic field described by $\mathbf{F}$. Let the particle's 4-velocity vector in spacetime be $\mathbf{u}$, let its 4-momentum be described by the *1-form* $\boldsymbol{\pi}$, and let its proper time (aka "wrist-watch time") be τ. *Then* $q\,\mathbf{F}(\ldots, \mathbf{u})$ *describes the electromagnetic force exerted on the particle*

$$\boxed{\frac{d\boldsymbol{\pi}}{d\tau} = q\,\mathbf{F}(\ldots, \mathbf{u}).} \tag{34.23}$$

This is a 1-form, awaiting the insertion of a vector, and (physically) the resulting number is the magnitude of the force acting in the direction of the inserted vector.

Contrast this with the more complicated formula—the ***Lorentz Force Law***—that one learns in introductory electrodynamics for the rate of change of the spatial vector momentum $\mathbf{p}$ of a particle that has spatial vector velocity $\mathbf{v}$:

$$\frac{d\mathbf{p}}{dt} = q\,(\underline{\mathbf{E}} + \mathbf{v} \times \underline{\mathbf{B}}).$$

In this classical formulation, *different observers in relative motion will all disagree about the values of* $\mathbf{p}$, $\underline{\mathbf{E}}$, $\mathbf{v}$, *and* $\underline{\mathbf{B}}$! (For a fuller discussion of this point, see Misner, Thorne, and Wheeler (1973, §3.3).) Contrast this with the elegance of the geometrical, observer-independent, spacetime law, (34.23)!

Before we introduce the Maxwell 2-form, let us state ***Maxwell's Equations*** of electromagnetism, which we shall divide into two pairs of equations.

The first pair describes the electromagnetic field in the absence of *sources* (meaning the (scalar) electric charge density ρ, and the (vector) current density $\underline{\mathbf{j}}$):

$$\boxed{\textit{Maxwell's Source-Free Equations}: \quad \begin{aligned} \nabla \bullet \underline{\mathbf{B}} &= 0, \\ \nabla \times \underline{\mathbf{E}} + \partial_t \underline{\mathbf{B}} &= 0. \end{aligned}} \tag{34.24}$$

As we shall see in Section 36.6, when expressed in terms of the Faraday 2-form, these two equations reduce to a *single*, elegant equation.

The second pair of Maxwell's equations describes the electromagnetic field generated by sources:

$$\boxed{\textit{Maxwell's Source Equations}: \quad \begin{aligned} \nabla \bullet \underline{\mathbf{E}} &= 4\pi\rho, \\ \nabla \times \underline{\mathbf{B}} - \partial_t \underline{\mathbf{E}} &= 4\pi\,\underline{\mathbf{j}}. \end{aligned}} \tag{34.25}$$

This second pair *also* reduces to a single, elegant equation—to be derived in Section 36.6—but this time in terms of a different 2-form that is closely to related to $\mathbf{F}$, and is denoted $\star\mathbf{F}$. This is the ***Maxwell 2-form***[10], which is *almost* (but not quite) the result of interchanging the roles of the electric and magnetic fields in the Faraday 2-form:

[10] As noted in footnote 9, this terminology is not universal: Penrose (2005) instead calls $\mathbf{F}$ the ***Maxwell 2-form***.

$$\boxed{\star\mathbf{F} = \mathbf{Maxwell} = \boldsymbol{\beta} \wedge \mathbf{d}t - \mathbb{E}.} \tag{34.26}$$

The star operator ($\star$) is the previously mentioned *Hodge duality operator*; see Exercise 15. It is a purely mathematical operation that (in n dimensions) maps a p-form to an $(n-p)$-form. Since we are in 4 dimensions, $\star$ maps the 2-form $\mathbf{F}$ to another 2-form $\star\mathbf{F}$, the ***(Hodge) dual of*** $\mathbf{F}$. If we apply the star operator a second time, in other words take the dual of Maxwell, then we recover Faraday (only with a minus):

$$\boxed{\star\star\mathbf{F} = -\mathbf{F}.} \tag{34.27}$$

Thus we may say that Maxwell and Faraday are the duals of one another—a nomenclature that we suspect both scientists would have relished, if only they had lived to see it.

Chapter 35

3-Forms

35.1 A 3-Form Requires Three Dimensions

Let us begin by recognizing that 3-forms require at least three dimensions in order to exist.

In $\mathbb{R}^2$, it is easy enough to create a tensor $\mathbf{H}$ of valence $\left\{\begin{smallmatrix} 0 \\ 3 \end{smallmatrix}\right\}$. For example, take any 1-form φ, and construct,

$$\mathbf{H} = \varphi \otimes \varphi \otimes \varphi \qquad \Longrightarrow \qquad \mathbf{H}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \varphi(\mathbf{v}_1)\, \varphi(\mathbf{v}_2)\, \varphi(\mathbf{v}_3).$$

As we see, this tensor is *totally symmetric*, meaning that swapping any pair of input vectors returns the same output.

But to create a *3-form*, Ξ, we require (by definition) that our tensor be *totally antisymmetric*, and this is impossible! Consider the components of Ξ:

$$\Xi_{ijk} = \Xi\,(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k).$$

In $\mathbb{R}^2$ there are only two basis vectors, $\{\mathbf{e}_1, \mathbf{e}_2\}$, so that means two of the slots of Ξ contain the *same* vector, but then antisymmetry demands that the output vanishes.

In order for Ξ_{ijk} not to vanish identically, it is therefore essential that all three indices be distinct: in other words, we must be in *at least* three dimensions. Clearly, exactly the same reasoning shows, quite generally, that the existence of a p-form requires at least p dimensions.

We can also turn this around and observe that *the highest degree form that can exist in* n *dimensions is an* n*-form.*

35.2 The Wedge Product of a 2-Form and 1-Form

The wedge product of two 1-forms yielded a 2-form, and we now seek to extend the definition of the wedge product so as to ensure that multiplication of a 2-form $\boldsymbol{\Psi}$ and a 1-form $\boldsymbol{\sigma}$ automatically yields a *3-form*, $\boldsymbol{\Psi} \wedge \boldsymbol{\sigma}$.

Let us begin with simplest possible (ill-fated) guess: how about

$$(\boldsymbol{\Psi} \wedge \boldsymbol{\sigma})(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) \overset{???}{=} \boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2)\, \boldsymbol{\sigma}(\mathbf{v}_3)\ ?$$

Well, this is clearly antisymmetric under $\mathbf{v}_1 \leftrightarrow \mathbf{v}_2$, but not under $\mathbf{v}_2 \leftrightarrow \mathbf{v}_3$, so let us *impose* antisymmetry, using the same trick we used in (33.8). That is, let us subtract the same expression with $\mathbf{v}_2 \leftrightarrow \mathbf{v}_3$:

$$(\boldsymbol{\Psi} \wedge \boldsymbol{\sigma})(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) \overset{???}{=} \boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2)\, \boldsymbol{\sigma}(\mathbf{v}_3) - \boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_3)\, \boldsymbol{\sigma}(\mathbf{v}_2)\ ?$$

But now the second term has ruined antisymmetry under $\mathbf{v}_1 \leftrightarrow \mathbf{v}_2$, so repeat the trick and subtract that same term, but with $\mathbf{v}_1 \leftrightarrow \mathbf{v}_2$:

$$(\boldsymbol{\Psi} \wedge \boldsymbol{\sigma})(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) \overset{???}{=} \boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2)\, \boldsymbol{\sigma}(\mathbf{v}_3) - \big[\boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_3)\, \boldsymbol{\sigma}(\mathbf{v}_2) - \boldsymbol{\Psi}(\mathbf{v}_2, \mathbf{v}_3)\, \boldsymbol{\sigma}(\mathbf{v}_1)\big]\ ?$$

Success!

$$\boxed{(\boldsymbol{\Psi}\wedge\boldsymbol{\sigma})(\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3)=\boldsymbol{\Psi}(\mathbf{v}_1,\mathbf{v}_2)\,\boldsymbol{\sigma}(\mathbf{v}_3)+\boldsymbol{\Psi}(\mathbf{v}_3,\mathbf{v}_1)\,\boldsymbol{\sigma}(\mathbf{v}_2)+\boldsymbol{\Psi}(\mathbf{v}_2,\mathbf{v}_3)\,\boldsymbol{\sigma}(\mathbf{v}_1).} \tag{35.1}$$

Note the cyclic permutation of the vectors, which makes the formula easy to remember.

We can play the same game to define a seemingly different 3-form, $(\boldsymbol{\sigma}\wedge\boldsymbol{\Psi})$, but this turns out to be the *same* 3-form, $(\boldsymbol{\Psi}\wedge\boldsymbol{\sigma})$! To see why, simply move each $\boldsymbol{\sigma}(\mathbf{v}_i)$ to the front of each term on the right-hand side of (35.1), and observe that this preserves the cyclic permutation:

$$(\boldsymbol{\sigma}\wedge\boldsymbol{\Psi})(\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3)=\boldsymbol{\sigma}(\mathbf{v}_1)\,\boldsymbol{\Psi}(\mathbf{v}_2,\mathbf{v}_3)+\boldsymbol{\sigma}(\mathbf{v}_2)\,\boldsymbol{\Psi}(\mathbf{v}_3,\mathbf{v}_1)+\boldsymbol{\sigma}(\mathbf{v}_3)\,\boldsymbol{\Psi}(\mathbf{v}_1,\mathbf{v}_2). \tag{35.2}$$

Therefore, although the wedge product is *antisymmetric when it combines two 1-forms*, i.e.,

$$(\boldsymbol{\varphi}\wedge\boldsymbol{\psi})=-(\boldsymbol{\psi}\wedge\boldsymbol{\varphi}),$$

it is *symmetric when it combines a 2-form* $\boldsymbol{\Psi}$ *and a 1-form* $\boldsymbol{\sigma}$:

$$\boxed{\boldsymbol{\Psi}\wedge\boldsymbol{\sigma}=\boldsymbol{\sigma}\wedge\boldsymbol{\Psi}.}$$

Here is the general rule, which we will justify shortly:

$$\boxed{\textit{If }\ \boldsymbol{\Psi}\textit{ is a } p\textit{-form and }\boldsymbol{\Omega}\textit{ is a }q\textit{-form, then }\boldsymbol{\Psi}\wedge\boldsymbol{\Omega}=(-1)^{pq}\,\boldsymbol{\Omega}\wedge\boldsymbol{\Psi}\,.} \tag{35.3}$$

You may easily check that the previous two equations conform to this rule.

Assuming (35.3) to be true, we also deduce that if the degree of $\boldsymbol{\Omega}$ is *odd*, then [exercise],

$$\boxed{\boldsymbol{\Omega}\wedge\boldsymbol{\Omega}=0.} \tag{35.4}$$

Of course the wedge product of a 1-form with itself was our first instance of this phenomenon: $\boldsymbol{\psi}\wedge\boldsymbol{\psi}=0$.

35.3 The Volume 3-Form

Now let us take $\boldsymbol{\Psi}=\mathbf{d}x^1\wedge\mathbf{d}x^2$ and $\boldsymbol{\sigma}=\mathbf{d}x^3$, and define the ***volume 3-form*** to be

$$\boxed{\boldsymbol{\mathcal{V}}\equiv(\mathbf{d}x^1\wedge\mathbf{d}x^2)\wedge\mathbf{d}x^3.} \tag{35.5}$$

We now justify the naming of $\boldsymbol{\mathcal{V}}$.

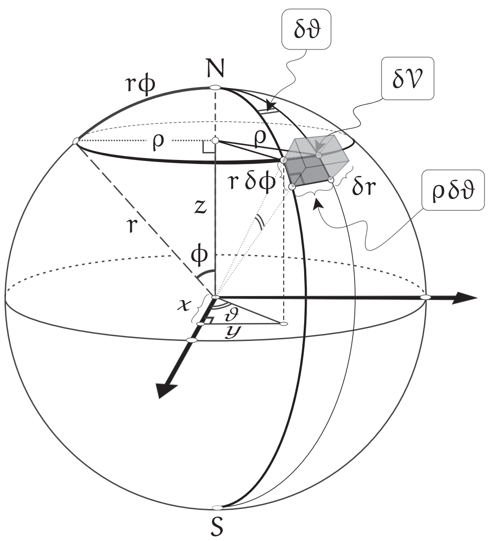
Look again at the standard formula (34.17) for the volume of a parallelepiped with edges $\mathbf{u}$, $\mathbf{v}$, and $\underline{\mathbf{\Omega}}$. If we apply our 3-form $\mathcal{V}$ to these three vectors, then (35.1) yields

$$\begin{aligned}
\mathcal{V}(\mathbf{u},\mathbf{v},\underline{\mathbf{\Omega}}) &= [(\mathbf{d}x^1\wedge\mathbf{d}x^2)\wedge\mathbf{d}x^3](\mathbf{u},\mathbf{v},\underline{\mathbf{\Omega}}) \\
&= (\mathbf{d}x^1\wedge\mathbf{d}x^2)(\mathbf{u},\mathbf{v})\,\mathbf{d}x^3(\underline{\mathbf{\Omega}}) + (\mathbf{d}x^1\wedge\mathbf{d}x^2)(\mathbf{v},\underline{\mathbf{\Omega}})\,\mathbf{d}x^3(\mathbf{u}) + (\mathbf{d}x^1\wedge\mathbf{d}x^2)(\underline{\mathbf{\Omega}},\mathbf{u})\,\mathbf{d}x^3(\mathbf{v}) \\
&= (u^1\,v^2 - u^2\,v^1)\,\Omega^3 + (v^1\,\Omega^2 - v^2\,\Omega^1)\,u^3 + (\Omega^1\,u^2 - \Omega^2\,u^1)\,v^3 \\
&= \Omega^1(u^2\,v^3 - u^3\,v^2) + \Omega^2(u^3\,v^1 - u^1\,v^3) + \Omega^3(u^1\,v^2 - u^2\,v^1) \\
&= \mathbf{\Omega}(\mathbf{u},\mathbf{v}) \\
&= \text{Volume of the parallelepiped with edges } \mathbf{u},\ \mathbf{v}, \text{ and } \underline{\mathbf{\Omega}}.
\end{aligned}$$

We shall see that the wedge product of these three 1-forms is *associative*, so the volume 3-form can be expressed as

$$\boxed{(\mathbf{d}x^1\wedge\mathbf{d}x^2)\wedge\mathbf{d}x^3 = \mathbf{d}x^1\wedge\mathbf{d}x^2\wedge\mathbf{d}x^3 = \mathbf{d}x^1\wedge(\mathbf{d}x^2\wedge\mathbf{d}x^3).}$$

35.4 The Volume 3-Form in Spherical Polar Coordinates



[35.1] *Geometric proof that* $\delta\mathcal{V} \asymp \delta r\ \ (r\,\delta\phi)\ (\rho\,\delta\vartheta) = r^2 \sin\phi\ \delta r\ \delta\phi\ \delta\vartheta$*, and derivation of formulas for* x*,* y*, and* z*.*

Once again, (35.5) reminds us of a classical expression: when doing a triple integral in space, we write the element of volume as $d\mathcal{V} = dx\ dy\ dz$. This connection can easily be made explicit, in the same way as before: applying $\mathbf{d}x\wedge\mathbf{d}y\wedge\mathbf{d}z$ to the vector edges of a small box with sides δx, δy, δz, we obtain its volume, $\delta\mathcal{V} = \delta x\ \delta y\ \delta z$.

If we switch to spherical polar coordinates, (r, ϕ, ϑ), then the classical element of volume becomes,

$$d\mathcal{V} = r^2\ \sin\phi\ dr\ d\phi\ d\vartheta.$$

Figure [35.1] contains a direct, geometrical proof, enabling us to *understand* this result. Note that this figure also provides the *formulas* for the coordinate change: since the distance of a point on the sphere from the NS, z-axis is $\rho = r\ \sin\phi$,

$$\begin{aligned}
x &= \rho\ \cos\vartheta = (r\ \sin\phi)\ \cos\vartheta, \\
y &= \rho\ \sin\vartheta = (r\ \sin\phi)\ \sin\vartheta, \\
z &= r\ \cos\phi.
\end{aligned}$$

Finally, note that the *order* (r, ϕ, ϑ) of the coordinates[1] has been chosen so that the movements resulting from small increases in the coordinates are *right handed*. Make sure you can see this.

---

[1]We are using the (logical!) convention of American mathematicians, whereby ϑ denotes the angle within the (x, y)-plane, *just as it does in* $\mathbb{R}^2$. However, British mathematicians, and essentially *all* scientists, interchange the meanings of ϑ and ϕ, so *their* order of the coordinates, and their volume 3-form, will appear to be different from ours. There are actually compelling historical reasons for adopting the physicists' choice—spherical harmonics immediately come to mind—and the conflict could be resolved if only we could *all* agree to change the angle in $\mathbb{R}^2$ to ϕ! Dray (2015) is a rare, brave adopter of this solution.

In the case of a more complicated change of coordinates, we would traditionally be *forced* to calculate the Jacobian. Well, no longer!

Let us use the example of spherical polar coordinates to demonstrate how (in more difficult cases) the volume 3-form saves us time, memorization, and even the need to *think!*

Well, a *little* thinking is certainly helpful: when multiplying out the wedge products below, we are *only* looking for ways to obtain $\mathbf{d}r \wedge \mathbf{d}\phi \wedge \mathbf{d}\vartheta$, because all other terms vanish. Towards the end of the calculation, we write $c_\vartheta \equiv \cos\vartheta$, and so forth, to save space, but we also recommend this as a clarifying notation in private, pencil-and-paper calculations.

$$\begin{aligned}
\mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}z &= \mathbf{d}(r\sin\phi\cos\vartheta) \wedge \mathbf{d}(r\sin\phi\sin\vartheta) \wedge \mathbf{d}(r\cos\phi) \\
&= [(\mathbf{d}r)\sin\phi\cos\vartheta + r\cos\phi\,\mathbf{d}\phi\cos\vartheta - r\sin\phi\sin\vartheta\,\mathbf{d}\vartheta] \\
&\wedge [(\mathbf{d}r)\sin\phi\sin\vartheta + r\cos\phi\,\mathbf{d}\phi\sin\vartheta + r\sin\phi\cos\vartheta\,\mathbf{d}\vartheta] \\
&\wedge [(\mathbf{d}r)\cos\phi - r\sin\phi\,\mathbf{d}\phi] \\
&= r^2\sin\phi\left[s_\phi^2\, c_\vartheta^2 + s_\phi^2\, s_\vartheta^2 + c_\phi^2\, s_\vartheta^2 + c_\phi^2\, c_\vartheta^2\right] \mathbf{d}r \wedge \mathbf{d}\phi \wedge \mathbf{d}\vartheta \\
&= r^2\sin\phi\,\mathbf{d}r \wedge \mathbf{d}\phi \wedge \mathbf{d}\vartheta.
\end{aligned}$$

Thinking geometrically enabled us to *understand* this result, while the "Devil's machine" only gave us the *answer*, and without any need for thought. Heed the warning that opened the Prologue: "...the danger to our soul is there, because when you pass over into algebraic calculation, essentially you stop thinking: you stop thinking geometrically, you stop thinking about the meaning."

These are but the first glimpses of the power of the "Devil's machine." You must therefore remain resolute, and on constant guard, lest you succumb to its temptations and thereby lose your soul!

35.5 The Wedge Product of Three 1-Forms and of p 1-Forms

To confirm the associativity we have just claimed, let us generalize the discussion, and define the wedge product of three *arbitrary* 1-forms. We wish to label the three 1-forms systematically, but we also wish to avoid any possible confusion with *components*, so let us call them $\underset{1}{\boldsymbol{\sigma}}, \underset{2}{\boldsymbol{\sigma}}, \underset{3}{\boldsymbol{\sigma}}$.

The formula for $\underset{1}{\boldsymbol{\sigma}} \wedge \underset{2}{\boldsymbol{\sigma}} \wedge \underset{3}{\boldsymbol{\sigma}}$ can be found by taking $\boldsymbol{\Psi} = \underset{1}{\boldsymbol{\sigma}} \wedge \underset{2}{\boldsymbol{\sigma}}$ in (35.1), but it is simpler to generalize the geometrical mapping (34.5) and the resulting geometrical interpretation given in [34.2].

That is, let us use the three 1-forms to define a mapping F of a vector $\mathbf{v}$ in $\mathbb{R}^n$ to a vector $F(\mathbf{v})$ in $\mathbb{R}^3$:

$$\boxed{\mathbf{v} \longrightarrow F(\mathbf{v}) \equiv \begin{bmatrix} \underset{1}{\boldsymbol{\sigma}}(\mathbf{v}) \\ \underset{2}{\boldsymbol{\sigma}}(\mathbf{v}) \\ \underset{3}{\boldsymbol{\sigma}}(\mathbf{v}) \end{bmatrix}} \tag{35.6}$$

Thus a parallelepiped in $\mathbb{R}^n$ with edges $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, is mapped by F to a parallelepiped in $\mathbb{R}^3$ with edges $F(\mathbf{v}_1)$, $F(\mathbf{v}_2)$, and $F(\mathbf{v}_3)$, and, by analogy with [34.2], we can define $\underset{1}{\boldsymbol{\sigma}} \wedge \underset{2}{\boldsymbol{\sigma}} \wedge \underset{3}{\boldsymbol{\sigma}}$ as the operator that yields the *volume* of this image parallelepiped:

$$\boxed{\left[\underset{1}{\boldsymbol{\sigma}}\wedge\underset{2}{\boldsymbol{\sigma}}\wedge\underset{3}{\boldsymbol{\sigma}}\right](\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3)=\det\begin{bmatrix} | & | & | \\ F(\mathbf{v}_1) & F(\mathbf{v}_2) & F(\mathbf{v}_3) \\ | & | & | \end{bmatrix}.} \tag{35.7}$$

Clearly, this can be immediately generalized to define the wedge product of p 1-forms:

$$\boxed{\left[\underset{1}{\boldsymbol{\sigma}}\wedge\cdots\wedge\underset{p}{\boldsymbol{\sigma}}\right](\mathbf{v}_1,\cdots,\mathbf{v}_p)=\det\begin{bmatrix} | & \cdots & | \\ F(\mathbf{v}_1) & \cdots & F(\mathbf{v}_p) \\ | & \cdots & | \end{bmatrix}.}$$

Equation (35.7) describes the *action* of a 3-form on three vectors, but can we abstract away the vectors to reveal the 3-form itself? Yes, but to do so, we need to deal directly with tensor products, just as we did in the original definition (34.3) of the wedge product of two 1-forms.

To achieve our goal, we must try to write down an expression that is antisymmetric under exchange of any pair of 1-forms. It is possible to do this [exercise] by working backwards from (35.7), but it is actually simpler to attack the problem head on. Try this yourself, *before* studying the answer, which is *here*:

$$\boxed{\begin{aligned}\underset{1}{\boldsymbol{\sigma}}\wedge\underset{2}{\boldsymbol{\sigma}}\wedge\underset{3}{\boldsymbol{\sigma}} \;=\;& \underset{1}{\boldsymbol{\sigma}}\otimes\underset{2}{\boldsymbol{\sigma}}\otimes\underset{3}{\boldsymbol{\sigma}}+\underset{2}{\boldsymbol{\sigma}}\otimes\underset{3}{\boldsymbol{\sigma}}\otimes\underset{1}{\boldsymbol{\sigma}}+\underset{3}{\boldsymbol{\sigma}}\otimes\underset{1}{\boldsymbol{\sigma}}\otimes\underset{2}{\boldsymbol{\sigma}}\\ &-\underset{2}{\boldsymbol{\sigma}}\otimes\underset{1}{\boldsymbol{\sigma}}\otimes\underset{3}{\boldsymbol{\sigma}}-\underset{3}{\boldsymbol{\sigma}}\otimes\underset{2}{\boldsymbol{\sigma}}\otimes\underset{1}{\boldsymbol{\sigma}}-\underset{1}{\boldsymbol{\sigma}}\otimes\underset{3}{\boldsymbol{\sigma}}\otimes\underset{2}{\boldsymbol{\sigma}}.\end{aligned}} \tag{35.8}$$

Note that the positive terms on the top line are the cyclic permutations of 123, whereas the negative terms on the second line are the noncyclic permutations. Each term on the second line is obtained by swapping the first two 1-forms in the cyclic term directly above it.

We leave it you to check that (35.8) is indeed equivalent to (35.7).

35.6 Basis 3-Forms

Just as $\{\mathbf{d}x^i\wedge\mathbf{d}x^j\}$ formed a basis for 2-forms, so $\{\mathbf{d}x^i\wedge\mathbf{d}x^j\wedge\mathbf{d}x^k\}$ forms a basis for 3-forms. In $\mathbb{R}^3$ there is only one such basis 3-form, namely, the volume 3-form, $\mathcal{V}=\mathbf{d}x^1\wedge\mathbf{d}x^2\wedge\mathbf{d}x^3$; all 3-forms in $\mathbb{R}^3$ are simply multiples of $\mathcal{V}$.

In $\mathbb{R}^4$, however, there are four distinct basis 3-forms, and a general 3-form $\boldsymbol{\Omega}$ can be decomposed into components as follows:

$$\begin{aligned}\boldsymbol{\Omega}=\;&\Omega_{123}\,(\mathbf{d}x^1\wedge\mathbf{d}x^2\wedge\mathbf{d}x^3)+\Omega_{234}\,(\mathbf{d}x^2\wedge\mathbf{d}x^3\wedge\mathbf{d}x^4)+\Omega_{341}\,(\mathbf{d}x^3\wedge\mathbf{d}x^4\wedge\mathbf{d}x^1)\\ &+\Omega_{412}\,(\mathbf{d}x^4\wedge\mathbf{d}x^1\wedge\mathbf{d}x^2).\end{aligned}$$

Let us now use such components to explain the rule (35.3). To make the argument vivid, let us take $p=2$ and $q=3$, so that $\boldsymbol{\Psi}$ is a 2-form and $\boldsymbol{\Omega}$ is a 3-form. When we form $\boldsymbol{\Psi}\wedge\boldsymbol{\Omega}$, a typical component $(\mathbf{d}x^{j_1}\wedge\mathbf{d}x^{j_2})$ of $\boldsymbol{\Psi}$ is multiplied by a typical component $(\mathbf{d}x^{k_1}\wedge\mathbf{d}x^{k_2}\wedge\mathbf{d}x^{k_3})$

of $\boldsymbol{\Omega}$. (We assume that we are in at least five dimensions, otherwise this product vanishes identically!)

Now we wish to transform this into a term of the *reversed* product, $\boldsymbol{\Omega} \wedge \boldsymbol{\Psi}$. To do so, we must begin by moving $\mathbf{d}x^{j_2}$ through the three 1-forms to its right, with each successive swap producing a change of sign:

$$
\begin{aligned}
(\mathbf{d}x^{j_1} \wedge \mathbf{d}x^{j_2}) \wedge (\mathbf{d}x^{k_1} \wedge \mathbf{d}x^{k_2} \wedge \mathbf{d}x^{k_3}) &= -\,\mathbf{d}x^{j_1} \wedge \mathbf{d}x^{k_1} \wedge \mathbf{d}x^{j_2} \wedge \mathbf{d}x^{k_2} \wedge \mathbf{d}x^{k_3} \\
&= +\,\mathbf{d}x^{j_1} \wedge \mathbf{d}x^{k_1} \wedge \mathbf{d}x^{k_2} \wedge \mathbf{d}x^{j_2} \wedge \mathbf{d}x^{k_3} \\
&= -\,\mathbf{d}x^{j_1} \wedge (\mathbf{d}x^{k_1} \wedge \mathbf{d}x^{k_2} \wedge \mathbf{d}x^{k_2}) \wedge \mathbf{d}x^{j_2}.
\end{aligned}
$$

With $q = 3$, this has resulted in a factor of $(-1)^3 = -1$, and in the general case the factor would be $(-1)^q$. To complete the reversal of $\boldsymbol{\Psi} \wedge \boldsymbol{\Omega}$ to create $\boldsymbol{\Omega} \wedge \boldsymbol{\Psi}$, we must now move $\mathbf{d}x^{j_1}$ through the 1-forms of $\boldsymbol{\Omega}$ so as to reform $(\mathbf{d}x^{j_1} \wedge \mathbf{d}x^{j_2})$ on the right. In so doing, we pick up a second factor of $(-1)^q$. In the general case, we must repeat this for each of the p 1-forms of $\boldsymbol{\Psi}$, picking up p factors of $(-1)^q$. Therefore, the net factor is $(-1)^{pq}$, which completes the proof of (35.3).

35.7 Is $\boldsymbol{\Psi} \wedge \boldsymbol{\Psi} \neq 0$ Possible?

One consequence of the rule (35.3), which we have now proved, is (35.4): if $\boldsymbol{\Psi}$ is of *odd* degree, then $\boldsymbol{\Psi} \wedge \boldsymbol{\Psi} = 0$.

Thus, if it were possible for $\boldsymbol{\Psi} \wedge \boldsymbol{\Psi} \neq 0$, then $\boldsymbol{\Psi}$ would have to be of *even* degree. The simplest possibility is therefore that $\boldsymbol{\Psi}$ is a 2-form, in which case $\boldsymbol{\Psi} \wedge \boldsymbol{\Psi}$ would be a 4-form. The smallest number of dimensions in which a 4-form can exist is 4, so let us take $\boldsymbol{\Psi}$ to be a 2-form in spacetime, with coordinates (t, x, y, z).

Consider,

$$\boldsymbol{\Psi} = \mathbf{d}t \wedge \mathbf{d}x + \mathbf{d}y \wedge \mathbf{d}z.$$

If $\mathcal{V}$ denotes the *volume 4-form* of spacetime, then [exercise],

$$\boldsymbol{\Psi} \wedge \boldsymbol{\Psi} = 2\, \mathbf{d}t \wedge \mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}z = 2\mathcal{V}.$$

Chapter 36

Differentiation

36.1 The Exterior Derivative of a 1-Form

In the context of Forms, a function f is taken to be a 0-form. When the ***exterior derivative*** $\mathbf{d}$ is applied to f, it yields the gradient 1-form $\mathbf{d}f$, which encodes *how fast* f *changes in all possible directions*.

If we wish to know the rate of change of f in a *specific* direction $\mathbf{u}$, we must submit it as a question to the oracle 1-form $\mathbf{d}f$, which then declares the answer to be (32.11):

$$\mathbf{d}f(\mathbf{u}) = \boldsymbol{\nabla}_{\mathbf{u}} f.$$

In summary, $\mathbf{d}$ measures the rate of change in all possible directions, and it increases the degree of the form by one, to allow for the input of an additional vector, along which to measure the rate of change of the form.

Expressed in this manner, our task is clear: we must extend the definition of $\mathbf{d}$ so that when it is applied to a p-form $\boldsymbol{\Psi}$ it yields a $(p+1)$-form $\mathbf{d}\boldsymbol{\Psi}$ that measures the rate of change of $\boldsymbol{\Psi}$ in all possible directions.

As is our custom, we shall begin with the simplest possible case, then gradually work our way up. So, let us try to differentiate a 1-form φ to obtain a 2-form $\mathbf{d}\varphi$, acting on pairs of vectors, $\mathbf{u}$ and $\mathbf{v}$.

In order for a rate of change to even make sense, we must imagine that the 1-form, and the vectors upon which it acts, are defined in at least a neighborhood of the point at which we seek the rate of change. Thus we take φ to be a 1-form *field*, and likewise take $\mathbf{u}$ and $\mathbf{v}$ to be vector *fields*.

Let us begin by imagining that $\mathbf{u}$ is the direction in which we want to know the rate of change of φ, the same role it played in $\mathbf{d}f(\mathbf{u}) = \boldsymbol{\nabla}_{\mathbf{u}} f$. Since $\mathbf{v}$ is the only other vector in play, the function whose rate of change will be sought can only be $\varphi(\mathbf{v})$. In other words, we are essentially forced to consider the rate of change along $\mathbf{u}$ of $\varphi(\mathbf{v})$:

$$\boldsymbol{\nabla}_{\mathbf{u}}\varphi(\mathbf{v}).$$

We are off to a good start, but we require that $\mathbf{d}\varphi$ be a *2-form*, so its action $\mathbf{d}\varphi(\mathbf{u}, \mathbf{v})$ on pairs of vectors must be *antisymmetric* in those vectors, which this expression is not. We therefore employ a trick that should be very familiar by now: we *impose* antisymmetry, by subtracting the same expression, but with $\mathbf{u} \leftrightarrow \mathbf{v}$:

$$\boldsymbol{\nabla}_{\mathbf{u}}\varphi(\mathbf{v}) - \boldsymbol{\nabla}_{\mathbf{v}}\varphi(\mathbf{u}). \tag{36.1}$$

This expression in now antisymmetric in the two vector inputs, so we are certainly getting closer, but this *still* cannot be the correct answer.

The problem is that $\mathbf{d}\varphi$ is supposed to measure the rate of change (in all directions) of the 1-form φ—the vector fields $\mathbf{u}$ and $\mathbf{v}$ should have nothing to do with this. And yet, even before we look into the matter in detail, it is clear that the variation of $\varphi(\mathbf{v})$, for example, will depend on both the variation of φ (which is the thing we care about) *and* on the variation of $\mathbf{v}$ (which we do *not* care about).

In order to define the exterior derivative $\mathbf{d}\varphi$, we must therefore discover how the variations of $\mathbf{u}$ and $\mathbf{v}$ contribute to (36.1), and then we must surgically *remove* their unwanted contributions!

In order to accomplish this, let us express the action of φ on a vector in terms of its vector counterpart, $\underline{\varphi}$, as we did in (34.15):

$$\varphi(\mathbf{u}) = \underline{\varphi} \cdot \mathbf{u}.$$

Then,

$$\begin{aligned}
\nabla_{\mathbf{u}}\varphi(\mathbf{v}) - \nabla_{\mathbf{v}}\varphi(\mathbf{u}) &= \nabla_{\mathbf{u}}(\underline{\varphi}\cdot\mathbf{v}) - \nabla_{\mathbf{v}}(\underline{\varphi}\cdot\mathbf{u}) \\
&= \left\{\mathbf{v}\cdot\nabla_{\mathbf{u}}\underline{\varphi} - \mathbf{u}\cdot\nabla_{\mathbf{v}}\underline{\varphi}\right\} + \left\{\underline{\varphi}\cdot\nabla_{\mathbf{u}}\mathbf{v} - \underline{\varphi}\cdot\nabla_{\mathbf{v}}\mathbf{u}\right\} \\
&= \left\{\mathbf{v}\cdot\nabla_{\mathbf{u}}\underline{\varphi} - \mathbf{u}\cdot\nabla_{\mathbf{v}}\underline{\varphi}\right\} + \left\{\underline{\varphi}\cdot[\mathbf{u},\mathbf{v}]\right\} \\
&= \left\{\mathbf{v}\cdot\nabla_{\mathbf{u}}\underline{\varphi} - \mathbf{u}\cdot\nabla_{\mathbf{v}}\underline{\varphi}\right\} + \varphi([\mathbf{u},\mathbf{v}]),
\end{aligned}$$

where $[\mathbf{u},\mathbf{v}]$ is the ***commutator***, pictured in [29.7], page 288.

The last term, $\varphi([\mathbf{u},\mathbf{v}])$, contains the variation in the vectors fields—*this* is what must be surgically removed, leaving only the *first*, bracketed term as our definition of the exterior derivative $\mathbf{d}\varphi$ of φ.

Having used the vector correspondence as a stepping stone, we can now dispense with it, and express the final result purely in terms of the 1-form φ:

$$\boxed{\mathbf{d}\varphi(\mathbf{u},\mathbf{v}) = \nabla_{\mathbf{u}}\varphi(\mathbf{v}) - \nabla_{\mathbf{v}}\varphi(\mathbf{u}) - \varphi([\mathbf{u},\mathbf{v}]).} \tag{36.2}$$

One advantage of this formula is that it is geometrically defined, independently of any choice of basis and coordinates. However, we also perceive two *problems* with this formula, one aesthetic, and one philosophical: (i) it is disappointingly complicated and (ii) it is very unclear what it *means*. We shall address the first problem immediately, and the second problem later in this chapter, and, at a deeper level, in the next chapter.

The key to the simplifying (36.2) is to abstract the vectors away entirely, leaving behind the 2-form $\mathbf{d}\varphi$ itself. Since we have gone to so much trouble to make (36.2) independent of the variations of the vector fields, let us simply *choose* the vector fields to be *constant*, in which case their commutator vanishes. Next, imagine that a basis has been chosen, enabling us to express the formula in terms of the Cartesian basis 1-forms, $\{\mathbf{d}x^k\}$.

In this case, writing $\varphi = \varphi_i\,\mathbf{d}x^i$, and using the abbreviation $\partial_k \equiv \partial_{x^k}$, we find that (36.2) reduces to,

$$\begin{aligned}
\mathbf{d}\varphi(\mathbf{u},\mathbf{v}) &= \nabla_{\mathbf{u}}\varphi(\mathbf{v}) - \nabla_{\mathbf{v}}\varphi(\mathbf{u}) \\
&= u^i\partial_i(v^j\,\varphi_j) - v^i\partial_i(u^j\,\varphi_j) \\
&= u^i v^j\,\partial_i\varphi_j - u^j v^i\,\partial_i\varphi_j \\
&= \partial_i\varphi_j\left\{\mathbf{d}x^i(\mathbf{u})\mathbf{d}x^j(\mathbf{v}) - \mathbf{d}x^j(\mathbf{u})\mathbf{d}x^i(\mathbf{v})\right\} \\
&= \partial_i\varphi_j\,(\mathbf{d}x^i\wedge\mathbf{d}x^j)(\mathbf{u},\mathbf{v}).
\end{aligned}$$

Therefore, abstracting away the arbitrary vector fields,

$$\boxed{\mathbf{d}\varphi = \partial_i\varphi_j\,(\mathbf{d}x^i\wedge\mathbf{d}x^j).} \tag{36.3}$$

Much better!

But we can do better still! Since $\mathbf{d}f = \partial_i f\,\mathbf{d}x^i$, we now deduce the following elegant, natural-looking, and highly *practical* expression for the action of $\mathbf{d}$:

$$\boxed{\mathbf{d}\varphi = \mathbf{d}\big(\varphi_j\,\mathbf{d}x^j\big) = \mathbf{d}\varphi_j \wedge \mathbf{d}x^j.} \tag{36.4}$$

36.2 The Exterior Derivative of a 2-Form and of a p-Form

The obvious next step is to try to discover the action of $\mathbf{d}$ on a *2-form*, $\boldsymbol{\Psi}$. This should be a 3-form, taking three vector fields as inputs: $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$. Clearly, $\mathbf{d}\boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ should be built out of elements such as $\boldsymbol{\nabla}_{\mathbf{v}_1}\boldsymbol{\Psi}(\mathbf{v}_2, \mathbf{v}_3)$, but how?

Since $\mathbf{d}$ increases the degree of a form by 1, it is analogous to a 1-form. The answer to the construction of $\mathbf{d}\boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ can therefore be found in our earlier work on the construction of the wedge product of a 1-form and a 2-form, as given in (35.2). Assuming constant vector fields from the outset, we therefore find that

$$\boxed{\mathbf{d}\boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \boldsymbol{\nabla}_{\mathbf{v}_1}\boldsymbol{\Psi}(\mathbf{v}_2, \mathbf{v}_3) + \boldsymbol{\nabla}_{\mathbf{v}_2}\boldsymbol{\Psi}(\mathbf{v}_3, \mathbf{v}_1) + \boldsymbol{\nabla}_{\mathbf{v}_3}\boldsymbol{\Psi}(\mathbf{v}_1, \mathbf{v}_2).}$$

Again, we can simplify this formula in a gratifying manner, by abstracting the vectors away to reveal the 3-form itself. We find [exercise] that (36.4) generalizes in a way that is clearly part of a larger pattern:

$$\boxed{\mathbf{d}\boldsymbol{\Psi} = \mathbf{d}\Big(\Psi_{ij}\,\mathbf{d}x^i \wedge \mathbf{d}x^j\Big) = \mathbf{d}\Psi_{ij} \wedge \mathbf{d}x^i \wedge \mathbf{d}x^j.} \tag{36.5}$$

It is now clear how $\mathbf{d}$ acts upon a p-form $\boldsymbol{\Phi}$ to create a $(p+1)$-form $\mathbf{d}\boldsymbol{\Phi}$:

$$\boxed{\mathbf{d}\boldsymbol{\Phi} = \mathbf{d}\Big(\Phi_{i_1 \ldots i_p}\,\mathbf{d}x^{i_1} \wedge \cdots \wedge \mathbf{d}x^{i_p}\Big) = \mathbf{d}\Phi_{i_1 \ldots i_p} \wedge \mathbf{d}x^{i_1} \wedge \cdots \wedge \mathbf{d}x^{i_p}.} \tag{36.6}$$

36.3 The Leibniz Rule for Forms

It is simple to verify [exercise] that if f is a function (0-form) then

$$\boxed{\mathbf{d}(f\,\boldsymbol{\Psi}) = (\mathbf{d}f) \wedge \boldsymbol{\Psi} + f\,\mathbf{d}\boldsymbol{\Psi}.} \tag{36.7}$$

However, the general Leibniz Rule (aka Product Rule) for Forms is much less obvious:

$$\boxed{\mathbf{d}(\boldsymbol{\Phi} \wedge \boldsymbol{\Psi}) = (\mathbf{d}\boldsymbol{\Phi}) \wedge \boldsymbol{\Psi} + (-1)^{\deg \boldsymbol{\Phi}}\,\boldsymbol{\Phi} \wedge (\mathbf{d}\boldsymbol{\Psi}),} \tag{36.8}$$

where deg $\boldsymbol{\Phi}$ denotes the degree of $\boldsymbol{\Phi}$.

For example, if deg $\mathbf{\Phi} = 1$, so that $\mathbf{\Phi} = \phi$ is a 1-form, then

$$\boxed{\mathbf{d}(\phi \wedge \mathbf{\Psi}) = (\mathbf{d}\phi) \wedge \mathbf{\Psi} - \phi \wedge (\mathbf{d}\mathbf{\Psi}).} \tag{36.9}$$

We first prove this simple example, for its proof will immediately generalize to provide an explanation of the full rule, (36.8).

To be concrete, suppose that $\mathbf{\Psi}$ is a 2-form. Then, using the Leibniz Rule for functions, (32.12), we find that,

$$\begin{aligned}
\mathbf{d}(\phi \wedge \mathbf{\Psi}) &= \mathbf{d}\left([\phi_i\, \mathbf{d}x^i] \wedge [\Psi_{jk}\, \mathbf{d}x^j \wedge \mathbf{d}x^k]\right) \\
&= \mathbf{d}(\phi_i\, \Psi_{jk}) \wedge \mathbf{d}x^i \wedge \mathbf{d}x^j \wedge \mathbf{d}x^k \\
&= (\Psi_{jk}\, \mathbf{d}\phi_i + \phi_i\, \mathbf{d}\Psi_{jk}) \wedge \mathbf{d}x^i \wedge \mathbf{d}x^j \wedge \mathbf{d}x^k \\
&= (\mathbf{d}\phi_i \wedge \mathbf{d}x^i) \wedge [\Psi_{jk}\, \mathbf{d}x^j \wedge \mathbf{d}x^k] + (\phi_i\, \mathbf{d}\Psi_{jk}) \wedge \mathbf{d}x^i \wedge \mathbf{d}x^j \wedge \mathbf{d}x^k \\
&= (\mathbf{d}\phi) \wedge \mathbf{\Psi} - [\phi_i \wedge \mathbf{d}x^i] \wedge [\mathbf{d}\Psi_{jk} \wedge \mathbf{d}x^j \wedge \mathbf{d}x^k] \\
&= (\mathbf{d}\phi) \wedge \mathbf{\Psi} - \phi \wedge (\mathbf{d}\mathbf{\Psi}).
\end{aligned}$$

The key to understanding the general formula (36.8) is the passage from the fourth to the fifth line, in which we pushed $\mathbf{d}\Psi_{jk}$ past $\mathbf{d}x^i$ (belonging to ϕ) in order to return it to its rightful place, next to its own $\mathbf{d}x^j \wedge \mathbf{d}x^k$, obtaining

$$\mathbf{d}x^i \wedge \mathbf{d}\Psi_{jk} = -\mathbf{d}\Psi_{jk} \wedge \mathbf{d}x^i.$$

In the general case, we must push $\mathbf{d}\Psi_{jk}$ past all the $\mathbf{d}x^i$'s belonging to $\mathbf{\Phi}$—numbering deg $\mathbf{\Phi}$—resulting in deg $\mathbf{\Phi}$ changes of sign. In other words, the sign of $\mathbf{\Phi} \wedge (\mathbf{d}\mathbf{\Psi})$ is $(-1)^{\deg \Phi}$, which completes the proof.

A helpful way of remembering the general formula is to think of $\mathbf{d}$ as a 1-form. Then, before $\mathbf{d}$ can *reach* $\mathbf{\Psi}$, on the right of $\mathbf{d}(\mathbf{\Phi} \wedge \mathbf{\Psi})$, *it must push past all the 1-forms of* $\mathbf{\Phi}$, resulting in deg $\mathbf{\Phi}$ changes of sign. As we have just seen, this mnemonic is actually quite close to the truth of the matter!

36.4 Closed and Exact Forms

36.4.1 A Fundamental Result: $\mathbf{d}^2 = 0$

Reconsider the formula (36.3):

$$\mathbf{d}\boldsymbol{\varphi} = \partial_i \varphi_j\, (\mathbf{d}x^i \wedge \mathbf{d}x^j).$$

Now let us take this 2-form and apply $\mathbf{d}$ for a *second* time, using (36.5):

$$\begin{aligned}
\mathbf{d}^2\boldsymbol{\varphi} &= \mathbf{d}\Big[\partial_i \varphi_j\, (\mathbf{d}x^i \wedge \mathbf{d}x^j)\Big] \\
&= \mathbf{d}\Big\{\partial_i \varphi_j\Big\} \wedge \mathbf{d}x^i \wedge \mathbf{d}x^j \\
&= \Big\{[\partial_k \partial_i \varphi_j]\, \mathbf{d}x^k \wedge \mathbf{d}x^i\Big\} \wedge \mathbf{d}x^j \\
&= 0,
\end{aligned}$$

because $\partial_k \partial_i \varphi_j = \partial_i \partial_k \varphi_j$, but $(\mathbf{d}x^k \wedge \mathbf{d}x^i) = -(\mathbf{d}x^i \wedge \mathbf{d}x^k)$.

By virtue of (36.6), it is clear that the above argument generalizes to arbitrary p-forms: $\mathbf{d}^2\boldsymbol{\Phi} = 0$. We may therefore abstract the p-form away, obtaining this absolutely fundamental result:

$$\text{Every form is annihilated by two applications of } \mathbf{d}\text{:} \qquad \boxed{\mathbf{d}^2 = 0.} \tag{36.10}$$

The computational *proof* is certainly short and clear, but what does the result itself actually *mean*? The resolution of this mystery must await our discussion of integration. (If you insist on ruining the suspense, see Section 37.4.)

36.4.2 Closed and Exact Forms

A form is called ***closed*** if its exterior derivative vanishes:

$$\boxed{\boldsymbol{\Upsilon} \text{ is } \textit{closed} \quad \Longleftrightarrow \quad \mathbf{d}\boldsymbol{\Upsilon} = 0.}$$

We shall eventually see that a closed form is a higher-dimensional analogue of the flux 2-form of an incompressible fluid flow.

A p-form is called ***exact*** if it is the exterior derivative of a $(p-1)$-form:

$$\boxed{\boldsymbol{\Upsilon} \text{ is } \textit{exact} \quad \Longleftrightarrow \quad \boldsymbol{\Upsilon} = \mathbf{d}\boldsymbol{\Psi} \quad (\text{for some} \boldsymbol{\Psi}).}$$

In the language of physics, we call $\boldsymbol{\Psi}$ a ***potential*** for $\boldsymbol{\Upsilon}$. If such a potential exists, it is far from unique, for if $\boldsymbol{\Theta}$ is an *arbitrary* $(p-2)$-form, and we change

$$\boldsymbol{\Psi} \rightsquigarrow \widetilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi} + \mathrm{d}\boldsymbol{\Theta},$$

then [exercise] $\widetilde{\boldsymbol{\Psi}}$ is *also* a potential: $\boldsymbol{\Upsilon} = \mathbf{d}\widetilde{\boldsymbol{\Psi}}$. This freedom in choosing the potential is called ***gauge freedom***, and the transformation $\boldsymbol{\Psi} \rightsquigarrow \widetilde{\boldsymbol{\Psi}}$ is called a ***gauge transformation***.

It follows immediately from (36.10) that

$$\boxed{\textit{Every exact form is closed:} \quad \boldsymbol{\Upsilon} = \mathbf{d}\boldsymbol{\Psi} \quad \Longrightarrow \quad \mathbf{d}\boldsymbol{\Upsilon} = 0}$$

In light of this, it is very natural to ask about a possible converse: *Is a closed form always exact?* This turns out to be a very interesting, deep question.

The short answer is that it depends on the *topology* of the region in which the form is defined.

> **Poincaré Lemma**. *If* $\mathbf{d}\boldsymbol{\Upsilon} = 0$ *throughout a simply connected region, then* $\boldsymbol{\Upsilon} = \mathbf{d}\boldsymbol{\Psi}$*, for some* $\boldsymbol{\Psi}$. (36.11)

Thus a potential *does* always exist *locally*. But global problems arise if the region is *not* simply connected, in which case it is possible that no such $\boldsymbol{\Psi}$ exists. The study of the closed Forms that are *not* exact then encodes detailed information about the topology of the space, called the ***de Rham cohomology***. This is the subject of Section 37.9.

Let us end this subsection with some practice with these concepts. We ask that you prove to yourself each of the following interesting facts:

If $\boldsymbol{\Upsilon}$ *and* $\boldsymbol{\Phi}$ *are closed, then* $\boldsymbol{\Upsilon} \wedge \boldsymbol{\Phi}$ *is closed, too.*

If $\boldsymbol{\Upsilon}$ *is closed, then* $\boldsymbol{\Upsilon} \wedge \mathbf{d}\boldsymbol{\Phi}$ *is closed for all* $\boldsymbol{\Phi}$.

If deg $\boldsymbol{\Phi}$ *is* even, *then* $\boldsymbol{\Phi} \wedge \mathbf{d}\boldsymbol{\Phi}$ *is closed.*

Hint: (35.4).

36.4.3 Complex Analysis: Cauchy–Riemann Equations

In VCA (pp. 216–217) we showed geometrically that a complex function $f(z) = u + iv$ is locally an amplitwist—as defined in Section 4.6—if and only if

$$\boxed{i\,\partial_x f = \partial_y f.} \tag{36.12}$$

Equating real and imaginary parts, we recover the more common form of the celebrated ***Cauchy–Riemann equations***, characterizing a conformal (aka complex analytic) mapping:

$$\boxed{\partial_x u = \partial_y v \qquad \text{and} \qquad \partial_x v = -\partial_y u.}$$

We now show that these equations may be elegantly (and consequentially) reformulated in the language of Forms.

Consider the complex-valued 1-form,

$$f\,\mathbf{d}z = (u + iv)(\mathbf{d}x + i\mathbf{d}y) = \left[u\,\mathbf{d}x - v\,\mathbf{d}y\right] + i\left[v\,\mathbf{d}x + u\,\mathbf{d}y\right].$$

Then,

$$\begin{aligned} \mathbf{d}\big(f\,\mathbf{d}z\big) &= \mathbf{d}f \wedge \mathbf{d}z \\ &= \big(\partial_x f\,\mathbf{d}x + \partial_y f\,\mathbf{d}y\big) \wedge \big(\mathbf{d}x + i\,\mathbf{d}y\big) \\ &= \big(i\,\partial_x f - \partial_y f\big)\,\mathcal{A}, \end{aligned}$$

where, as usual, $\mathcal{A} = \mathbf{d}x \wedge \mathbf{d}y$ is the area 2-form.

Therefore, by virtue of the compact form of the Cauchy–Riemann equations, (36.12), we conclude that

The complex function f *is locally an amplitwist if and only if the 1-form* $f\,\mathbf{d}z$ *is* closed*:* $\mathbf{d}\big(f\,\mathbf{d}z\big) = 0$. (36.13)

The central result regarding integration of Forms is called (in this book[1]) *The Fundamental Theorem of Exterior Calculus*; it is the main result of the next chapter. Armed with that theorem, we shall see (in Section 37.7) that (36.13) *instantly* implies the pivotal result of Complex Analysis—*Cauchy's Theorem!*

The above formulation is valuable even if we are interested in complex functions that are *not* locally an amplitwist (i.e., that are ***nonanalytic***). In particular, we note for future use that reflection across the real axis is *anticonformal*, so the mapping $f(z)=\overline{z}=x-iy$ is *not* subject to $\mathbf{d}(f\,\mathbf{d}z)=0$. Indeed, we find [exercise] that

$$\boxed{\mathbf{d}(\overline{z}\,\mathbf{d}z)=2i\,\boldsymbol{\mathcal{A}}.} \tag{36.14}$$

36.5 Vector Calculus via Forms

In this section we simply apply the exterior derivative to form fields in $\mathbb{R}^3$, and discover that the fundamental operations and identities of Vector Calculus emerge automatically, as if by magic—automagically!

As we explained in Sections 34.6 and 34.7, the exterior calculus works perfectly in *all* dimensions and on *all* p-forms, but *only* in $\mathbb{R}^3$ can a 2-form masquerade as a vector, and *only* in $\mathbb{R}^3$ can the wedge product of two 1-forms masquerade as the vector product of two vectors—these are the twin spells that conjure Vector Calculus into existence!

First, we remind the reader of the notation: $\underline{\boldsymbol{\varphi}}$ is the vector corresponding ($\rightleftarrows$) to the 1-form $\boldsymbol{\varphi}$, and $\underline{\boldsymbol{\Psi}}$ is the vector corresponding to the flux 2-form $\boldsymbol{\Psi}$, and $\boldsymbol{\mathcal{V}}=\mathbf{d}x^1\wedge\mathbf{d}x^2\wedge\mathbf{d}x^3$ is the volume 3-form.

We have seen that the Vector Calculus concept of the gradient vector $\boldsymbol{\nabla}f$ corresponds to the gradient 1-form $\mathbf{d}f$ of the 0-form f. Now let us step up one level, and take the exterior derivative of a 1-form, writing out explicitly the three resulting components of (36.3) that occur in $\mathbb{R}^3$; finally, let us *interpret* the answer as the flux 2-form of a flow in space:

$$\begin{aligned}
\mathbf{d}\varphi &= \partial_i\varphi_j\,(\mathbf{d}x^i\wedge\mathbf{d}x^j)\\
&= (\partial_2\varphi_3-\partial_3\varphi_2)\,(\mathbf{d}x^2\wedge\mathbf{d}x^3)+(\partial_3\varphi_1-\partial_1\varphi_3)\,(\mathbf{d}x^3\wedge\mathbf{d}x^1)+(\partial_1\varphi_2-\partial_2\varphi_1)\,(\mathbf{d}x^1\wedge\mathbf{d}x^2)\\
&\rightleftarrows \begin{bmatrix}\partial_2\varphi_3-\partial_3\varphi_2\\ \partial_3\varphi_1-\partial_1\varphi_3\\ \partial_1\varphi_2-\partial_2\varphi_1\end{bmatrix},
\end{aligned}$$

by virtue of (34.10).

Lo and behold, out has popped the ***curl*** of the vector field corresponding to the 1-form! To summarize,

$$\boxed{\mathbf{d}\varphi \quad \rightleftarrows \quad \boldsymbol{\nabla}\times\underline{\boldsymbol{\varphi}}=\mathbf{curl}\,\underline{\boldsymbol{\varphi}}=\begin{bmatrix}\partial_1\\ \partial_2\\ \partial_3\end{bmatrix}\times\begin{bmatrix}\varphi_1\\ \varphi_2\\ \varphi_3\end{bmatrix}.} \tag{36.15}$$

[1]The more common name is, instead, the Generalized Stokes's Theorem, the twisted origin of which is described in Section 37.3.2.

Thus if φ is *closed*, meaning $\mathbf{d}\varphi = 0$, then $\underline{\varphi}$ has vanishing curl: if $\underline{\varphi}$ is pictured as the velocity of a fluid flow in space, then if we insert a tiny ball—*not* a point particle—into the flow, it will be carried off with velocity $\underline{\varphi}$, but *it will not spin*. In this case the flow is called ***irrotational***. If $\mathbf{d}\varphi \neq 0$, then the ball will spin, and $\mathbf{curl}\,\underline{\varphi}$ points along the spin axis, and it magnitude gives double[2] the rate of spin. In this context, $\mathbf{curl}\,\underline{\varphi}$ is called the ***vorticity*** vector.

Alternatively, we may picture the closed form as corresponding to a ***conservative force field***, with the attendant advantage that the 1-form φ itself can immediately be given meaning, as the *work*, exactly as in the gravitational example of Section 32.3.1. There, φ was closed by virtue of the rather uninteresting reason that it was *constant*. In that gravitational case, the path-independent work done transporting a particle along a path from p to q could therefore be obtained simply by applying φ to the direct path along $\overrightarrow{pq}$.

In accord with the Poincaré lemma, in this gravitational example, the closed 1-form φ is *exact*: if h again represents height, so that gh is the potential energy of a unit mass, then $\varphi = \mathbf{d}(gh) = g\,\mathbf{d}h$. This connects naturally with the classical expression $g\,dh$, for if the particle is raised δh by moving it along the short vector $\mathbf{v}$, then the work done is

$$g\,\mathbf{d}h(\mathbf{v}) = g\,\delta h.$$

As we will discuss in the next chapter, for a *nonconstant* φ, the work done along a path must be evaluated by breaking the path down into lots of small vectors, applying φ to each one, and summing. As we shall discuss, if φ is *closed*, then the work done in transporting a particle around a closed loop vanishes.

Next, let us take the exterior derivative of a 2-form $\boldsymbol{\Psi}$, but labelling its component with the corresponding vector $\underline{\boldsymbol{\Psi}}$, as we did in (34.10):

$$\boldsymbol{\Psi} = \Psi^1\,(\mathbf{d}x^2 \wedge \mathbf{d}x^3) + \Psi^2\,(\mathbf{d}x^3 \wedge \mathbf{d}x^1) + \Psi^3\,(\mathbf{d}x^1 \wedge \mathbf{d}x^2) \quad \rightleftarrows \quad \underline{\boldsymbol{\Psi}} = \Psi^j\,\mathbf{e}_j = \begin{bmatrix} \Psi^1 \\ \Psi^2 \\ \Psi^3 \end{bmatrix}.$$

Then,

$$\begin{aligned}
\mathbf{d}\boldsymbol{\Psi} &= \mathbf{d}\Psi^1 \wedge \mathbf{d}x^2 \wedge \mathbf{d}x^3 + \mathbf{d}\Psi^2 \wedge \mathbf{d}x^3 \wedge \mathbf{d}x^1 + \mathbf{d}\Psi^3 \wedge \mathbf{d}x^1 \wedge \mathbf{d}x^2 \\
&= \partial_1\Psi^1\,\mathbf{d}x^1 \wedge \mathbf{d}x^2 \wedge \mathbf{d}x^3 + \partial_2\Psi^2\,\mathbf{d}x^2 \wedge \mathbf{d}x^3 \wedge \mathbf{d}x^1 + \partial_2\Psi^3\,\mathbf{d}x^3 \wedge \mathbf{d}x^1 \wedge \mathbf{d}x^2 \\
&= \left(\partial_1\Psi^1 + \partial_2\Psi^2 + \partial_2\Psi^3\right)\mathbf{d}x^1 \wedge \mathbf{d}x^2 \wedge \mathbf{d}x^3 \\
&= \left\{ \begin{bmatrix} \partial_1 \\ \partial_2 \\ \partial_3 \end{bmatrix} \bullet \begin{bmatrix} \Psi^1 \\ \Psi^2 \\ \Psi^3 \end{bmatrix} \right\} \mathcal{V}.
\end{aligned}$$

Lo and behold, out has popped the ***divergence*** of the vector field flow corresponding to the flux 2-form! To summarize,

$$\boxed{\mathbf{d}\boldsymbol{\Psi} = (\mathbf{div}\,\underline{\boldsymbol{\Psi}})\,\mathcal{V} = (\boldsymbol{\nabla} \bullet \underline{\boldsymbol{\Psi}})\,\mathcal{V}.} \tag{36.16}$$

Thus if $\boldsymbol{\Psi}$ is *closed*, meaning $\mathbf{d}\boldsymbol{\Psi} = 0$, then $\underline{\boldsymbol{\Psi}}$ has vanishing divergence, corresponding to the flow of an incompressible fluid in a region where no fluid is being pumped in or sucked out, so that exactly as much fluid flows out of a closed surface as flows into it. In general, applying $\mathbf{d}\boldsymbol{\Psi}$ to

[2]See Feynman et al. (1963, Vol. 1, §40).

a very small parallelepiped ultimately yields the *net flux of fluid out of it*. This will be explained in the next chapter.

The fundamental identity of exterior differentiation is (36.10):

$$\mathbf{d}^2 = 0,$$

and this has important implications for vector fields in $\mathbb{R}^3$. If in (36.15) we take $\varphi = \mathbf{d}\mathrm{f}$, then we instantly obtain a classical result of Vector Calculus:

$$\boxed{\mathbf{d}^2\varphi = 0 \qquad \Longleftrightarrow \qquad \mathrm{curl}\left(\mathrm{grad}\,\mathrm{f}\right) = \boldsymbol{\nabla} \times \boldsymbol{\nabla}\mathrm{f} = 0.}$$

If in (36.16) we take $\boldsymbol{\Psi} = \mathbf{d}\varphi$, then we instantly obtain a second classical result of Vector Calculus:

$$\boxed{\mathbf{d}^2\,\boldsymbol{\Psi} = 0 \qquad \Longleftrightarrow \qquad \mathrm{div}\left(\mathrm{curl}\,\underline{\varphi}\right) = \boldsymbol{\nabla}\bullet\left(\boldsymbol{\nabla} \times \underline{\varphi}\right) = 0.}$$

Working with Forms obviates the need to remember these identities, as well as the many more complicated ones that are routinely employed by working scientists when using Vector Calculus. Furthermore, if we *do* wish to employ such an identity, it may be *derived* using Forms, and often much more quickly and elegantly than by direct vector methods. We illustrate this point with two examples of differential identities, leaving others to Exercise 14.

First, however, we shall state a very useful *algebraic* link between the world of Forms and the world of ($\mathbb{R}^3$) vectors, leaving the simple proof to you:

$$\boxed{\varphi \wedge \boldsymbol{\Psi} = (\underline{\varphi}\bullet\underline{\boldsymbol{\Psi}})\,\mathcal{V}} \tag{36.17}$$

As our first example, suppose we wish to find an identity for $\boldsymbol{\nabla}\bullet\left[\mathrm{f}\,\underline{\boldsymbol{\Psi}}\right]$. According to (36.16), we should therefore evaluate $\mathbf{d}\left[\mathrm{f}\,\boldsymbol{\Psi}\right]$, using (36.7), like this:

$$\left(\boldsymbol{\nabla}\bullet\left[\mathrm{f}\,\underline{\boldsymbol{\Psi}}\right]\right)\mathcal{V} = \mathbf{d}\left[\mathrm{f}\,\boldsymbol{\Psi}\right] = \left(\mathbf{d}\mathrm{f}\right)\wedge\boldsymbol{\Psi} + \mathrm{f}\,\mathbf{d}\boldsymbol{\Psi} = \left[(\boldsymbol{\nabla}\mathrm{f})\bullet\underline{\boldsymbol{\Psi}} + \mathrm{f}\,\boldsymbol{\nabla}\bullet\underline{\boldsymbol{\Psi}}\right]\mathcal{V},$$

yielding the familiar identity,

$$\boxed{\boldsymbol{\nabla}\bullet\left[\mathrm{f}\,\underline{\boldsymbol{\Psi}}\right] = (\boldsymbol{\nabla}\mathrm{f})\bullet\underline{\boldsymbol{\Psi}} + \mathrm{f}\,\boldsymbol{\nabla}\bullet\underline{\boldsymbol{\Psi}}.}$$

In this case, it is almost as easy [exercise] to derive the identity directly. So let us move on to an example in which the advantages of Forms start to emerge clearly.

Suppose we wish to find an identity for $\boldsymbol{\nabla} \times \left[\mathrm{f}\,\underline{\varphi}\right]$. According to (36.15), this corresponds to the flux 2-form

$$\mathbf{d}\left[\mathrm{f}\,\varphi\right] = \mathbf{d}\mathrm{f}\wedge\varphi + \mathrm{f}\,\mathbf{d}\varphi,$$

so we *effortlessly* deduce another familiar identity of Vector Calculus:

$$\boxed{\boldsymbol{\nabla} \times \left[\mathrm{f}\,\underline{\varphi}\right] = \boldsymbol{\nabla}\mathrm{f} \times \underline{\varphi} + \mathrm{f}\,\boldsymbol{\nabla} \times \underline{\varphi}.}$$

As the complexity of the vector identity increases, so does the simplifying power of Forms; see Exercise 14.

36.6 Maxwell's Equations

We end this chapter with a beautiful and *profound* application of the above ideas: the encapsulation of Maxwell's complete laws of electromagnetism in just *two*, supremely beautiful equations.

We begin with the source-free equations, (34.24), which we restate here, for your convenience:

$$\begin{aligned} \nabla \bullet \underline{\mathbf{B}} &= 0, \\ \nabla \times \underline{\mathbf{E}} + \partial_t \underline{\mathbf{B}} &= 0. \end{aligned}$$

Following the example of Baez and Muniain (1994, §5), let us use $\mathbf{d}_S$ to denote the *spatial* part of the spacetime exterior derivative, so that

$$\mathbf{d}f = \mathbf{d}_S f + \partial_t f \, \mathbf{d}t.$$

Now consider the exterior derivative of the Faraday 2-form, (34.22):

$$\mathbf{dF} = \mathbf{d}(\varepsilon \wedge \mathbf{d}t) + \mathbf{dB}.$$

Let us evaluate these two terms separately.

First,

$$\begin{aligned} \mathbf{d}(\varepsilon \wedge \mathbf{d}t) &= \mathbf{d}(\varepsilon) \wedge \mathbf{d}t \\ &= \mathbf{d}_S(\varepsilon) \wedge \mathbf{d}t \\ &= \Big[\text{flux 2-form of } \boldsymbol{\nabla} \times \underline{\mathbf{E}}\Big] \wedge \mathbf{d}t, \end{aligned}$$

by virtue of (36.15).

Second,

$$\begin{aligned} \mathbf{dB} &= \mathbf{d}_S \mathbf{B} + \Big[\partial_t B_x \, \mathbf{d}t \wedge (\mathbf{d}y \wedge \mathbf{d}z) + \cdots\Big] \\ &= (\boldsymbol{\nabla} \bullet \underline{\mathbf{B}}) \, \mathcal{V} + \mathbf{d}t \wedge \Big[\partial_t B_x \, (\mathbf{d}y \wedge \mathbf{d}z) + \cdots\Big] \\ &= (\boldsymbol{\nabla} \bullet \underline{\mathbf{B}}) \, \mathcal{V} + \mathbf{d}t \wedge \partial_t \mathbf{B} \end{aligned}$$

by virtue of (36.16).

But $\mathbf{d}t \wedge (\partial_t \mathbf{B}) = (\partial_t \mathbf{B}) \wedge \mathbf{d}t$, by virtue of (35.3), because $\partial_t \mathbf{B}$ is a 2-form. Therefore, since

$$\partial_t \mathbf{B} = \text{flux 2-form of } \partial_t \underline{\mathbf{B}},$$

combining the previous two results yields

$$\mathbf{dF} = (\boldsymbol{\nabla} \bullet \underline{\mathbf{B}}) \, \mathcal{V} + \Big\{\text{flux 2-form of } \Big[\boldsymbol{\nabla} \times \underline{\mathbf{E}} + \partial_t \underline{\mathbf{B}}\Big]\Big\} \wedge \mathbf{d}t.$$

Therefore, reaching upward, we may now touch the face of God:

> *Maxwell's Source-Free Equations state that the Faraday 2-form is* closed*:*
>
> $$\mathbf{dF} = 0.$$

As we have stated, and will discuss further in the next chapter, *locally*, every closed form is exact. In other words, there exists a 1-form[3] potential, $\mathbf{A}$, such that

$$\boxed{\boxed{\mathbf{F} = \mathbf{d}\mathbf{A}.}}$$

NOTE: The classical literature on electrodynamics instead employs the corresponding *vector field* $\underline{\mathbf{A}}$, which is called the ***vector potential***.

Next, we turn our attention to the second pair of Maxwell's Equations, (34.25), which describe the fields generated by *sources*—an electrical charge density ρ, and a current density $\underline{\mathbf{j}}$:

$$\begin{aligned} \nabla \bullet \underline{\mathbf{E}} &= 4\pi\rho, \\ \nabla \times \underline{\mathbf{B}} - \partial_t \underline{\mathbf{E}} &= 4\pi \underline{\mathbf{j}}. \end{aligned}$$

In order to see what these equations are truly trying to show us, we must first introduce the *spacetime* version of the sources, combined into a single 1-form:

$$\mathbf{J} = -\rho\, \mathbf{d}t + \mathbf{j},$$

where the (spatial) 1-form[4] $\mathbf{j}$ corresponds to the (spatial) vector current density $\underline{\mathbf{j}}$:

$$\underline{\mathbf{j}} = \begin{bmatrix} j^1 \\ j^2 \\ j^3 \end{bmatrix} \qquad \rightleftarrows \qquad \mathbf{j} = j^1\, \mathbf{d}x^1 + j^2\, \mathbf{d}x^2 + j^3\, \mathbf{d}x^3.$$

Thus when $\mathbf{J}$ acts on a purely spatial 4-vector, it yields the amount of current flowing in that spatial direction, but when it acts on the time-axis 4-velocity of a stationary observer, it describes the charge flowing with the observer in time, i.e., the charge density ρ at the observer's location. More generally, if $\mathbf{u}$ is the 4-velocity of an observer, then she measures the charge density to be $\mathbf{J}(\mathbf{u})$.

In Exercise 15 you will see that in 4-dimensional spacetime, the Hodge star operator ($\star$) maps a p-form to a ***dual*** $(4-p)$-form. In particular, the dual of the source density 1-form $\mathbf{J}$ is the 3-form,

$$\star\mathbf{J} = -\rho\, \mathcal{V} + \left[\text{flux 2-form of } \underline{\mathbf{j}}\right] \wedge \mathbf{d}t, \tag{36.18}$$

and the dual of the Faraday 2-form is the Maxwell 2-form, (34.26):

$$\star\mathbf{F} = \mathbf{Maxwell} = \boldsymbol{\beta} \wedge \mathbf{d}t - \mathbf{E}.$$

Since Maxwell can be obtained from Faraday by interchanging the electric and magnetic fields, *and* changing the sign of the first term, the entire calculation above of $\mathbf{dF}$ may be coopted to simply write down the answer for $\mathbf{d}\star\mathbf{F}$:

$$\begin{aligned} \mathbf{d}\star\mathbf{F} &= -(\boldsymbol{\nabla}\bullet\underline{\mathbf{E}})\,\mathcal{V} + \left\{\text{flux 2-form of } \left[\boldsymbol{\nabla}\times\underline{\mathbf{B}} - \partial_t\underline{\mathbf{E}}\right]\right\} \wedge \mathbf{d}t \\ &= 4\pi\left[-\rho\,\mathcal{V} + \left\{\text{flux 2-form of } \underline{\mathbf{j}}\right\} \wedge \mathbf{d}t\right], \end{aligned}$$

by virtue of the two Maxwell source equations above.

[3]This notation is universal, so in this case we must abandon our convention of using lowercase Greek letters to represent 1-forms.

[4]This notation is universal, so, once again, we must abandon our convention of using lowercase Greek letters to represent 1-forms.

We have arrived at a very remarkable—and a very *beautiful*—conclusion:

The Maxwell 2-form is subject to this Law of Nature:

$$\mathbf{d} \star \mathbf{F} = 4\pi \star \mathbf{J}.$$

For further discussion of Maxwell's equations in the language of Forms, see Hubbard and Hubbard (2009, §6.11), Baez and Muniain (1994, §§1,5), and Misner, Thorne, and Wheeler (1973, § 4).

Chapter 37

Integration

37.1 The Line Integral of a 1-Form

37.1.1 Circulation and Work

The *circulation* $\mathcal{C}_{\mathsf{K}}(\underline{\varphi})$ along a directed curve K of a fluid flow with velocity $\underline{\varphi}$ was previously defined ((27.2), p. 263) as the integral of the component of the fluid flow *along* K:

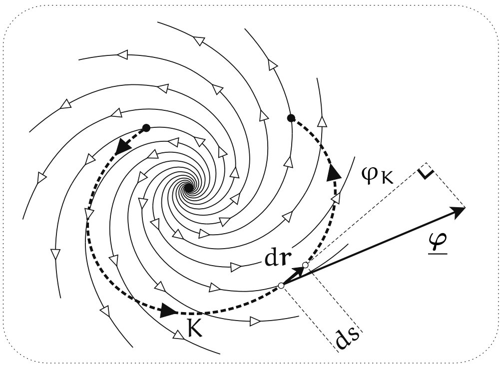
$$\mathcal{C}_{\mathsf{K}}(\underline{\varphi}) \equiv \int_{\mathsf{K}} \underline{\varphi} \bullet \mathbf{dr},$$

as illustrated in [37.1]. Alternatively, and equally importantly, we may think of $\underline{\varphi}$ as a *force field*, in which case exactly the same integral is now interpreted as the *work* done by the force field in moving a particle along K.

We have chosen to illustrate this idea in $\mathbb{R}^2$ in order to keep things simple, but we should stress that there is nothing intrinsically 2-dimensional about this concept: imagine the fluid to be the ocean, or the force field to be gravity. Indeed, everything we are about to discuss applies (without change) to 1-forms acting on vectors in $\mathbb{R}^n$.

Translating this into the language of Forms, we may write,

$$\varphi(\mathbf{dr}) = \underline{\varphi} \bullet \mathbf{dr} = \varphi_{\mathsf{K}}\, \mathrm{ds}.$$



[37.1] *The work done in moving a particle a distance* ds *along the path* K *is* $\varphi_{\mathsf{K}}\,\mathrm{ds}$*, where* φ_{K} *is the component of the force field* $\underline{\varphi}$ *in the direction of* K.

Here, as illustrated in [37.1], φ_{K} is the projection of $\underline{\varphi}$ onto the direction of K, and $\mathrm{ds} = |\mathbf{dr}|$. In other words, the 1-form φ acts on a small piece[1] $\mathbf{dr}$ of the directed curve, yielding the component of the force in the direction of the movement, times the length of that small movement, which is the work done in carrying out that movement. The complete integral may then be thought of as the total work done, obtained by summing the increments of work, as the 1-form φ eats its way along the directed curve K, one small $\mathbf{dr}$ bite at a time.

We stress that *only the component of the force in the direction of movement does any work*. Thus, for example, if K were, instead, an orthogonal trajectory of the field lines, then no work would be done, i.e., the integral would vanish.

[1]Here we are being lazy: we *should* write $\delta\mathbf{r}$, and use ultimate equalities.

Thought of in this way, we have arrived at our definition of the integral of 1-form φ along K, written

$$\boxed{\boxed{\int_K \varphi \equiv \mathcal{C}_K(\underline{\varphi}) \equiv \int_K \varphi_K \, ds.}} \tag{37.1}$$

In this definition, the absence (on the left) of d (*something*) under the integral sign is unfamiliar and perhaps disconcerting at first: just remember that this is not needed, since it is *understood* that φ is going to eat its way along K, one small $d\mathbf{r}$ bite at a time.

In the illustrated case, it is clear that K is always moving in roughly the same direction as the flow or force field, so the integral is clearly positive. But suppose we traverse the *same curve, but in the opposite direction* denoted $-K$. Now we are swimming upstream against the current, or, using the other interpretation, we are pushing the particle against the resistance of the force field. Thus the integral is now negative. More precisely, since $\varphi(-d\mathbf{r}) = -\varphi(d\mathbf{r})$, it follows that

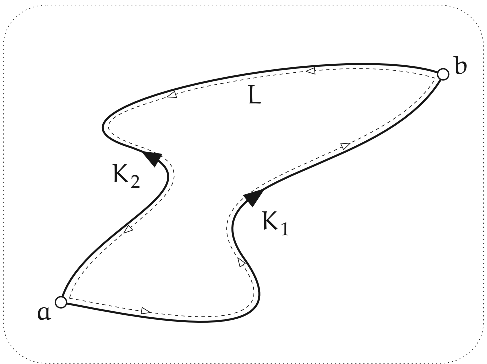
$$\boxed{\int_{-K} \varphi = -\int_K \varphi.} \tag{37.2}$$

37.1.2 Path-Independence $\iff$ Vanishing Loop Integrals

If K connects two fixed points a and b, then, in general, the value of $\mathcal{C}_K(\underline{\varphi})$ will depend on the specific path K from a to b that is chosen. However, for many important physical fluid flows and force fields, we have ***path independence***, which means what it says, and says what it means: all paths K from a to b yield the *same* value of the integral.

This concept can be rephrased in an important and useful way. As illustrated in [37.2], consider any two paths K_1 and K_2 connecting a to b. Then we can create a *closed loop* L by first travelling along K_1 from a to b, and then travelling back from b to a along $-K_2$, so that $L = K_1 - K_2$.

But then path-independence implies that the loop integral *vanishes*:



[37.2] *If the integral is independent of the path K between a and b, then the integral around the closed loop L vanishes. Conversely, if all loop integrals vanish, then the integral is path-independent.*

$$\int_L \varphi = \int_{K_1 - K_2} \varphi = \int_{K_1} \varphi - \int_{K_2} \varphi = 0.$$

Conversely, suppose that we know that all loop integrals vanish. It follows that the integral is path-independent. For if K_1 and K_2 are any two paths connecting a to b, then we may form the closed loop $L = K_1 - K_2$, and the fact that $\mathcal{C}_L(\underline{\varphi}) = 0$ then implies

$$\int_{K_1} \varphi = \int_{K_2} \varphi.$$

In brief,

$$\boxed{\textit{Path-independence} \iff \textit{vanishing loop integrals.}} \tag{37.3}$$

Look again at [37.1], and imagine that we extend K so that it closes up to become an elliptical loop encircling the vortex. Clearly the integral around this loop does *not* vanish, and is in fact positive. Thus here we have an example in which the integral of φ *does* depend on the path connecting two points.

37.1.3 The Integral of an Exact Form: $\varphi = \mathbf{d}f$

If φ is *exact* then, by definition, $\varphi = \mathbf{d}f$, for some function f. If $\delta\mathbf{r}$ is a short, ultimately vanishing, movement along the integration path K, then its contribution to the integral is

$$\varphi(\delta\mathbf{r}) = \mathbf{d}f(\delta\mathbf{r}) \asymp \delta f \equiv \text{change in f from the tail to the tip of } \delta\mathbf{r}.$$

Thus the complete integral is imply the net change in f from the start of K to the end of K. To summarize,

> *The integral of an exact 1-form $\varphi = \mathbf{d}f$ is path-independent, and is given by the change in f along K:*
> $$\int_K \mathbf{d}f = f(b) - f(a).$$ (37.4)

In the language of physics, $\varphi = \mathbf{d}f$ corresponds to a ***conservative*** force field, $\underline{\varphi} = \nabla f$, and f is the ***potential energy***. Then the change in potential energy $[f(b) - f(a)]$ is the path-independent amount of work done by the force field in carrying the particle from a to b, so (37.4) is simply another way of writing the familiar fact that

$$\int_K (\nabla f) \bullet \mathbf{dr} = f(b) - f(a). \tag{37.5}$$

As discussed by Feynman et al. (1963, Vol. 1, §14-4), at the most fundamental level *all* the forces in Nature are conservative, in accordance with the fundamental *principle of conservation of energy*.

37.2 The Exterior Derivative as an Integral

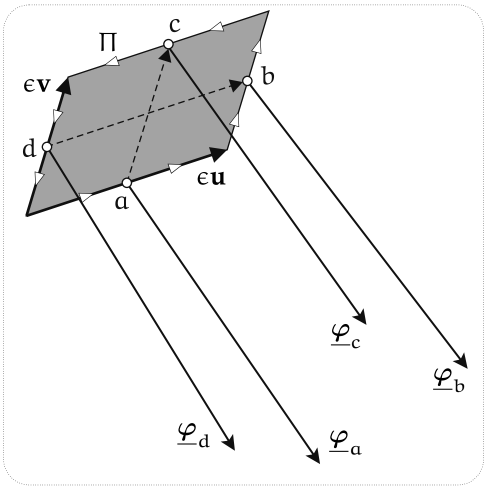
37.2.1 d(1-Form)

From our very first exposure to calculus, we are taught that differentiation and integration are *inverse operations*—this is the Fundamental Theorem of Calculus. Indeed, we have just seen (or so it would seem) a fresh instance of this idea in (37.4).

It may therefore come as disorienting shock to learn that, when it comes to Forms, exterior differentiation *is* integration—around a small closed loop, or over a small closed surface—*not* its inverse!

Let $\Pi(\epsilon\mathbf{u}, \epsilon\mathbf{v})$ be the oriented boundary of a small—ultimately vanishing, as ϵ vanishes—parallelogram with first edge $\epsilon\mathbf{u}$ and second edge $\epsilon\mathbf{v}$, as illustrated in [37.3]. Let us define $\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v})$ to be the integral of the 1-form φ around Π:

$$\boxed{\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}) \equiv \oint_{\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v})} \varphi.}$$



[37.3] *The integral of the 1-form φ around the small parallelogram Π is ultimately equal to its exterior derivative, $\mathbf{d}\varphi$, applied to the edges of Π.*

We will now show that this integral is a 2-*form*, and not just any 2-form—it is the *exterior derivative* of φ! That is,

$$\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}) \asymp \mathbf{d}\varphi(\epsilon\mathbf{u}, \epsilon\mathbf{v}).$$

Abstracting the vectors away,

$$\boxed{\boldsymbol{\Omega} \asymp \mathbf{d}\varphi.} \tag{37.6}$$

That $\boldsymbol{\Omega}$ does indeed possess the antisymmetry required of a 2-form follows immediately from the fact that $\Pi(\epsilon\mathbf{v}, \epsilon\mathbf{u}) = -\Pi(\epsilon\mathbf{u}, \epsilon\mathbf{v})$, for this implies that

$$\boldsymbol{\Omega}(\epsilon\mathbf{v}, \epsilon\mathbf{u}) = \oint_{\Pi(\epsilon\mathbf{v},\epsilon\mathbf{u})} \varphi = \oint_{-\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v})} \varphi = -\oint_{\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v})} \varphi = -\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}).$$

It is also possible to confirm geometrically that the linearity requirements are met, directly from the definition, but we shall instead cut directly to the chase, proving the *full* result, (37.6), thereby confirming linearity, and (re)confirming antisymmetry, in the process. To do so, we need only generalize the argument we previously gave in [27.2], page 263.

First, however, let us try to picture a flag flying over our ultimate destination, so that we have some idea in which direction we should march. In order for $\Pi(\epsilon\mathbf{u}, \epsilon\mathbf{v})$ to be a *closed* parallelogram loop, it is necessary—as explained in [29.7], page 288, and again in the upcoming [37.4]—that the *commutator* of its sides vanishes:

$$[\epsilon\mathbf{u}, \epsilon\mathbf{v}] = 0.$$

This is readily achieved by choosing the vector fields to be constant. In this event, our very first expression for the exterior derivative, (36.2), reduces to

$$\mathbf{d}\varphi(\epsilon\mathbf{u}, \epsilon\mathbf{v}) = \boldsymbol{\nabla}_{\epsilon\mathbf{u}}\, \varphi(\epsilon\mathbf{v}) - \boldsymbol{\nabla}_{\epsilon\mathbf{v}}\, \varphi(\epsilon\mathbf{u}), \tag{37.7}$$

and *this* is our flag.

With our objective clear, let us return to the evaluation of the integral $\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v})$. Since we will ultimately shrink the parallelogram to nothing by letting ϵ go to zero, we may use a Riemann sum with just a single term per side of the parallelogram. Usually a Riemann sum is used to evaluate an integral along some fixed interval, breaking it down into a larger and larger number of smaller and smaller subintervals, so that the precise point at which we evaluate φ on each subinterval becomes irrelevant in this limit.

However, that is *not* what we are doing here: we seek the *limiting form*—pun intended—of the integral as ϵ vanishes, *and the integral vanishes with it*. If we simply pick a *random* point on each side, then the error dies away as ϵ^2, and that is simply not good enough. For it is intuitively clear that the contributions to Ω from opposite sides of Π will *cancel* to order ϵ, so that Ω itself will be of order ϵ^2—the *same* order as the error incurred in a random Riemann sum.

The answer is to instead choose the *midpoints* of the sides, for then *the error incurred dies away much faster—in fact as* ϵ^3. (This is proved geometrically in VCA, p. 382.) By using midpoints, we therefore immediately attain the accuracy we require[2] to find a net result for Ω that is of order ϵ^2.

Therefore, as illustrated in [37.3], let us approximate $\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v})$ using these midpoints, a, b, c, d, and let the corresponding vector at a be $\underline{\varphi}_a$, and so on, as illustrated. Then, taking all derivatives to be evaluated at the centre of the parallelogram,

$$\begin{aligned}
\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}) &\asymp \varphi_a(\epsilon\mathbf{u}) + \varphi_b(\epsilon\mathbf{v}) + \varphi_c(-\epsilon\mathbf{u}) + \varphi_d(-\epsilon\mathbf{v}) \\
&= \Big[\varphi_b(\epsilon\mathbf{v}) - \varphi_d(\epsilon\mathbf{v})\Big] - \Big[\varphi_c(\epsilon\mathbf{u}) - \varphi_a(\epsilon\mathbf{u})\Big] \\
&\asymp \boldsymbol{\nabla}_{\epsilon\mathbf{u}}\, \varphi(\epsilon\mathbf{v}) - \boldsymbol{\nabla}_{\epsilon\mathbf{v}}\, \varphi(\epsilon\mathbf{u}) \\
&= \mathbf{d}\varphi(\epsilon\mathbf{u}, \epsilon\mathbf{v}),
\end{aligned}$$

by virtue of (37.7). Abstracting away the vectors, this completes the proof of (37.6).

We stress that there is nothing intrinsically 2-dimensional about the preceding argument: picture the $\underline{\varphi}$ vectors in [37.3] as sticking out of the page of the book, and pick up and tilt the book so that Π takes up any position in space!

If we now think of the vector field $\underline{\varphi}$ as a fluid flow in space, then the result (37.6) attaches a wonderfully concrete and vivid *meaning* to the exterior derivative:

> *If fluid flows with velocity $\underline{\varphi}$, then its circulation Ω around a small (ultimately vanishing) parallelogram is ultimately equal to the result of applying the flux 2-form $\mathbf{d}\varphi$ to its edges.* (37.8)

For future use, let us express this crucial result as an explicit formula in the language of classical Vector Calculus. Let $\widehat{\mathbf{n}}$ be the unit normal to the parallelogram, its direction determined by the right-hand rule, curling our fingers from $\mathbf{u}$ to $\mathbf{v}$. Also, let $\delta\mathcal{A}$ be the area of the parallelogram:

$$\delta\mathcal{A} = |\epsilon\, \mathbf{u} \times \epsilon\, \mathbf{v}|,$$

so that

$$\epsilon\, \mathbf{u} \times \epsilon\, \mathbf{v} = \widehat{\mathbf{n}}\, \delta\mathcal{A}.$$

The flux of $\boldsymbol{\nabla} \times \underline{\varphi}$ through the parallelogram is its component perpendicular to the parallelogram (i.e., its component in the direction of $\widehat{\mathbf{n}}$) multiplied by the area of the parallelogram:

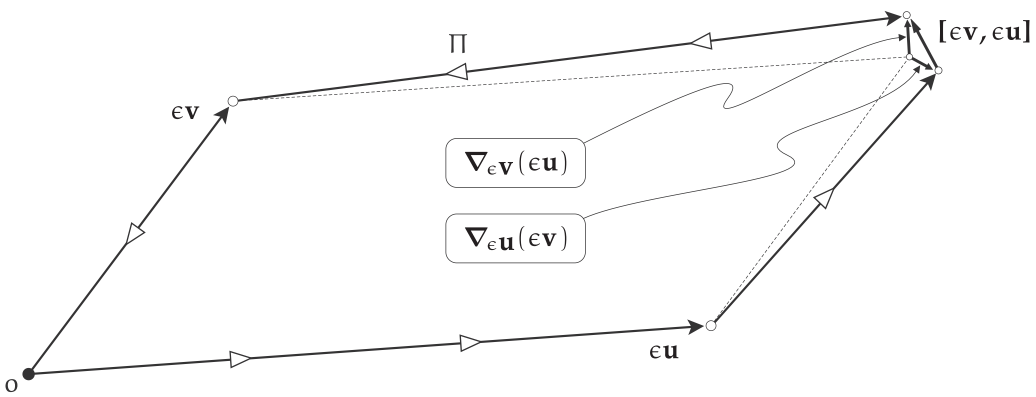
$$\boxed{\text{flux} \asymp (\boldsymbol{\nabla} \times \underline{\varphi}) \bullet \widehat{\mathbf{n}}\, \delta\mathcal{A}.} \tag{37.9}$$

Then, combining (36.15), (37.8), and (37.9),

$$\boxed{\oint_{\Pi(\epsilon\mathbf{u}, \epsilon\mathbf{v})} \varphi = \boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}) \asymp \mathbf{d}\varphi(\epsilon\mathbf{u}, \epsilon\mathbf{v}) = (\boldsymbol{\nabla} \times \underline{\varphi}) \bullet \widehat{\mathbf{n}}\, \delta\mathcal{A}.} \tag{37.10}$$

---

[2]We first used this reasoning in VCA (pp. 412–413) to provide a direct, geometrical proof of Cauchy's Theorem.



[37.4] *General vector fields do not create a closed parallelogram. However, we may create a closed loop* Π *by bridging the gap with the commutator:* $[\epsilon\mathbf{v}, \epsilon\mathbf{u}] = \boldsymbol{\nabla}_{\epsilon\mathbf{v}}\, \epsilon\mathbf{u} - \boldsymbol{\nabla}_{\epsilon\mathbf{u}}\, \epsilon\mathbf{v}$.

We end with another bonus: a *geometrical* explanation of the original, full formula for the exterior derivative, (36.2), *including* its formerly mysterious commutator term. Recall that we first arrived at the following formula for $\mathbf{d}\varphi$ by means of a purely formal calculation (albeit an enlightened calculation):

$$\mathbf{d}\varphi(\mathbf{u},\mathbf{v}) = \boldsymbol{\nabla}_{\mathbf{u}}\varphi(\mathbf{v}) - \boldsymbol{\nabla}_{\mathbf{v}}\varphi(\mathbf{u}) - \varphi([\mathbf{u},\mathbf{v}]).$$

We chose to simplify the foregoing exposition by insisting on constant vector fields, but what if $\mathbf{u}$ and $\mathbf{v}$ are *not* constant? Then, as illustrated in [37.4], the vector fields do not generally create a closed loop. However, we may *create a closed loop* Π *by bridging the gap with the commutator* $[\epsilon\mathbf{v}, \epsilon\mathbf{u}]$. The bulk of the foregoing analysis goes through unchanged, but now *the integral* $\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v})$ *gains one extra contribution from the gap-closing commutator edge:*

$$\varphi([\epsilon\mathbf{v}, \epsilon\mathbf{u}]) = -\varphi([\epsilon\mathbf{u}, \epsilon\mathbf{v}]),$$

thereby explaining the formula for $\mathbf{d}\varphi$!

37.2.2 d*(2-Form)*

Let us repeat the above analysis for the integral of a 2-form $\boldsymbol{\Psi}$ over the oriented 2-dimensional boundary $\Pi(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w})$ of a small—ultimately vanishing, as ϵ vanishes—right-handed parallelepiped with first edge $\epsilon\mathbf{u}$, second edge $\epsilon\mathbf{v}$, and third edge $\epsilon\mathbf{w}$.

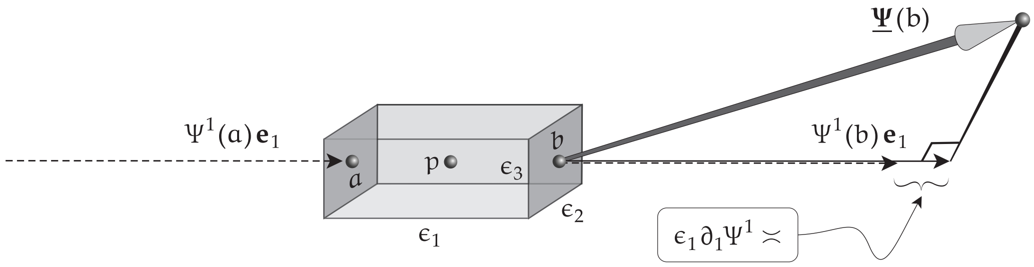
Let us define $\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w})$ to be the integral of φ over Π:

$$\boxed{\boldsymbol{\Omega}(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w}) \equiv \iint\limits_{\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v},\epsilon\mathbf{w})} \boldsymbol{\Psi}.}$$

In the language of classical Vector Calculus, this is the net *outward flux* of $\underline{\boldsymbol{\Psi}}$ through Π:

$$\boxed{\Omega(\Pi) = \iint\limits_{\Pi} \underline{\boldsymbol{\Psi}} \bullet \widehat{\mathbf{n}}\, \mathrm{d}\mathcal{A},}$$

where $\widehat{\mathbf{n}}$ is the unit normal pointing *out* of the parallelepiped.



[37.5] *Flux out of right face* $\asymp \epsilon_2\epsilon_3\,\Psi^1(b)$. *Flux out of left face* $\asymp -\epsilon_2\epsilon_3\,\Psi^1(a)$. *Therefore, the net outflux through this pair of faces is* $\asymp [\partial_1\Psi^1\epsilon_1]\,\epsilon_2\epsilon_3 = [\partial_1\Psi^1]\,\mathcal{V}$.

In light of the 1-form result (37.6), we anticipate that this integral is a *3-form*, and not just any 3-form—it should be the *exterior derivative* of $\boldsymbol{\Psi}$:

$$\iint_{\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v},\epsilon\mathbf{w})} \boldsymbol{\Psi} = \boldsymbol{\Omega}(\epsilon\mathbf{u},\epsilon\mathbf{v},\epsilon\mathbf{w}) \asymp \mathbf{d}\boldsymbol{\Psi}(\epsilon\mathbf{u},\epsilon\mathbf{v},\epsilon\mathbf{w}) = (\boldsymbol{\nabla}\bullet\underline{\boldsymbol{\Psi}})\,\mathcal{V}(\epsilon\mathbf{u},\epsilon\mathbf{v},\epsilon\mathbf{w}),$$

by virtue of (36.16). So, abstracting the vectors away, we are led to believe that

$$\boxed{\boldsymbol{\Omega} \asymp \mathbf{d}\boldsymbol{\Psi} = (\boldsymbol{\nabla}\bullet\underline{\boldsymbol{\Psi}})\,\mathcal{V}.} \tag{37.11}$$

This is indeed true, and the proof is simplest and most transparent in the case that the edges of the parallelepiped (which we take to be centred at p) are orthogonal, directed along the orthonormal basis vectors: $\{\epsilon_1\,\mathbf{e}_1, \epsilon_2\,\mathbf{e}_2, \epsilon_3\,\mathbf{e}_3\}$, so that its volume is $\mathcal{V} = \epsilon_1\epsilon_2\epsilon_3$.

As illustrated in [37.5], consider the flux of $\underline{\boldsymbol{\Psi}}$ through the two faces that are orthogonal to $\mathbf{e}_1$, that are separated by distance ϵ_1, and whose centres are a and b. Only the component $\Psi^1\mathbf{e}_1$ carries fluid across these these faces; the other two components flow parallel to the faces. (For the sake of definiteness, in this figure we have assumed that $\Psi^1 > 0$.)

To evaluate the flux out of the right-hand face, we multiply its area $\epsilon_2\epsilon_3$ by the orthogonal component $\Psi^1(b)$ of the velocity at the centre of the face, so that

$$\text{flux out of right face} \asymp \epsilon_2\epsilon_3\,\Psi^1(b).$$

For the left-hand face, the outward-pointing normal $\widehat{\mathbf{n}} = -\mathbf{e}_1$ is *opposite* to the outward normal of the right face, so

$$\textit{outward}\text{ velocity of fluid on left face} = \underline{\boldsymbol{\Psi}}\bullet\widehat{\mathbf{n}} = -\Psi^1(a).$$

Therefore,

$$\text{flux } \textit{out} \text{ of left face} = -(\text{flux } \textit{into} \text{ left face}) \asymp -\epsilon_2\epsilon_3\,\Psi^1(a).$$

Note that in [37.5] we have drawn $\Psi^1(a)\mathbf{e}_1$ (dashed) with its arrowhead *at* a, both to declutter the picture and to make it clear that this represents flow *into* the box. Note that we have also drawn a *copy* of $\Psi^1(a)\mathbf{e}_1$ at b, to facilitate an upcoming comparsion to $\Psi^1(b)\mathbf{e}_1$

Adding these outfluxes,

$$\begin{aligned} \text{NET out-flux through this pair of faces} &\asymp [\Psi^1(b) - \Psi^1(a)]\,\epsilon_2\epsilon_3 \\ &\asymp [\partial_1\Psi^1\epsilon_1]\,\epsilon_2\epsilon_3 \\ &= [\partial_1\Psi^1]\,\mathcal{V}, \end{aligned}$$

in which the derivative is evaluated at the centre p of the box.

Of course exactly the same reasoning applies to the pair of faces orthogonal to $\mathbf{e}_2$, yielding a net outflux of $[\partial_2\Psi^2]\mathcal{V}$ through that pair. The final pair likewise contributes $[\partial_3\Psi^3]\mathcal{V}$. Thus,

$$\Omega(\Pi) = \text{Total flux out of } \Pi \asymp \left[\partial_1\Psi^1 + \partial_2\Psi^2 + \partial_3\Psi^3\right]\mathcal{V} = (\boldsymbol{\nabla} \bullet \underline{\boldsymbol{\Psi}})\mathcal{V},$$

thereby proving (37.11).

The results (37.6) and (37.11) are part of a pattern that continues on into higher dimensions. Though visualization fails us, it is nevertheless true that $\mathbf{d}$(3-form) applied to a small, compact, ultimately vanishing region of a 4-dimensional space yields the integral of the 3-form over its 3-dimensional boundary, and so on. For more on these higher-dimensional results, see the *Further Reading* section at the end of this book.

37.3 Fundamental Theorem of Exterior Calculus (Generalized Stokes's Theorem)

37.3.1 Fundamental Theorem of Exterior Calculus

All of the integral theorems you learned in Vector Calculus—(37.5), Green's Theorem, Stokes's Theorem, and Gauss's Theorem—are merely special cases of *one* elegant theorem about Forms.

The Fundamental Theorem of Exterior Calculus:
$$\boxed{\int_{\mathrm{R}} \mathbf{d}\boldsymbol{\Phi} = \int_{\partial \mathrm{R}} \boldsymbol{\Phi}} \qquad (37.12)$$

On the left, the $(p+1)$-form $\mathbf{d}\boldsymbol{\Phi}$ is integrated over a compact (oriented) $(p+1)$-dimensional region R, and, on the right, the p-form $\boldsymbol{\Phi}$ is integrated over the (oriented) p-dimensional *boundary* $\partial\mathrm{R}$ of R. Note that we must dispense with multiple integral signs, instead using a single integral sign on both sides of this equation, because the degree p is general and unknown.

WARNING: What we have called the ***Fundamental Theorem of Exterior Calculus*** (hereinafter, FTEC) is instead called, in essentially all other books, the ***Generalized Stokes's Theorem***, often abbreviated as GST.

37.3.2 Historical Aside

Unquestionably, Stokes made great contributions to science, but it is singularly inappropriate that *this* theorem should bear his name, and his name alone. Vladimir Arnol'd (1989, p. 192) refers to (37.12)—tongue-in-cheek—as the "*Newton-Leibniz-Gauss-Green-Ostrogradskii-Stokes-Poincaré formula*"! But, out of this list, Stokes actually had the *least* to do with the theorem.

In 1854, Cambridge University's Smith's Prize examination was devised by Stokes. Question number 8 asked the students to prove the original, pregeneralized result—the one we now call *Stokes's Theorem*; this was the very first appearance of the result in print. One of the candidates who sat for that examination was none other than James Clerk Maxwell—who tied for first place, by the way—and the initial naming of the theorem after Stokes likely stemmed from this event!

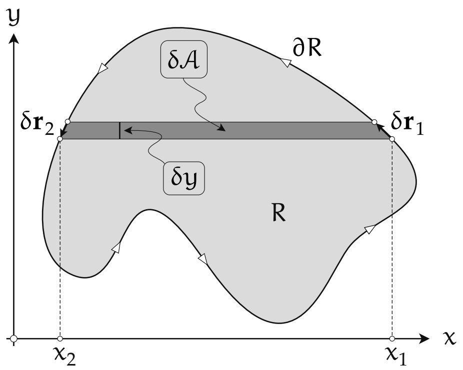
But Stokes only knew of the result because William Thomson (Lord Kelvin) had sent it to him in a letter, on the 2nd of July, 1850! But Kelvin seems to have been led to the result by studying Green ...! You begin to see what a pickle we are in; for extra pickles, see Katz (1979) and Crowe (1985).

An Alexandrian solution to this Gordian Knot of a naming conundrum was proposed by N.M.J. Woodhouse to Roger Penrose: jettison *all* these names, and call it instead the *Fundamental Theorem of Exterior Calculus*. That is the solution adopted in Penrose and Rindler (1984), and subsequently in Penrose (2005), and we hereby follow those precedents.

As for the history of the general result, (37.12), we shall only say this: the first complete statement occurred in a lecture course delivered by Cartan in Paris in 1936–1937, in which he explicitly pointed out that all the results of Vector Calculus were special cases; these lectures first appeared in print in Cartan (1945).

37.3.3 Example: Area

Returning to the result itself, let us begin with an elementary geometrical example of the theorem in action.



[37.6] *The contributions to* $\oint_{\partial R} x\,\mathbf{d}y$ *from* $\delta\mathbf{r}_1$ *and* $\delta\mathbf{r}_2$ *sum to* $\delta\mathcal{A}$.

As illustrated in [37.6], in $\mathbb{R}^2$ consider the integral of the 1-form $x\,\mathbf{d}y$ around the counterclockwise boundary ∂R of the region R of area $\mathcal{A}(R)$.

Let us evaluate the contributions to the integral from the illustrated matching pair of movements $\delta\mathbf{r}_1$ and $\delta\mathbf{r}_2$ sandwiched between the horizontal lines at heights y and $y+\delta y$:

$$(x\,\mathbf{d}y)(\delta\mathbf{r}_1)+(x\,\mathbf{d}y)(\delta\mathbf{r}_2)=(x_1-x_2)\,\delta y \asymp \delta\mathcal{A}.$$

Thus,

$$\oint_{\partial R} x\,\mathbf{d}y=\mathcal{A}(R).$$

(Check that this argument continues to work properly even if we slide the shaded horizonal strip down until it splits into two strips.)

This is indeed in accord with the Fundamental Theorem of Exterior Calculus (FTEC), (37.12), for it predicts that

$$\oint_{\partial R} x\,\mathbf{d}y=\iint_R \mathbf{d}(x\,\mathbf{d}y)=\iint_R \mathbf{d}x\wedge\mathbf{d}y=\iint_R \boldsymbol{\mathcal{A}}=\mathcal{A}(R).$$

This also makes physical sense. According to the correspondence (34.10), $\mathbf{d}x\wedge\mathbf{d}y$ corresponds to a flow in the z-direction, orthogonal to the (x,y)-plane, with unit speed. Thus its flux through a region of area $\mathcal{A}$ is equal to $\mathcal{A}$.

37.4 The Boundary of a Boundary Is Zero!

Before we prove the Fundamental Theorem of Exterior Calculus—or FTEC—let us see how it resolves a mystery we have been living with for some time: *why* is $\mathbf{d}^2=0$? The computational *proof* of this fundamental result was short and easy (see Section 36.4.1), but what does the result itself actually *mean*?!

If we apply FTEC *twice*, we find that

$$0=\int_R \mathbf{d}^2\boldsymbol{\Phi}=\int_{\partial R}\mathbf{d}\boldsymbol{\Phi}=\int_{\partial(\partial R)}\boldsymbol{\Phi}.$$

But since $\boldsymbol{\Phi}$ and R are arbitrary, this can only be true if

$$\boxed{\partial^2 = 0 \quad \Longleftrightarrow \quad \textit{The boundary of a boundary is zero!}}$$

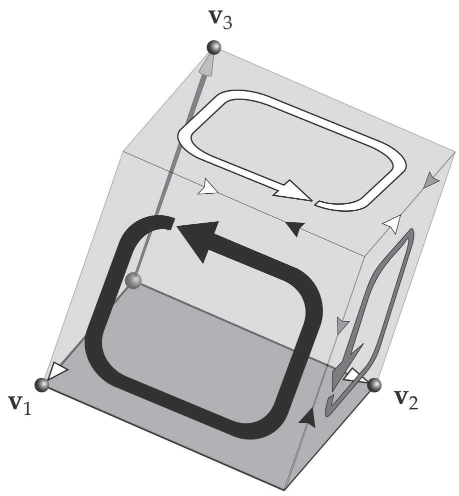
Conversely, if we can understand this geometrical statement, we will have explained the geometrical *meaning* of $\mathbf{d}^2 = 0$:

$$\boxed{\mathbf{d}^2 = 0 \quad \Longleftrightarrow \quad \partial^2 = 0.}$$

Let us explain this in 3-2-1 dimensions. Consider [37.7], which shows a right-handed, 3-dimensional parallelepiped with edges $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. Its boundary consists of six 2-dimensional faces, with indicated orientations. Finally, the boundary of the boundary consists of 12 edges, *each of which is traversed once in one direction, and once in the opposite direction.*

Thus, as illustrated, we see that the 1-dimensional boundary of the 2-dimensional boundary of the 3-dimensional solid is indeed *zero!*

For an explanation in 4-3-2 dimensions and beyond, see Misner, Thorne, and Wheeler (1973, §15). As they explain, but we shall not, this is also the key to understanding the Differential Bianchi Identity, (29.17).



[37.7] *The boundary of a boundary is zero:* $\partial^2 = 0$.

37.5 The Classical Integral Theorems of Vector Calculus

Of course the above geometrical explanation of $\mathbf{d}^2 = 0$ hinges on FTEC, and we have yet to explain why *that* is true! But before we offer a proof of FTEC, we shall ramp up the suspense still further, in order to convince you that the Fundamental Theorem of Exterior Calculus truly is *"Fundamental"!* To that end, we now verify our earlier claim that this single theorem contains *all* of the classical theorems of Vector Calculus as special cases.

37.5.1 $\boldsymbol{\Phi} = $ *0-Form*

Taking $\boldsymbol{\Phi} = f$ to be a function (0-form) in FTEC, $\mathbf{d}\boldsymbol{\Phi} = \mathbf{d}f$ is a 1-form, so R must be taken to be a 1-dimensional curve K, and its boundary ∂K must be taken to be its end points, say, a and b. Since K points at b and points away from a, we count the arrival point as $+b$, and the departure point as $-a$. Then FTEC yields the previously established (37.4), and its classical vector equivalent (37.5):

$$\int_K \mathbf{d}f = f(b) - f(a) \quad \Longleftrightarrow \quad \int_K (\boldsymbol{\nabla} f) \cdot \mathbf{dr} = f(b) - f(a).$$

37.5.2 $\boldsymbol{\Phi} = $ 1-Form

Taking $\boldsymbol{\Phi} = \boldsymbol{\varphi}$ to be a 1-form in FTEC, $\mathbf{d}\boldsymbol{\Phi} = \mathbf{d}\boldsymbol{\varphi}$ is the flux 2-form of the vector field $\boldsymbol{\nabla} \times \underline{\varphi}$, so R must be taken to be a an oriented 2-dimensional surface $\mathcal{S}$, and its boundary $\partial\mathcal{S}$ must be taken to be its 1-dimensional boundary curve.

Let us begin with the simplest case, and historically first case, in which the surface $\mathcal{S}$ is simply a planar region R of the (x, y)-plane, exactly as in our area example above, [37.6]. Here the vector field is also planar, so that the 1-form is given by

$$\boldsymbol{\varphi} = \varphi_x \, \mathbf{d}x + \varphi_y \, \mathbf{d}y.$$

Then the corresponding flux 2-form is

$$\begin{aligned} \mathbf{d}\boldsymbol{\varphi} &= \mathbf{d}\varphi_x \wedge \mathbf{d}x + \mathbf{d}\varphi_y \wedge \mathbf{d}y \\ &= \partial_y \varphi_x \, \mathbf{d}y \wedge \mathbf{d}x + \partial_x \varphi_y \, \mathbf{d}x \wedge \mathbf{d}y \\ &= \left(\partial_x \varphi_y - \partial_y \varphi_x\right) \boldsymbol{\mathcal{A}}. \end{aligned}$$

According to (34.10), this $\mathbf{d}\boldsymbol{\varphi}$ corresponds to a flow in the z-direction, orthogonal to the (x, y)-plane, with velocity $\left(\partial_x \varphi_y - \partial_y \varphi_x\right)$. According to FTEC, its flux through R is equal to the circulation of $\underline{\varphi}$ around ∂R. In other words, if we begin and end with classical notation, we see that FTEC implies ***Green's Theorem***:[3]

$$\begin{aligned} \oint_{\partial R} \varphi_x \, dx + \varphi_y \, dy &= \oint_{\partial R} \boldsymbol{\varphi} \\ &= \iint_R \mathbf{d}\boldsymbol{\varphi} \\ &= \iint_R \left(\partial_x \varphi_y - \partial_y \varphi_x\right) \boldsymbol{\mathcal{A}} \\ &= \iint_R \left(\partial_x \varphi_y - \partial_y \varphi_x\right) dx \, dy. \end{aligned}$$

This is a generalized version of the area example in [37.6].

Turning now to the case of a more general, curved surface $\mathcal{S}$ in space, let $\widehat{\mathbf{n}}$ be the unit normal to surface, the choice of direction being determined by the orientation of the surface via the right-hand rule. Then, by virtue of (37.9), FTEC yields ***Stokes's Theorem***:

$$\begin{aligned} \oint_{\partial\mathcal{S}} \underline{\varphi} \bullet \mathbf{dr} &= \oint_{\partial\mathcal{S}} \boldsymbol{\varphi} \\ &= \iint_{\mathcal{S}} \mathbf{d}\boldsymbol{\varphi} \\ &= \iint_{\mathcal{S}} (\boldsymbol{\nabla} \times \underline{\varphi}) \bullet \widehat{\mathbf{n}} \, d\mathcal{A}. \end{aligned}$$

This theorem provides the crucial bridge between the macroscopic electromagnetic phenomena that Faraday observed in his laboratory, and the microscopic differential equations to which Maxwell ultimately reduced them mathematically.

[3] Green never actually wrote down the 2-dimensional formula that now bears his name, but it *is* a logical consequence of a 3-dimensional formula that he did discover. The first known publication of Green's Theorem (but without any proof) was by Cauchy, in 1846. See Katz (1979).

Take ***Faraday's Law of Electromagnetic Induction***, which Faraday discovered in 1831:

> *If the surface $\mathcal{S}$ spans a loop of wire $\partial\mathcal{S}$, and the field lines of a varying magnetic field pass through it, then an electromotive force is induced in the loop that is equal to the negative of the rate of change of the magnetic flux through $\mathcal{S}$.*

Expressed in mathematical form,

$$\oint_{\partial\mathcal{S}} \underline{\mathbf{E}} \bullet \mathrm{d}\mathbf{r} = -\partial_t \iint_{\mathcal{S}} \underline{\mathbf{B}} \bullet \widehat{\mathbf{n}}\, \mathrm{d}\mathcal{A}.$$

But then Stokes's Theorem allows us to re-express this as

$$\iint_{\mathcal{S}} \left[\boldsymbol{\nabla} \times \underline{\mathbf{E}} + \partial_t \underline{\mathbf{B}}\right] \bullet \widehat{\mathbf{n}}\, \mathrm{d}\mathcal{A} = 0.$$

But if this is to be true for all surfaces $\mathcal{S}$, then the integrand itself must vanish, and we thereby find that Faraday's Law of 1831 is equivalent to one of Maxwell's equations of 1873:

$$\boldsymbol{\nabla} \times \underline{\mathbf{E}} + \partial_t \underline{\mathbf{B}} = 0. \tag{37.13}$$

37.5.3 $\boldsymbol{\Phi} = $ 2-Form

If we take $\boldsymbol{\Phi} = \boldsymbol{\Psi}$ to be a flux 2-form in FTEC, then, by virtue of (37.11), $\mathbf{d}\boldsymbol{\Psi} = (\boldsymbol{\nabla} \bullet \underline{\boldsymbol{\Psi}})\,\mathcal{V}$ is the 3-form describing the source density. Now R must be taken to be a an oriented volume V, and its boundary $\mathcal{S} = \partial V$ must be taken to be its 2-dimensional boundary surface. So, if we begin and end with classical notation, we see that FTEC implies ***Gauss's Theorem***, also known as the ***Divergence Theorem:***

$$\iiint_V (\boldsymbol{\nabla} \bullet \underline{\boldsymbol{\Psi}})\, \mathrm{d}\mathcal{V} = \iiint_V \mathbf{d}\boldsymbol{\Psi} = \oiint_{\partial V} \boldsymbol{\Psi} = \oiint_{\mathcal{S}} \underline{\boldsymbol{\Psi}} \bullet \widehat{\mathbf{n}}\, \mathrm{d}\mathcal{A}. \tag{37.14}$$

37.6 Proof of the Fundamental Theorem of Exterior Calculus

We shall prove the 2-dimensional form (Stokes's Theorem) of FTEC, and we shall also extend the proof (essentially unchanged) to the 3-dimensional case (Gauss's Theorem). But we shall not develop the concepts, terminology, and notation required to explicitly extend the proof, in a uniform manner, to all dimensions. For that, see the *Further Reading* section at the end of this book.

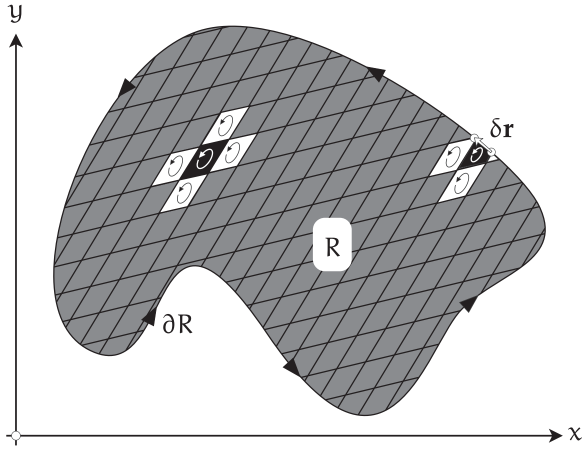
Consider [37.8], which shows a region R of the (x, y)-plane divided up into small (ultimately vanishing) parallelograms, each of which has first edge $\epsilon\mathbf{u}$ and second edge $\epsilon\mathbf{v}$. The orientation associated with this ordering is indicated by the swirl within each cell. Note that the parallelograms *abutting* the boundary ∂R are cut off into irregular shapes.

Now let us sum the integrals of the 1-form φ around all of the parallelograms and truncated parallelograms into which R has been divided. We shall collectively refer the parallelograms and the cut-off parallelograms as the ***cells*** of R.

Consider the illustrated, black, interior parallelogram, and the four white copies that abut it. Every black edge is matched with an oppositely directed white edge of a neighbouring cell, and so the integrals along these two edges *cancel*. The only edges that are *not* cancelled are the small pieces of the boundary, ∂R, such as the illustrated $\delta\mathbf{r}$. Therefore,

$$\oint_{\partial R} \varphi = \sum_{\text{cells in R}} \oint_{\partial(\text{cell})} \varphi \tag{37.15}$$

One might naively guess that the integral around each cell should be of order ϵ, since each edge is of that order. But that cannot be correct, as one can see from the following rough argument.



[37.8] *The circulations along the edges of the interior, black parallelogram are all cancelled by the oppositely directed edges of the abutting white parallelograms. Summing over all cells, the only surviving contributions are from the uncancelled boundary edges, such as the illustrated* $\delta\mathbf{r}$.

We know that for a general 1-form, the integral round ∂R—hence the sum (37.15)—will be nonzero and finite. This leads us to believe that each term must die away with the reciprocal dependence on ϵ as governs the growth of the number of terms in the series. But the number of terms grows as *(fixed area of* R, *divided by the area of each cell)*—that is, as $(1/\epsilon^2)$. Thus the magnitude of each term is expected to die away as ϵ^2. If our original guess had been correct, the order of the sum in (37.15) would have been $\epsilon(1/\epsilon^2)$, yielding an *infinite* result as the parallelograms shrunk. Conversely, any contributions to the terms involving powers of ϵ *greater* than two, cannot have any influence on the final result.

In fact, there is no need to appeal to such rough reasoning, because we previously derived the precise result (37.8) governing the integral of φ around such a shrinking parallelogram, and this is the key to understanding and proving FTEC. We see that, as anticipated, the contribution from each parallelogram is indeed of order ϵ^2, and is given by the application of the 2-form $\mathbf{d}\varphi$ to the edges of the parallelogram. Thus, (37.15) becomes

$$\oint_{\partial R} \varphi = \sum_{\text{cells in R}} \mathbf{d}\varphi(\epsilon\mathbf{u}, \epsilon\mathbf{v}).$$

But, in the limit that ϵ vanishes, this sum is precisely the definition of the integral of $\mathbf{d}\varphi$ over R, thereby completing our proof of the Fundamental Theorem of Exterior Calculus:

$$\oint_{\partial R} \varphi = \iint_{R} \mathbf{d}\varphi.$$

Now let us generalize this argument to prove the Stokes's Theorem form of FTEC. To do so, we need only suppose that the region R upon which our parallelograms are drawn is actually a rubber membrane, spanning a stiff but bendable wire, ∂R. We may now deform this wire ∂R out of $\mathbb{R}^2$ into space, so that R takes up the form of some curved surface in $\mathbb{R}^3$. We may then further deform the surface R as we please, while leaving the boundary wire in place.

The two critical ingredients of the above proof were the cancellation of all interior edges, and (37.8). *Both* of these remain true in this new 3-dimensional context, the latter taking the form (37.10):

$$\oint_{\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v})} \varphi \asymp \mathbf{d}\varphi(\epsilon\mathbf{u}, \epsilon\mathbf{v}) = (\boldsymbol{\nabla} \times \underline{\varphi}) \bullet \widehat{\mathbf{n}}\, \delta\mathcal{A}.$$

Thus we have proved the Fundamental Theorem of Exterior Calculus in this case, thereby also proving Stokes's Theorem in the process:

$$\oint_{\partial\mathcal{S}} \underline{\varphi} \bullet \mathrm{d}\mathbf{r} = \oint_{\partial\mathcal{S}} \varphi = \iint_{\mathcal{S}} \mathbf{d}\varphi = \iint_{\mathcal{S}} (\nabla \times \underline{\varphi}) \bullet \widehat{\mathbf{n}}\, \mathrm{d}\mathcal{A}.$$

Finally, let us step up a dimension to see why the above proof of FTEC goes through essentially unchanged, and yields, in the process, a proof of Gauss's Theorem. Let V be a compact region of $\mathbb{R}^3$ with boundary surface $\mathcal{S} = \partial V$. Now fill V into parallelepipeds with edges $\{\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w}\}$, and volume $\mathcal{V}(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w})$. The ones at the surface are truncated, just as happened with the parallelograms of [37.8]. Again, let us collectively refer the interior parallelepipeds and truncated ones at the surface as the cells of V.

Now consider the flux of the 2-form $\mathbf{\Psi}$ out of an interior cell. Each of its faces is also the face of a neighbouring cell, of *opposite orientation*: the outward normal of our cell points *into* the neighbouring cell. Thus if we sum the fluxes out of all the cells, the fluxes of all the interior faces *cancel*: physically, the fluid that flows out of a face flows into its neighbour. Thus the only faces that are not cancelled are those comprising part of the boundary surface, and therefore,

$$\oiint_{\partial V} \mathbf{\Psi} = \sum_{\text{cells in } V} \oiint_{\partial(\text{cell})} \mathbf{\Psi}.$$

But we know from (37.11) that the flux out of a small, ultimately vanishing parallelepiped is ultimately of order ϵ^3 and is obtained by applying the 3-form $\mathbf{d\Psi}$ to its edges:

$$\oiint_{\Pi(\epsilon\mathbf{u},\epsilon\mathbf{v},\epsilon\mathbf{w})} \mathbf{\Psi} \asymp \mathbf{d\Psi}(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w}) = (\nabla \bullet \underline{\mathbf{\Psi}})\, \mathcal{V}(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w}),$$

Thus,

$$\oiint_{\partial V} \mathbf{\Psi} \asymp \sum_{\text{cells in } V} \mathbf{d\Psi}(\epsilon\mathbf{u}, \epsilon\mathbf{v}, \epsilon\mathbf{w}).$$

Finally, as ϵ vanishes, the right-hand side becomes the integral of $\mathbf{d\Psi}$ over V. Thus we have proved the Fundamental Theorem of Exterior Calculus in this case.

Beginning and ending with the notation of classical Vector Calculus, we have also proved Gauss's Theorem in the process:

$$\oiint_{\mathcal{S}} \underline{\mathbf{\Psi}} \bullet \widehat{\mathbf{n}}\, \mathrm{d}\mathcal{A} = \oiint_{\partial V} \mathbf{\Psi} = \iiint_{V} \mathbf{d\Psi} = \iiint_{V} (\nabla \bullet \underline{\mathbf{\Psi}})\, \mathrm{d}\mathcal{V}.$$

37.7 Cauchy's Theorem

Suppose that the region R depicted in [37.8] resides not in $\mathbb{R}^2$ but in the *complex* plane, and that a complex function $f(z)$ is analytic throughout it, so that its local geometric effect is an amplitwist. Then, as we saw in (36.13), page 397, the fact that $f(z)$ is locally an amplitwist is equivalent to the statement that the 1-form $f\,\mathbf{d}z$ is *closed*: $\mathbf{d}(f\,\mathbf{d}z) = 0$.

Therefore, FTEC immediately yields ***Cauchy's Theorem***:

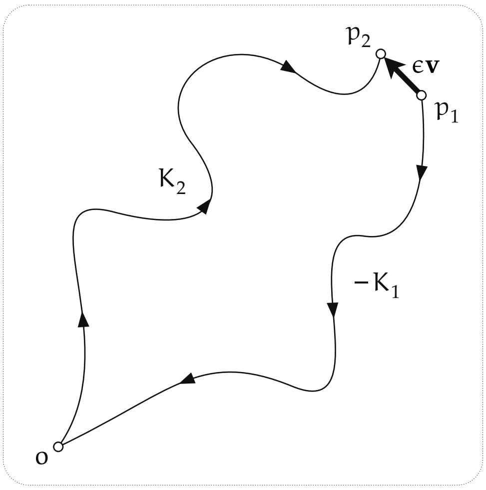
$$\boxed{\oint_{\partial R} f\,\mathbf{d}z = \iint_{R} \mathbf{d}(f\,\mathbf{d}z) = 0.}$$

FTEC can also be used to obtain interesting results when $f(z)$ is *not* locally an amplitwist. For example, in the case of the anticonformal mapping $f(z) = \overline{z}$, we saw in (36.14) that $\mathbf{d}(\overline{z}\,\mathbf{d}z) = 2i\,\boldsymbol{\mathcal{A}}$, so now FTEC yields

$$\boxed{\oint_{\partial R} \bar{z}\,\mathbf{d}z = \iint_R \mathbf{d}(\bar{z}\,\mathbf{d}z) = 2i \iint_R \mathcal{A} = 2i\,\mathcal{A}(R).}$$

For geometrical and physical explanations of this result, see VCA.

37.8 The Poincaré Lemma for 1-Forms



[37.9] *If* $\mathbf{d}\varphi = 0$ *then* $f(p) \equiv \int_o^p \varphi$ *is path-independent, and* $\varphi = \mathbf{d}f$.

Suppose that φ is *closed* throughout a simply connected region R of $\mathbb{R}^n$, i.e., $\mathbf{d}\varphi = 0$. Then, by FTEC, if L is a closed loop in R and we span it with a surface S, so that $L = \partial S$, then

$$\oint_L \varphi = \int_S \mathbf{d}\varphi = 0.$$

This implies, by virtue of (37.3), that $\int \varphi$ is *path-independent*.

Now let us pick an arbitrary fixed point o and treat it as our origin. Thinking of $\underline{\varphi}$ as a force field, we may then define a potential energy function $f(p)$ as the path-independent work done in carrying a particle from o to p:

$$\boxed{f(p) \equiv \int_o^p \varphi.}$$

We will now show that φ is *exact*, and is given by

$$\boxed{\varphi = \mathbf{d}f,} \tag{37.16}$$

thereby proving the Poincaré Lemma in the case of a 1-form.

Let p_1 and p_2 be two neighbouring points, as illustrated in [37.9], and let the short (ultimately vanishing) connecting vector between them be $\epsilon\mathbf{v} = \overrightarrow{p_1 p_2}$. Then the energy difference δf between the two points is

$$\delta f = f(p_2) - f(p_1) \asymp \mathbf{d}f(\epsilon\mathbf{v}).$$

But, as illustrated in [37.9], if K_1 is a path from o to p_1, and K_2 is a path from o to p_2, then $-K_1 + K_2$ is a path from p_1 to p_2. And since $\int \varphi$ is path-independent, we may replace this long, indirect route with the short, direct route along $\epsilon\mathbf{v}$. Therefore,

$$\mathbf{d}f(\epsilon\mathbf{v}) \asymp \int_{-K_1+K_2} \varphi = \int_{\epsilon\mathbf{v}} \varphi \asymp \varphi(\epsilon\mathbf{v}).$$

Since $\mathbf{v}$ is arbitrary, we may abstract it away, completing the proof of (37.16), and with it the proof of the Poincaré Lemma, (36.11), for 1-forms.

37.9 A Primer on de Rham Cohomology

37.9.1 Introduction

What if the region R is *not* simply connected? Then it turns out that a closed Form need *not* be exact, and the study of the Forms that are closed but not exact yields detailed information on the topology of R.

This information is encoded in what are called the ***de Rham cohomology groups***, $H^k(R)$, named after Georges de Rham (1903–1990), who discovered them in 1931. Here k refers to k-forms defined on R that are closed but not exact.

We will only provide a small taste of the subject, beginning with the 1-forms that are closed but not exact, yielding the *first* de Rham cohomology group, $H^1(R)$.

37.9.2 A Special 2-Dimensional Vortex Vector Field

In $\mathbb{R}^2$, with polar coordinates (r, θ), consider the special circular, counterclockwise vortex[4] centred at the origin, shown in [37.10], in which the speed of the flow is chosen to be $|\underline{\varphi}| = (q/2\pi r)$, where q is a constant that measures the strength of the vortex. The reason for this special choice will be revealed in just a moment.

As we approach the origin, the speed goes to infinity, so the origin is a ***singularity***, and we exclude this one point: the flow is defined in the ***punctured plane***, $\mathbb{R}^2 -$ origin.

Since the speed along the circle K_r of radius r is $(q/2\pi r)$, and the length of K_r is $2\pi r$, the circulation around it is

$$\boxed{\mathcal{C} = \oint_{K_r} \varphi = q,}$$

independent of the radius r. Indeed, it was precisely in order to make the circulation independent of the radius that we needed to choose the speed of the flow to be inversely proportional to the radius.

The mere fact that $\mathcal{C} \neq 0$ implies that

> *The 1-form φ of the vortex cannot be exact: $\varphi \neq \mathbf{d}f$, for any f.*

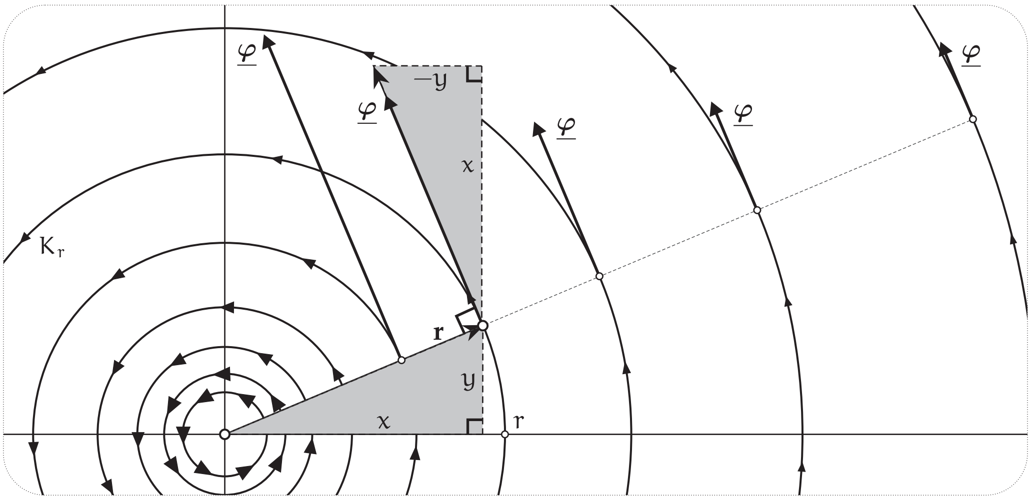
Nevertheless, in a moment we will show that

> *The 1-form φ of the vortex is closed: $\mathbf{d}\varphi = 0$.*

The existence of such a 1-form, that is closed but not exact, is made possible by the *puncturing* of the plane. A loop that encircles the vortex's singularity at the origin cannot be shrunk to a point without crossing the origin, so the punctured plane is *not* simply connected, and therefore there is no conflict with the Poincaré Lemma.

---

[4]Here we prefer the more evocative word ***vortex*** over the more technically precise term ***centre***. See [19.4], page 200.



[37.10] *A special circular vortex vector field* $\underline{\varphi}$ *with speed* $|\underline{\varphi}| = (q/2\pi r)$. *Note the decreasing speed as we move away from the origin. The direction of* $\underline{\varphi}$ *is obtained by rotating the radius vector* $\mathbf{r}$ *through a right angle, but since this has length* $r = |\mathbf{r}|$, *we must multiply it by* $(q/2\pi r^2)$. *Thus the 1-form of the vortex is* $\varphi = (q/2\pi r^2)$ $\left[-y\,\mathbf{d}x + x\,\mathbf{d}y\right]$. *The origin is a singularity of infinite speed, and must be excluded.*

37.9.3 The Vortex 1-Form Is Closed

Figure [37.10] shows a typical radius vector $\mathbf{r}$ of length r, which is then rotated by $(\pi/2)$ to obtain a vector in the direction of the vortex flow. But this vector has length r, whereas we wish to construct a vortex that has speed $(q/2\pi r)$, so we must multiply this vector by $(q/2\pi r^2)$.

Expressed as a 1-form, the vortex vector field is therefore

$$\boxed{\varphi = \frac{q}{2\pi r^2}\left[-y\,\mathbf{d}x + x\,\mathbf{d}y\right],} \tag{37.17}$$

where $r^2 = x^2 + y^2$.

Therefore,

$$\begin{aligned}\left[\frac{2\pi}{q}\right]\mathbf{d}\varphi &= -\frac{1}{r^4}(\mathbf{d}r^2)\wedge\left[-y\,\mathbf{d}x + x\,\mathbf{d}y\right] + 2\,\frac{1}{r^2}\,\mathbf{d}x\wedge\mathbf{d}y\\ &= -\frac{1}{r^4}\,2\,(x\,\mathbf{d}x + y\,\mathbf{d}y)\wedge\left[-y\,\mathbf{d}x + x\,\mathbf{d}y\right] + 2\,\frac{1}{r^2}\,\mathbf{d}x\wedge\mathbf{d}y\\ &= -\frac{1}{r^4}\,2\,(x^2+y^2)\,\mathbf{d}x\wedge\mathbf{d}y + 2\,\frac{1}{r^2}\,\mathbf{d}x\wedge\mathbf{d}y\\ &= 0,\end{aligned}$$

as claimed.

37.9.4 Geometrical Meaning of the Vortex 1-Form

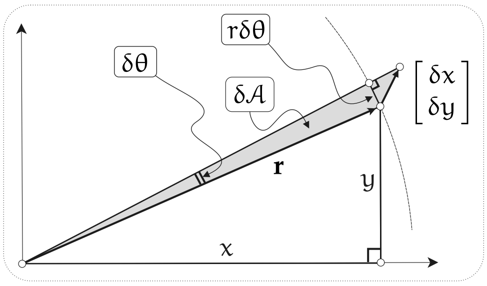
We presented the above calculation in order to provide useful practice with the mechanics of exterior differentiation. However, this calculation is actually unnecessary, for it is possible to instead give a simple geometrical *reason* why φ is closed.

Consider [37.11], which shows a short, ultimately vanishing vector $\begin{bmatrix} \delta x \\ \delta y \end{bmatrix}$ emanating from the tip of $\mathbf{r} = \begin{bmatrix} x \\ y \end{bmatrix}$, subtending an angle of $\delta\theta$ at the origin. Let the shaded area of the triangle with these two edges be $\delta\mathcal{A}$.

On the one hand, we may evaluate $\delta\mathcal{A}$ as one-half of the determinant of the triangle's edges, and on the other hand it is ultimately equal to the area of the triangle with base r and height $r\,\delta\theta$, so,

$$\frac{1}{2}(-y\,\delta x + x\,\delta y) = \frac{1}{2}\det\begin{bmatrix} x & \delta x \\ y & \delta y \end{bmatrix}$$

$$= \delta\mathcal{A} \asymp \frac{1}{2}r(r\,\delta\theta).$$



[37.11] *The area $\delta\mathcal{A}$ of the shaded triangle can be evaluated via the determinant of its edges, yielding $\delta\mathcal{A} = \frac{1}{2}(-y\,\delta x + x\,\delta y)$, or geometrically from the figure, yielding $\delta\mathcal{A} \asymp \frac{1}{2}r^2\,\delta\theta$. Equating these expressions, we find that the vortex 1-form is $\varphi = \frac{q}{2\pi}\,\mathbf{d}\theta$.*

The vortex 1-form (37.17) therefore has a beautifully simple geometrical meaning:

$$\boxed{\varphi = \frac{q}{2\pi}\,\mathbf{d}\theta,}$$

making it crystal clear *why* φ is closed: $\mathbf{d}\varphi = \frac{q}{2\pi}\,\mathbf{d}^2\theta = 0$.

But *wait*—wasn't the entire point of this example that φ is *not* exact? Sweet baby Newton, what the dickens is going on here?!

The resolution of this appararent paradox is somewhat subtle. Although $\mathbf{d}\theta$ has a perfectly well-defined geometrical meaning, it *cannot* be interpreted as ($\mathbf{d}$ of θ), because θ is not even a *function!* Which is to say, any given point must be assigned infinitely many angles.

If we restrict our attention to a simply connected region S that does not contain the singularity at the origin, such as the one shown in [37.12], then we may (nonuniquely) define a genuine, single-valued angle function θ on S, with $\theta_1 \leqslant \theta \leqslant \theta_2$, as illustrated. Now the Poincaré Lemma applies, so the closed 1-form φ must be exact, and we already know the explicit formula: $\varphi = \frac{q}{2\pi}\,\mathbf{d}\theta$.

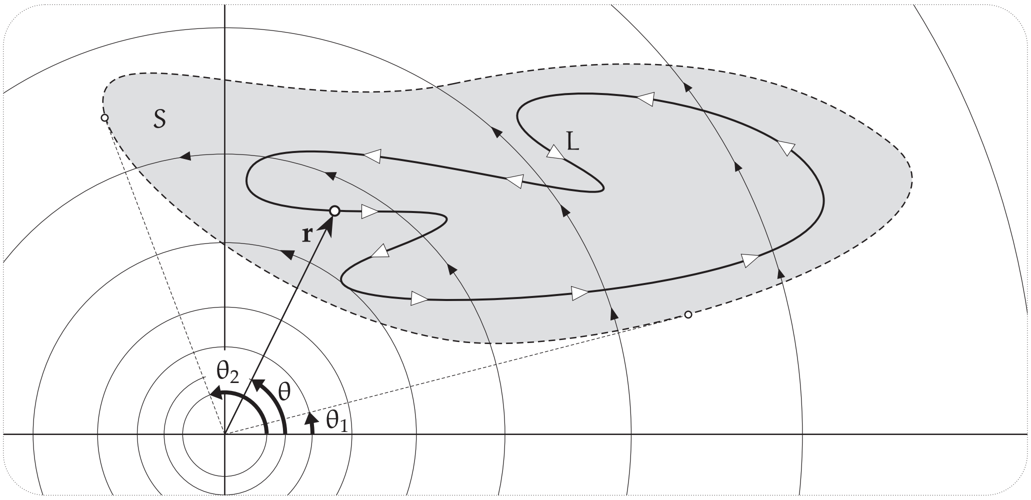
Furthermore, the circulation of the vortex around any loop L contained within S does indeed vanish. For, as the illustrated particle traverses L, the position vector $\mathbf{r}$ rocks back and forth, but when it returns to its starting point, the *net* change in its angle θ is zero, so $\mathcal{C} = \oint_L \frac{q}{2\pi}\,\mathbf{d}\theta = 0$.

But if we attempt to do this in the entire punctured plane, then we immediately run into trouble. Yes, we can *try* to define θ by insisting that $0 \leqslant \theta < 2\pi$, for example, but then θ is not even continuous, let alone differentiable. Thus it is the global topology of the region on which φ is defined that prevents it from being exact throughout that region.

37.9.5 The Topological Stability of the Circulation of a Closed 1-Form

We previously saw that the circulation $\mathcal{C} = \oint_{K_r} \varphi$ of our special vortex was independent of the size of the circle K_r, and the new geometric interpretation of φ makes it obvious why: as we traverse any K_r, the angle increases by 2π, and therefore $\int_0^{2\pi} \frac{q}{2\pi}\,d\theta = q$.

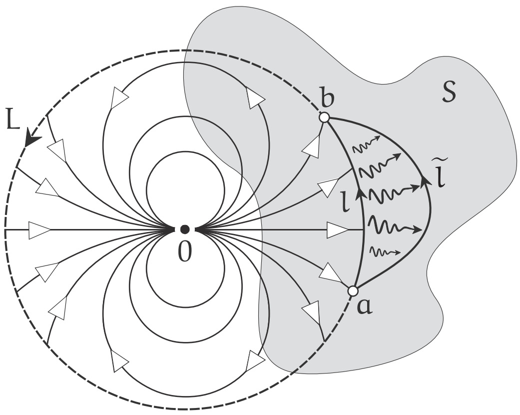
Indeed, this geometrical interpretation makes it clear that the size-independence of $\mathcal{C}$ is in fact a very special case of a far more general phenomenon: if we gradually deform such a circle into *any* shape whatsoever, but being careful not to make it cross the singularity at the origin as it evolves, then the value of the integral does not change!



[37.12] *The simply connected region* S *does not contain the origin, so we we may define a single-valued angle function* θ *on it, where* $\theta_1 \leqslant \theta \leqslant \theta_2$. *Then* $\boldsymbol{\varphi} = \frac{q}{2\pi}\,\mathbf{d}\theta$ *is exact on* S. *As* $\mathbf{r}$ *traces the loop* L, *it rocks back and forth, but when it returns to its starting point, the* net *change in the angle* θ *is zero, so* $\mathcal{C} = \oint_L \frac{q}{2\pi}\,\mathbf{d}\theta = 0$.

In other words,

> *All simple loops that enclose the vortex's singularity at the origin have the same circulation,* $\mathcal{C} = q$. *On the other hand, any loop that does not enclose the singularity has vanishing circulation:* $\mathcal{C} = 0$. (37.18)



[37.13] *The simply connected region* S *does not contain the singularity of the closed dipole field* $\boldsymbol{\chi}$. *The Poincaré Lemma then tells us that* $\oint_L \boldsymbol{\chi}$ *is invariant under the deformation of* l *into* $\widetilde{l}$.

Observe that the distinction between these two kinds of loops is purely *topological*.

NOTE: If a loop goes around the origin m times, then one says that it has ***winding number*** m, in which case $\mathcal{C} = mq$. For much more on the concept of the winding number, and its many applications, see VCA.

In fact, the topological stability of the integral, as expressed in (37.18), is true of *all* closed 1-forms, not just our specific vortex. To understand this broader result, consider the dipole field $\boldsymbol{\chi}$ shown in [37.13], assumed to be closed: $\mathbf{d}\boldsymbol{\chi} = 0$. There is no special significance to this particular field; we simply wish to make it clear that we may apply the following reasoning to *any* closed 1-form.

Now let S be any simply connected region that does *not* contain the singularity at the origin, such as the shaded region shown in [37.13]. Since $\boldsymbol{\chi}$ is closed within S, the Poincaré Lemma implies that $\boldsymbol{\chi} = \mathbf{d}f$ is exact within S, and hence its integral is path-independent within $\mathcal{S}$.

Now, as illustrated, consider an origin-centred circle L, part of which is contained in S. If we take the illustrated arc l (connecting a to b) and deform it into the alternative route $\widetilde{l}$, then the path-independence tells us that

$$\int_{\tilde{l}} \chi = \int_{l} \chi = f(b) - f(a).$$

Thus if $\widetilde{L}$ denotes the new deformed loop, in which l has been deformed into $\tilde{l}$, then

$$\boxed{\oint_{\widetilde{L}} \chi = \oint_{L} \chi.}$$

It is intuitively clear (and we prove it in VCA) that we may now continue deforming L into *any* simple loop around the singularity, and since the circulation does not change throughout this deformation, we have proved the generalized version of (37.18):

Deformation Theorem for Closed 1-Forms. *The circulation (called the* **(de Rham) period***) of a closed 1-form along a loop is invariant under a continuous deformation of the loop.* (37.19)

Given this independence of the circulation from the specific shape of the loop encircling the singularity, we are entitled to think of it as being *a property of the singularity itself*. This is a generalization of the concept in Complex Analysis of the ***residue*** of a complex analytic function $f(z)$ at each of its singularities. Indeed, in VCA (Ch. 11) we explicitly describe residues in terms of the circulation and flux of the Pólya vector field, $\overline{f(z)}$.

37.9.6 The First de Rham Cohomology Group

Now let us return to the definition of the ***first de Rham cohomology group***, for the current case, $R = (\mathbb{R}^2 - \text{origin})$, written $H^1(\mathbb{R}^2 - \text{origin})$. An element of this group is not an individual 1-form, rather it is an *equivalence class* of 1-forms, in which

Two 1-forms are considered fundamentally the same, "equivalent," if they have the same circulation around all loops.

(NOTE: The technical terminology (which we shall not employ) is to say that the two 1-forms are ***cohomologous***.)

For example, suppose we add to the vortex 1-form φ a constant 1-form, corresponding to a steady flow in some direction, which itself clearly has no circulation around any loop. Then the circulation of the new flow is the sum of the circulation of the old flow and the circulation of the steady flow (which vanishes). Therefore, the circulation is unaltered around all loops, and therefore this new 1-form belongs to the same equivalence class as φ.

More generally, we may add *any* flow that has vanishing circulation around all loops. In other words, we may add any 1-form that is *exact*. To sum up,

Let $\widetilde{\varphi} \sim \varphi$ denote the equivalence of the 1-forms $\widetilde{\varphi}$ and φ. Then,

$$\widetilde{\varphi} \sim \varphi \quad \Longleftrightarrow \quad \widetilde{\varphi} - \varphi = \mathbf{d}f, \quad \textit{for some } f.$$

Each equivalence class is defined by its value of the circulation, $\mathcal{C}$. (NOTE: In standard texts on de Rham cohomology, the circulation $\mathcal{C}$ is called the ***period*** of the equivalence class, and the class itself is called a ***cohomology class***.)

TERMINOLOGY WARNING: Historically, the use of the term ***period*** in this context stems from the theory of *elliptic integrals*. But in physics we are accustomed to the period referring to the time it takes for a pendulum to swing back and forth, or the time it takes for a planet to go once around the Sun. In looking at [37.10], we should therefore warn you that the de Rham period *has nothing to do with* the time it takes for a particle of fluid to perform one orbit around the singularity. Indeed, the time T_r it takes to go once around K_r *depends on* r, and is given by $T_r = 4\pi^2(r^2/q)$, whereas the fixed de Rham ***period*** is $\mathcal{C} = q$.

All of the above reasoning only depended on FTEC, not on the specific form of the vortex with which we chose to illustrate the idea. Thus it all carries through to more general 1-forms φ that are closed but not exact.

If we add two such 1-forms, φ_1 and φ_2, from two different equivalence classes, then we obtain a new 1-form from the class that has period

$$\mathcal{C}(\varphi_1 + \varphi_2) = \mathcal{C}(\varphi_1) + \mathcal{C}(\varphi_2).$$

Thus, the group $H^1(\mathbb{R}^2 - \text{origin})$ is isomorphic to the set of real periods under addition, and we may simply write

$$\boxed{H^1(\mathbb{R}^2 - \text{origin}) = \mathbb{R}.}$$

Next, suppose we step up a dimension and consider 1-forms in $\mathbb{R}^3$. If we remove the origin from $\mathbb{R}^3$ then we *can* contract every loop to a point, because now we have enough room to slip a contracting loop *around* the origin, without having to pass through it—so now the space *is* simply connected again. Thus there are no closed 1-forms that are not exact, and we record this fact as

$$\boxed{H^1(\mathbb{R}^3 - \text{origin}) = 0.} \tag{37.20}$$

On the other hand, if we remove the entire z-axis (for example) from $\mathbb{R}^3$ then we have the analogue of the punctured $\mathbb{R}^2$: it is *not* simply connected. Indeed, to deepen the analogy, imagine the flow [37.10] extending upward out of the page, creating a 3-dimensional vortex swirling around the z-axis, with the fluid rotating in concentric cylinders centred on the singular z-axis, where the velocity goes to infinity. Indeed, the analogy with the punctured plane is perfect:

$$\boxed{H^1(\mathbb{R}^3 - z\text{-axis}) = \mathbb{R}.}$$

37.9.7 The Inverse-Square Point Source in $\mathbb{R}^3$

As a prelude to our discussion of the second de Rham cohomology group, $H^2(R)$, we now define a special vector field with a singularity at the origin of $\mathbb{R}^3$. Let $\widehat{\mathbf{r}}$ be the unit radial vector:

$$\widehat{\mathbf{r}} = \frac{1}{r}\,\mathbf{r} = \frac{1}{r}\begin{bmatrix} x \\ y \\ z \end{bmatrix}, \qquad \text{where} \quad r = \sqrt{x^2 + y^2 + z^2}.$$

We now present to you perhaps the single most important vector field in all of physics, the ***inverse-square point source***:

$$\boxed{\underline{\boldsymbol{\Psi}} = \frac{q}{4\pi}\,\frac{\widehat{\mathbf{r}}}{r^2},}$$

where q is a constant that determines the strength of the field. Since the field has a singularity at the origin, this point must be excluded: the field is defined on $\mathbb{R}^3 - \text{origin}$.

This field has (at least!) three very important, physically distinct interpretations:

- If fluid is pumped into 0 at rate q, and flows radially and symmetrically outward, then $\underline{\boldsymbol{\Psi}}$ is the velocity of the fluid flow.
- If a point electric charge q sits at 0, Coulomb's Law says that $\underline{\boldsymbol{\Psi}}$ is its electric field.
- If we reverse the direction of the field, then Newton's Law of Universal Gravitation says that $-\underline{\boldsymbol{\Psi}}$ is the attractive gravitational field of the point mass q located at 0.

  As Newton proved in the *Principia*, this field also represents (very remarkably!) the external gravitational field of *any* spherically symmetric body of mass q, such as the Earth or the Sun.

In $\mathbb{R}^3$ we may interpret any vector field as either a 1-form or a 2-form. Our focus in the next section will be on the 2-form, but first let us discuss the 1-form ψ corresponding to the vector field $\underline{\boldsymbol{\Psi}}$:

$$\boxed{\psi = \frac{q}{4\pi r^3}\left(x\,\mathbf{d}x + y\,\mathbf{d}y + z\,\mathbf{d}z\right).}$$

Let us view this in terms of the third physical interpretation: gravity. Then instead of thinking of the line integral along an open path J as representing circulation, we should now think of it representing *work*:

$$\mathcal{W}_J = \text{work} = \int_J \psi.$$

We know that the gravitational force is *conservative*: carrying a mass around a loop within this gravitational field *must* result in zero net work, or else we could construct a perpetual motion machine! Thus we anticipate that ψ is closed. We leave it as an exercise to confirm by direct calculation that, indeed, $\mathbf{d}\psi = 0$.

However, we can instead demonstrate this by showing that ψ is exact: $\psi = \mathbf{d}f$. More explicitly, let us show that the potential energy is

$$\boxed{f = \text{potential energy} = -\frac{q}{4\pi r}.}$$

As a lemma, first observe [exercise] that

$$\mathbf{d}r = \frac{1}{r}\left(x\,\mathbf{d}x + y\,\mathbf{d}y + z\,\mathbf{d}z\right).$$

Therefore,

$$\begin{aligned}\mathbf{d}f &= -\frac{q}{4\pi}\,\mathbf{d}\left[\frac{1}{r}\right]\\ &= \frac{q}{4\pi r^2}\,\mathbf{d}r\\ &= \frac{1}{4\pi r^3}\big(x\,\mathbf{d}x + y\,\mathbf{d}y + z\,\mathbf{d}z\big)\\ &= \psi.\end{aligned}$$

Thus if J is any path connecting **a** to **b**, then

$$\mathcal{W}_J = \int_J \psi = \int_J \mathbf{d}f = f(\mathbf{b}) - f(\mathbf{a}) = \frac{q}{4\pi}\left[\frac{1}{|\mathbf{a}|} - \frac{1}{|\mathbf{b}|}\right].$$

Note that, unlike our previous vortex example, here the potential f is a well-defined function throughout $(\mathbb{R}^3 - \text{origin})$. Thus this result is in keeping with the claim, (37.20), that the first de Rham cohomology group is trivial in this case.

37.9.8 The Second de Rham Cohomology Group

Next, let us calculate a specific example of the *second* de Rham cohomology group, $H^2(R)$, which measures the 2-forms on R that are closed but not exact. Let us stay in $\mathbb{R}^3$, and consider the flux of a closed 2-form $\mathbf{\Psi}$ out of a closed surface $\mathcal{S}$.

Let us continue thinking about the inverse-square point source, defined on $R = (\mathbb{R}^3 - \text{origin})$, but let us now switch to thinking of it as a *fluid flow*, in which case it is natural to represent the field as a flux 2-form, via (34.10) and (34.12):

$$\boxed{\mathbf{\Psi} = \frac{q}{4\pi}\frac{1}{r^3}\big(x\,\mathbf{d}y\wedge\mathbf{d}z + y\,\mathbf{d}z\wedge\mathbf{d}x + z\,\mathbf{d}x\wedge\mathbf{d}y\big).}$$

That this field cannot be exact can be seen by looking at the flux out of an origin-centred sphere $\mathcal{S}_r$ of radius r. The sphere has area $4\pi r^2$, and the field is orthogonal to it, so the flux Ω is simply the product of the area and the field strength:

$$\boxed{\Omega = \oint\!\!\oint_{\mathcal{S}} \mathbf{\Psi} = q,}$$

independent of r.

Despite the fact that $\mathbf{\Psi}$ is not exact, we will now prove that it is *closed*:

$$\begin{aligned}\left[\frac{4\pi}{q}\right]\mathbf{d\Psi} &= \frac{3}{r^3}\mathcal{V} - \frac{3}{r^4}\,\mathbf{d}r\wedge\big(x\,\mathbf{d}y\wedge\mathbf{d}z + y\,\mathbf{d}z\wedge\mathbf{d}x + z\,\mathbf{d}x\wedge\mathbf{d}y\big)\\ &= \frac{3}{r^3}\left[\mathcal{V} - \frac{1}{r^2}\big(x\,\mathbf{d}x + y\,\mathbf{d}y + z\,\mathbf{d}z\big)\wedge\big(x\,\mathbf{d}y\wedge\mathbf{d}z + y\,\mathbf{d}z\wedge\mathbf{d}x + z\,\mathbf{d}x\wedge\mathbf{d}y\big)\right]\\ &= \frac{3}{r^3}\left[\mathcal{V} - \frac{1}{r^2}(x^2+y^2+z^2)\mathcal{V}\right]\\ &= 0.\end{aligned}$$

Next, let us show that (37.21) generalizes:

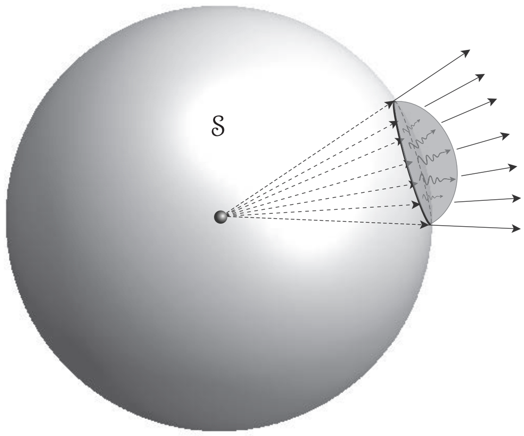
> **Deformation Theorem for Closed 2-Forms**. *The flux (**period**) of a closed 2-form out of a closed 2-surface is invariant under a continuous deformation of the surface.* (37.21)

Figure [37.14] shows fluid flowing radially outward through the sphere $\mathcal{S}$ from the inverse-square point source at the origin. A circular patch of $\mathcal{S}$ is deformed outward—indicated by squiggly arrows—to create the illustrated bump or blister on the surface of $\mathcal{S}$.

All the fluid that flows into the base of the blister also flows out of its outer surface, so if $\widetilde{\mathcal{S}}$ denotes the new deformed surface, then,

$$\boxed{\oint\!\!\oint_{\widetilde{\mathcal{S}}} \boldsymbol{\Psi} = \oint\!\!\oint_{\mathcal{S}} \boldsymbol{\Psi}.}$$

This argument easily generalizes to apply to *any* closed 2-form $\boldsymbol{\Psi}$. Since $\mathbf{d}\boldsymbol{\Psi} = 0$ within the blister, FTEC implies that the net *outflux* from it vanishes. But this can be rephrased as we just did: what flows into the blister must flow out. Thus we have proved (37.21): the flux out of $\mathcal{S}$ is invariant under deformations of $\mathcal{S}$ that do not force it to pass through a singularity of $\boldsymbol{\Psi}$.



[37.14] *Because the 2-form is closed, what flows into the blister must flow out. So the flux out of $\mathcal{S}$ is invariant under deformations.*

Returning to the specific inverse-square point source, a closed surface enclosing the singularity at the origin cannot be contracted to a point without crossing the origin, so all surfaces containing the singularity have the same flux q. This topologically stable flux measures the strength of the source at the origin, which has the three physically distinct interpretations listed above. (Again, we repeat that more advanced texts on de Rham cohomology theory call this source strength the *period*.)

The definition of equivalence classes of 2-forms is completely analogous to that for 1-forms; that is, *two 2-forms* $\widetilde{\boldsymbol{\Psi}}$ *and* $\boldsymbol{\Psi}$ *are* equivalent *if they have the same flux out of every closed surface*. Expressed in more conventional terms,

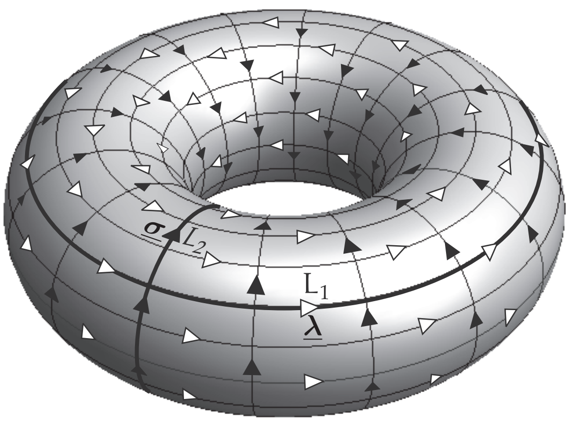
$$\boxed{\widetilde{\boldsymbol{\Psi}} \sim \boldsymbol{\Psi} \quad \Longleftrightarrow \quad \widetilde{\boldsymbol{\Psi}} - \boldsymbol{\Psi} = \mathbf{d}\phi, \qquad \text{for some 1-form } \phi.}$$

If we add two 2-forms then their fluxes add, too. So,

$$\boxed{H^2(\mathbb{R}^3 - \text{origin}) = \mathbb{R}.}$$

37.9.9 The First de Rham Cohomology Group of the Torus

What if the nonexact, closed Forms are defined on a manifold other than $\mathbb{R}^n$, such as closed 2-surface of genus g?



[37.15] *The loop* L_1 *around the equator of the torus, and the loop* L_2 *through the hole of the torus, are topologically distinct: we cannot deform one into the other. The equivalence class of a closed 1-form flow on the torus therefore requires the specification of* two *independent periods.*

Consider the torus shown in [37.15]. Here we see two different possible flows on the surface: the white flow $\underline{\boldsymbol{\lambda}}$ goes around the axis of symmetry of the torus, in the direction the equator(s), while the black flow $\underline{\boldsymbol{\sigma}}$ goes through the hole. We shall assume that both flows are *closed*:

$$\boxed{\mathbf{d}\boldsymbol{\lambda} = 0 = \mathbf{d}\boldsymbol{\sigma}.}$$

Now let L_1 be a closed loop streamline of the white flow $\underline{\boldsymbol{\lambda}}$, and let L_2 be a closed-loop streamline of the black flow $\underline{\boldsymbol{\sigma}}$, as illustrated. As we discussed in connection with [19.9], L_1 and L_2 are topologically *distinct*: we cannot continuously deform one into the other. However, *any* loop that goes once around the axis of symmetry (and zero times through the hole) is topologically equivalent to L_1, and *any* loop that goes once through the hole (and zero times around the axis of symmetry) is topologically equivalent to L_2.

Next, observe that neither of these flows can be exact. Since the white vector field $\underline{\boldsymbol{\lambda}}_1$ flows directly along L_1, it has a nonzero circulation along it. Let us call this period $\omega_1(\boldsymbol{\lambda})$, the subscript identifying which of the topologically distinct loops is being considered:

$$\omega_1(\boldsymbol{\lambda}) = \oint_{L_1} \boldsymbol{\lambda} \neq 0.$$

The Deformation Theorem (37.19) then tells us that this period will be the same for any loop into which we may deform L_1. That is, the period $\omega_1(\boldsymbol{\lambda})$ is a topological property of the white vector field, associated with any loop that goes once around the axis of symmetry (and zero times through the hole).

On the other hand, $\underline{\boldsymbol{\lambda}}$ is everywhere orthogonal to L_2, so it has zero circulation along it:

$$\omega_2(\boldsymbol{\lambda}) = \oint_{L_2} \boldsymbol{\lambda} = 0.$$

Furthermore, by the Deformation Theorem, this will be true of any loop that goes once through the hole (and zero times around the axis of symmetry).

A completely symmetrical analysis applies to the black vector field $\underline{\boldsymbol{\sigma}}$ through the hole:

$$\omega_1(\boldsymbol{\sigma}) = \oint_{L_1} \boldsymbol{\sigma} = 0, \qquad \text{and} \qquad \omega_2(\boldsymbol{\sigma}) = \oint_{L_2} \boldsymbol{\sigma} \neq 0,$$

and again these are true of any loops into which L_1 and L_2 may be deformed.

Now consider a more general vector field $\underline{\phi}$ on the torus, obtained by taking a linear combination of $\underline{\boldsymbol{\lambda}}$ and $\underline{\boldsymbol{\sigma}}$:

$$\phi = a\boldsymbol{\lambda} + b\boldsymbol{\sigma}, \tag{37.22}$$

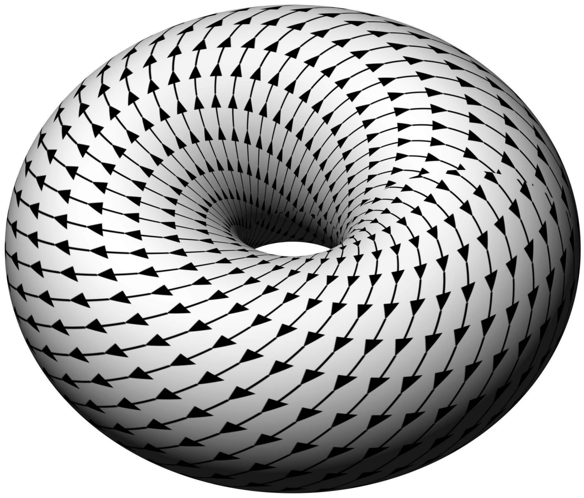
where a and b are constants. This both flows through the hole *and* around the axis of symmetry, as illustrated in [37.16]. But it is still closed, because

$$\mathbf{d}\phi = a\,\mathbf{d}\boldsymbol{\lambda} + b\,\mathbf{d}\boldsymbol{\sigma} = 0,$$

and thus its periods are still topologically defined.

But unlike either of the fields from which it was built, $\underline{\phi}$ has *two* nonvanishing, independent periods:

$$\begin{aligned} \omega_1(\phi) &= a\,\omega_1(\boldsymbol{\lambda}) \neq 0, \\ \omega_2(\phi) &= b\,\omega_2(\boldsymbol{\sigma}) \neq 0. \end{aligned}$$



[37.16] *The equivalence class of a closed 1-form ϕ on the torus is determined by two independent, topological periods: the circulation $\omega_1(\phi)$ around the axis of symmetry, and the circulation $\omega_2(\phi)$ through the hole. Courtesy of RokerHRO, CC by SA 3.0.*

This characterization of ϕ by two independent periods also applies to a completely general closed 1-form on the torus, even if ϕ doesn't happen to have the special form (37.22). To sum up,

> *The equivalence class of a closed 1-form ϕ on the torus is determined by two topologically defined periods: $\omega_1(\phi)$ measures the circulation around the axis of symmetry, and $\omega_2(\phi)$ measures the circulation through the hole.*

We may therefore neatly identify the equivalence class of ϕ with a 2-dimensional vector, which we shall call the ***period vector***, $\omega(\phi)$, given by

$$\omega(\phi) \equiv \begin{bmatrix} \omega_1(\phi) \\ \omega_2(\phi) \end{bmatrix}.$$

Finally, observe that *if we add two flows from two different equivalence classes, their period vectors add, as follows:*

$$\omega(\widetilde{\phi} + \phi) = \begin{bmatrix} \omega_1(\widetilde{\phi} + \phi) \\ \omega_2(\widetilde{\phi} + \phi) \end{bmatrix} = \begin{bmatrix} \omega_1(\widetilde{\phi}) \\ \omega_2(\widetilde{\phi}) \end{bmatrix} + \begin{bmatrix} \omega_1(\phi) \\ \omega_2(\phi) \end{bmatrix} = \omega(\widetilde{\phi}) + \omega(\phi).$$

We have thus found the first de Rham cohomology group of the torus:

$$\boxed{H^1(\text{torus}) = \mathbb{R}^2.}$$

Entire books are devoted to de Rham's theory, but we must come to a full stop here, lest *this* book become one of them! We have glimpsed the tip of a massive, beautiful iceberg, and we must reluctantly turn hard about. But *you* should feel free to go exploring deep beneath the surface, with the *Further Reading* section as your guide.

Chapter 38

Differential Geometry via Forms

38.1 Introduction: Cartan's Method of Moving Frames

At long last, we are finally ready to put the "Devil's machine"[1] to work in the service of geometry!

In this final chapter of this final Act, we shall apply the machinery of Cartan's Forms to Differential Geometry, just as Cartan himself did, reproving by computation many of the same fundamental results that were proved geometrically in the first four Acts. Furthermore, we shall see that Forms naturally lead us to many *new* results. We begin by outlining our battle plans, and their historical origins.

Recall that Frenet in 1847, and (independently) Serret in 1851, both analyzed curves in $\mathbb{R}^3$ by attaching an orthonormal frame $(\mathbf{T}, \mathbf{N}, \mathbf{B})$ to the curve, finding that its rate of change along the curve is given by the *Frenet–Serret Equations*, (9.3), which we repeat here:

$$\begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}' = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix} = [\Omega] \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}. \tag{38.1}$$

Thus, as this frame moves along the curve, its rate of rotation is fully encoded as the matrix $[\Omega]$, which is skew-symmetric (i.e., $[\Omega]^{\mathsf{T}} = -[\Omega]$).

The success of this enterprise hinged upon two ideas:

IDEA 1: Adapt the frame to the curve in a geometrically meaningful way, called the ***Frenet Frame***: $\mathbf{T}$ is tangent to the curve, and $\mathbf{N}$ is the *principal normal*, lying within the instantaneous plane of motion, and pointing at the instantaneous centre of curvature. $\mathbf{B}$ is then determined by the right hand rule.

IDEA 2: Express the rate of change of the frame in terms of the frame *itself*.

Around 1880, Jean-Gaston Darboux took the next step, extending these ideas to *surfaces*:

IDEA 3: Adapt the frame to a *surface*, by choosing one of the vectors to be the *normal* to the surface; the remaining two vectors are then automatically tangent vectors to the surface.

Starting around 1900, Élie Cartan developed Forms and brought them to bear on Differential Geometry, his efforts reaching substantial completion in Cartan (1928), though the final *perfection* of his ideas was perhaps Cartan (1945):

IDEA 4: Free the frame completely, and allow it to vary in *any (differentiable) way whatsoever* throughout $\mathbb{R}^3$. Instead of calling it the *Frenet Frame*, we now call it ***Cartan's moving frame field***, referring to this approach as a whole as ***Cartan's method of moving frames.*** As we shall see shortly, this again yields a skew-symmetric matrix, but it is now a matrix of *1-forms*.

[1]See Prologue.

IDEA 5: Replace the moving frame with its *dual frame of 1-forms*, and look at its rate of change in an arbitrary direction in space. Recapitulating IDEA 2, express this rate of change of the dual basis in terms of the dual basis *itself*.

Let us pause in order to comment on IDEA 4, which would appear to be an ill-fated step *backwards*. Frenet, Serret, and Darboux had all adapted their frames to a geometrical object in a geometrically meaningful way, so it made sense that their examination of its rate of change would yield geometrically meaningful information about that object.

But (diabolical) *Cartan* instead instructs us to do the following. Take your right hand, and arrange your thumb (imagined Satanically elongated!), forefinger, and middle finger into an orthonormal frame. Now wave your hand about in space, twisting your hand as you go![2]

Even if fleetingly, you have created an arbitrary *moving frame field*, at least at each point of the region of space that your hand passed through. Now imagine an entire region of space *filled* (differentiably) with such an arbitrary moving frame field. How on earth can anything mathematically interesting arise from looking at the arbitrary gyrations of your three fingers?!

Very remarkably, Cartan found that such an *arbitrary* moving frame field is subject to *two* extremely elegant laws, called *Cartan's First* and *Second Structural Equations.*[3] These are the subject of Section 38.4.

Most authors pass over the existence of these laws in silence, as if to say, "Keep moving! Nothing to see here." But when *I* first saw these two equations as a student, I was *shocked*—perceiving them to be magical, and possibly *black*-magical, at that! Even now, I cannot offer you a truly compelling explanation for the existence of these beautiful laws, but consider this . . . .

To specify a unit vector in $\mathbb{R}^3$, or, equivalently, a point on the unit sphere, we need *two* numbers—say, longitude and latitude. Thus three arbitrary unit vectors require *six* numbers. But now consider Cartan's *orthonormal* moving frame field, and suppose we are told just *one* of his vectors (using two numbers). We then know that the tips of the remaining two, orthogonal vectors of the frame lie somewhere on the unit circle orthogonal to the given vector. Thus the remaining two vectors can be pinned down with the specification of a *single* angle (of either one).

In other words, the extra structure provided by the *orthogonality* of Cartan's moving frame field has reduced the amount of data by *half.* To phrase this the other way around, Cartan's *orthonormal* moving frame field has *twice* the structure of a nonorthogonal field.

IDEA 6: Having risen above the fray to obtain his completely general Structural Equations for an arbitrary moving frame field, Cartan then descended to Earth, and adapted them to specific geometrical objects.

When Cartan adapted them to a curve, he simply recovered (38.1). But when he adapted his moving frame field to a surface—Darboux's IDEA 3—Cartan's Structural Equations immediately yielded all the *Fundamental Equations of a Surface*, but in a much more compact and elegant form than had been obtained by Gauss, Darboux, Codazzi, and other early masters of Differential Geometry—all thanks to their expression in Cartan's language of *Forms*. These Form equations are the subject of Section 38.5.

In this chapter we shall follow closely in the footsteps of the trail-blazing text of UCLA Professor Barrett O'Neill (1924–2011), *Elementary Differential Geometry* (second revised edition), the first edition of which appeared in 1966.

Today, more than a half-century later, O'Neill's work remains, in our view, the single most clear-eyed, elegant, and (ironically) *modern* treatment of the subject available—present company

[2]We humbly suggest that you do not attempt this in public!

[3]These are also commonly referred to as *Cartan's Structure Equations.*

excepted!—at the undergraduate level. Indeed, as of this writing, it still stands essentially[4] *alone* in its pioneering embrace of Forms as the principal means of communicating Differential Geometry to an undergraduate audience.

We shall carry our investigations substantially further than O'Neill's text, calculating the Riemann tensor of an n-manifold using curvature 2-forms, and also providing many *geometrical* explanations that he does not. Nevertheless, *O'Neill investigates many very interesting topics that we shall not.* Therefore, we *strongly* recommend O'Neill's pioneering work as a companion or follow-up text to this final chapter.

38.2 Connection 1-Forms

38.2.1 Notational Conventions and Two Definitions

Before we get going, let us remind you of some of our (ongoing) notational conventions:

- Vectors are denoted by lowercase bold Roman letters, e.g., $\mathbf{v}$.
- 1-forms are denoted by lowercase bold Greek letters, e.g., $\boldsymbol{\theta}$.
- 2-forms are denoted by uppercase bold Greek letters, e.g., $\boldsymbol{\Psi}$.
- Matrices are denoted by square brackets, e.g., $[A] = [a_{ij}]$, where a_{ij} denotes the matrix entry in *row* i, and *column* j.

To avoid any possibility of confusion while you become acclimatized to Cartan's ideas, we shall depart from usual practice and dispense with the Einstein summation convention, instead writing out summation signs, or even writing out the sums explicitly. This is actually a very small sacrifice because we shall see that once we specialize to 2-surfaces many of the "sums" reduce to a *single* term!

Finally, here are two definitions, one old and one new:

- As usual, let $\{\mathbf{e}_j\}$ be the fixed Euclidean orthonormal basis, with dual 1-form basis $\{\mathbf{d}x^i\}$. Thus, as we saw in (32.12), $\mathbf{d}x^i(\mathbf{e}_j) = \delta^i_j$.
- Let $\{\mathbf{m}_j\}$ be Cartan's (arbitrary, but differentiable) orthonormal ***moving frame field***, and let its dual 1-form basis be $\{\boldsymbol{\theta}^i\}$. Then, by definition, $\boldsymbol{\theta}^i(\mathbf{m}_j) = \delta^i_j$.

  Note that while [32.6] showed that a general, nonorthogonal basis does not allow us to pair an *individual* basis vector with a matching 1-form, this is no longer true of our *orthonormal* moving frame field: we *can* think of $\boldsymbol{\theta}^i$ as the dual of $\mathbf{m}_i$.

38.2.2 Connection 1-Forms

Let us apply IDEA **2** to Cartan's moving frame field. That is, let us *express the rate of change of the moving frame field in terms of the frame field itself*.

In the case of the *Frenet–Serret Equations*, the Frenet frame was only defined along the curve, and we only needed to look at its rate of change along that curve, but *now* we must examine how our moving frame field rotates as we move off in some *general* direction $\mathbf{v}$ in space:

$$\begin{aligned}
\nabla_{\mathbf{v}}\mathbf{m}_1 &= c_{11}\,\mathbf{m}_1 + c_{12}\,\mathbf{m}_2 + c_{13}\,\mathbf{m}_3\\
\nabla_{\mathbf{v}}\mathbf{m}_2 &= c_{21}\,\mathbf{m}_1 + c_{22}\,\mathbf{m}_2 + c_{23}\,\mathbf{m}_3\\
\nabla_{\mathbf{v}}\mathbf{m}_3 &= c_{31}\,\mathbf{m}_1 + c_{32}\,\mathbf{m}_2 + c_{33}\,\mathbf{m}_3.
\end{aligned}$$

[4] Darling (1994) is the closest thing to an exception that we know of, but it smacks more of a graduate text.

If we collect these coefficients into the matrix $[C] \equiv [c_{ij}]$, and make the abbreviation,

$$[\mathbf{m}] \equiv \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \end{bmatrix},$$

then we can rewrite these equations more neatly and compactly as

$$\boldsymbol{\nabla}_{\mathbf{v}}[\mathbf{m}] = [C][\mathbf{m}].$$

Since the moving frame field is (by definition) orthonormal, it follows that these coefficients—making up the entries of the matrix $[C]$—are given by

$$c_{ij} = \left(\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i\right) \bullet \mathbf{m}_j.$$

But these coefficients depend on the choice of $\mathbf{v}$, so we change our notation[5] and write them as *functions* of $\mathbf{v}$:

$\boldsymbol{\omega}_{ij}(\mathbf{v}) \equiv \left(\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i\right) \bullet \mathbf{m}_j$ *is the initial rate at which* $\mathbf{m}_i$ *tips towards* $\mathbf{m}_j$ *as the frame moves along* $\mathbf{v}$. (38.2)

Perhaps you will not be surprised to learn that these $\boldsymbol{\omega}_{ij}$ are *1-forms!*

The $\boldsymbol{\omega}_{ij}$ *are called the* **connection 1-forms**, *and* $\boldsymbol{\omega}_{ji} = -\boldsymbol{\omega}_{ij}$.

WARNING: the two subscripts attached to $\boldsymbol{\omega}_{ij}$ immediately (but *wrongly*!) make one think of the components of tensor of valence $\left\{ \begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right\}$. Let us therefore *stress* that these are all *1-forms*, housed within a matrix of 1-forms, and the subscripts serve to identify the *row and column* of each 1-form entry within this matrix.

We begin by verifying that they are indeed 1-forms:

$$\begin{aligned} \boldsymbol{\omega}_{ij}(a\,\mathbf{v} + b\,\mathbf{w}) &= \left(\boldsymbol{\nabla}_{a\,\mathbf{v}+b\,\mathbf{w}}\,\mathbf{m}_i\right) \bullet \mathbf{m}_j \\ &= \left(a\,\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i + b\,\boldsymbol{\nabla}_{\mathbf{w}}\mathbf{m}_i\right) \bullet \mathbf{m}_j \\ &= a\left(\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i\right) \bullet \mathbf{m}_j + b\left(\boldsymbol{\nabla}_{\mathbf{w}}\mathbf{m}_i\right) \bullet \mathbf{m}_j \\ &= a\,\boldsymbol{\omega}_{ij}(\mathbf{v}) + b\,\boldsymbol{\omega}_{ij}(\mathbf{w}). \end{aligned}$$

Next, to prove $\boldsymbol{\omega}_{ji} = -\boldsymbol{\omega}_{ij}$, we must show that $\boldsymbol{\omega}_{ji}(\mathbf{v}) = -\boldsymbol{\omega}_{ij}(\mathbf{v})$ for every vector $\mathbf{v}$:

$$0 = \boldsymbol{\nabla}_{\mathbf{v}}\delta^i_j = \boldsymbol{\nabla}_{\mathbf{v}}\left(\mathbf{m}_i \bullet \mathbf{m}_j\right) = \boldsymbol{\nabla}_{\mathbf{v}}\left(\mathbf{m}_i\right) \bullet \mathbf{m}_j + \left(\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_j\right) \bullet \mathbf{m}_i = \boldsymbol{\omega}_{ij}(\mathbf{v}) + \boldsymbol{\omega}_{ji}(\mathbf{v}).$$

A *geometrical* explanation of this result is implicit in the upcoming diagram, [38.2].

[5]Frustratingly, the *order* of the indices varies by author. We shall employ the notation of O'Neill (2006), whereas Dray (2015), for example, has i and j reversed: his $\boldsymbol{\omega}_{ji}$ is our $\boldsymbol{\omega}_{ij}$.

It follows that

$$\omega_{11} = \omega_{22} = \omega_{33} = 0.$$

Therefore, the matrix [C] of coefficients is *a skew-symmetric matrix of 1-forms, with only* three *independent entries*, ω_{12}, ω_{13}, ω_{23}:

$$[\omega] \equiv \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22} & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33} \end{bmatrix} = \begin{bmatrix} 0 & \omega_{12} & \omega_{13} \\ -\omega_{12} & 0 & \omega_{23} \\ -\omega_{13} & -\omega_{23} & 0 \end{bmatrix}.$$

If we take it as understood that $[\omega](\mathbf{v}) = [\omega(\mathbf{v})]$ means that each 1-form entry within the matrix acts on $\mathbf{v}$, then the rotation of the moving frame field as it moves off in a general direction $\mathbf{v}$ can be elegantly described by the so-called ***Connection Equations***:

$$\boxed{\boxed{\nabla_{\mathbf{v}}[\mathbf{m}] = [\omega(\mathbf{v})][\mathbf{m}].}} \tag{38.3}$$

In component form, this states

$$\boxed{\nabla_{\mathbf{v}}\mathbf{m}_i = \sum_j \omega_{ij}(\mathbf{v})\,\mathbf{m}_j.} \tag{38.4}$$

For fear that elegance become the enemy of clarity, let us simply spell out these equations *in full*:

$$\begin{aligned} \nabla_{\mathbf{v}}\mathbf{m}_1 &= \qquad\qquad\qquad\quad \omega_{12}(\mathbf{v})\,\mathbf{m}_2 + \omega_{13}(\mathbf{v})\,\mathbf{m}_3 \\ \nabla_{\mathbf{v}}\mathbf{m}_2 &= -\omega_{12}(\mathbf{v})\,\mathbf{m}_1 \qquad\qquad\qquad + \omega_{23}(\mathbf{v})\,\mathbf{m}_3 \\ \nabla_{\mathbf{v}}\mathbf{m}_3 &= -\omega_{13}(\mathbf{v})\,\mathbf{m}_1 - \omega_{23}(\mathbf{v})\,\mathbf{m}_2. \end{aligned}$$

We can use **IDEA 6** to recover the Frenet–Serret equations from these general equations, as follows: we simply ***adapt*** the moving frame to the curve, by choosing

$$\mathbf{m}_1 = \mathbf{T}, \qquad \mathbf{m}_2 = \mathbf{N}, \qquad \mathbf{m}_3 = \mathbf{B}, \qquad \text{and} \qquad \mathbf{v} = \mathbf{T}.$$

In comparing these equations with (38.1), the only mystery is the absence of a term corresponding to $\omega_{13}(\mathbf{v})$.

To understand this, recall that $\mathbf{B}$ is the normal to the instantaneous plane of motion of the curve (*osculating plane*); the orthogonal pair $(\mathbf{T}, \mathbf{N})$ rotate together within this plane, about the axis $\mathbf{B}$. This means that $\mathbf{T} = \mathbf{m}_1$ does not tip in the direction of $\mathbf{B} = \mathbf{m}_3$, so (38.2) implies $\omega_{13}(\mathbf{v}) = 0$, thereby explaining the mystery.

38.2.3 WARNING: *Notational Hazing Rituals Ahead!*

First the *good* news: the connection 1-forms are (*almost*) universally written ω_{ij}, just as we have written them, though often an index is raised ($\omega^i{}_j$) in order to be able to employ the Einstein summation convention—we too shall adopt this notation in due course.

Our use of $\boldsymbol{\theta}^i$ to denote the dual basis of 1-forms follows O'Neill (2006). Happily, this notation is widespread, but other perfectly reasonable alternatives exist e.g., Dray (2015) writes $\boldsymbol{\theta}^i$ as $\boldsymbol{\sigma}^i$.

Now the *bad* news: several highly respected mathematicians[6]—including more than one Fields Medalist—have elected to employ a notation that seems to us to have been expressly and perversely *designed* to induce maximum befuddlement of the hapless student—instead of $\boldsymbol{\theta}^i$, these mathematicians write $\boldsymbol{\omega}^i$! Yes, they use the *same* Greek letter to denote the connection 1-forms $\boldsymbol{\omega}^i{}_j$ and the basis 1-forms $\boldsymbol{\omega}^i$, with only the number of indices as clue that the symbols represent *entirely different concepts!*

Although rarer, there exists an equally perverse alternative notation—of which Chern et al. (1999) provide an exemplar—in which the dual basis is written $\boldsymbol{\theta}^i$, just as we have written it, but in which the *same* Greek letter is used to denote the *connection* 1-forms, written $\boldsymbol{\theta}^i{}_j$!

It is very hard to see these extraordinary notational follies as anything other than *hazing rituals*, designed to test the determination of impudent neophytes who would dare to master the learned treatises of their elders and betters!

We feel compelled to quote one of our scientific heroes, Cornelius Lanczos (1893–1974). In the preface to his (wonderful) *Variational Principles of Mechanics*, Lanczos (1970) writes,

> Many of the scientific treatises of today are formulated in a half-mystical language, as though to impress the reader with the uncomfortable feeling that he is in the permanent presence of a superman. The present book is conceived in a humble spirit and is written for humble people.

Well, tiptoeing in the giant footsteps of Lanczos, *we too* shall speak to you humbly, as plainly and as clearly as possible, *welcoming* you into the glorious garden of Cartan's discoveries. And, for that reason, we shall denote different concepts with different letters.

38.3 The Attitude Matrix

38.3.1 The Connection Forms via the Attitude Matrix

The ***attitude*** in space of the USS Enterprise[7] (or, less interestingly, *any* rigid body) means its *orientation* within $\mathbb{R}^3$.

To describe the attitude of Cartan's moving frame field $\{\mathbf{m}_j\}$, we need only specify the rotation A of the fixed Euclidean frame $\{\mathbf{e}_j\}$ that brings it into coincidence with $\{\mathbf{m}_j\}$.

This rotation is described by the ***attitude matrix*** $[A] = [a_{ij}]$, in which the entries are *functions* of position, since the whole idea is that the attitude of the $\mathbf{m}$-frame *varies* in space. Thus, if we write

$$[\mathbf{e}] \equiv \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix},$$

then

$$\boxed{[\mathbf{m}] = [A]\,[\mathbf{e}].}$$

It follows from our geometrical interpretation of the transpose of a matrix (given in Ex. 12, p. 221) that the matrix of a rotation must satisfy

---

[6]Sadly, my beloved Misner, Thorne, and Wheeler (1973) must *also* be counted amongst these sinners! Saddest of all, the *origin* of this notation was none other than Cartan (1927) himself! Cartan's works have *always* "enjoyed" the reputation of being very hard to understand; I would argue that his choice of notation did not help matters!

[7]To avoid any ambiguity, we refer here to NCC-1701.

$$\boxed{[A]^T = [A]^{-1}.} \tag{38.5}$$

Recall from Linear Algebra that such a matrix is called ***orthogonal***.

It is intuitively clear that the rate of change of [A] determines the rate of change of the **m**-frame, and thus it must determine the connection 1-forms. To derive an explicit formula, let us first agree that the exterior derivative of a matrix is the matrix of the exterior derivatives of its entries:

$$\mathbf{d}[A] = \mathbf{d}[a_{ij}] = [\mathbf{d}a_{ij}].$$

We will now show that the matrix of connection 1-forms is given by

$$\boxed{[\boldsymbol{\omega}] = \big(\mathbf{d}[A]\big)\,[A]^T.} \tag{38.6}$$

Since $[a_{ij}]^T = [a_{ji}]$, the matrix equation (38.6) is equivalent to this component equation:

$$\boxed{\omega_{ij} = \textstyle\sum_k \big(\mathbf{d}a_{ik}\big)\, a_{jk}.} \tag{38.7}$$

To prove this, we first compute the rate of rotation of the moving frame itself:

$$\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i = \boldsymbol{\nabla}_{\mathbf{v}} \sum_k a_{ik}\, \mathbf{e}_k = \sum_k (\mathbf{d}a_{ik})(\mathbf{v})\, \mathbf{e}_k,$$

because $\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{e}_k = 0$.

According to the definition (38.2) of the connection 1-forms, we must now determine how much of this tipping of $\mathbf{m}_i$ is in the direction of $\mathbf{m}_j$. Since

$$\mathbf{m}_j = \sum_l a_{jl}\, \mathbf{e}_l,$$

we find that

$$\mathbf{m}_j \bullet \mathbf{e}_k = \sum_l a_{jl}\, \mathbf{e}_l \bullet \mathbf{e}_k = a_{jk},$$

and therefore,

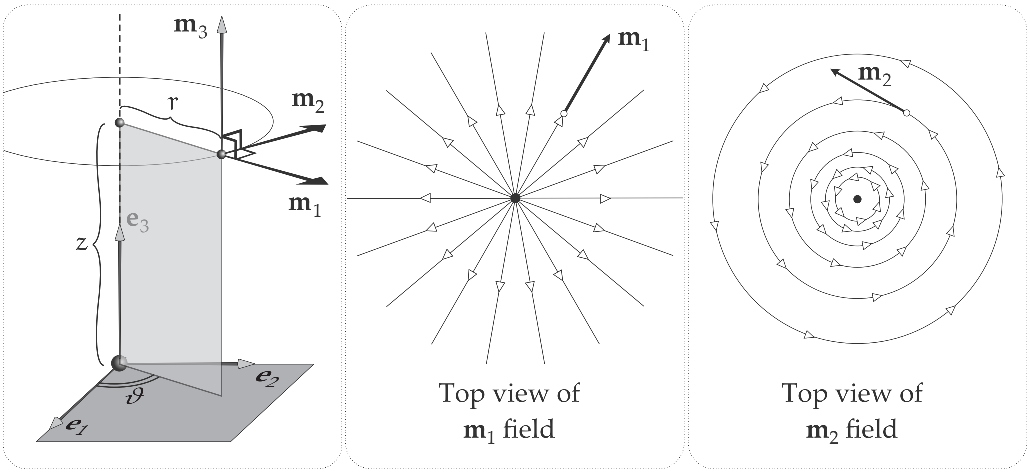
$$\omega_{ij}(\mathbf{v}) \equiv \big(\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i\big) \bullet \mathbf{m}_j = \sum_k (\mathbf{d}a_{ik})(\mathbf{v})\, \mathbf{e}_k \bullet \mathbf{m}_j = \sum_k \big(\mathbf{d}a_{ik}(\mathbf{v})\big)\, a_{jk}.$$

Abstracting away the arbitrary vector **v**, we have therefore proved (38.7), and, with it, the matrix equation (38.6).

38.3.2 Example: The Cylindrical Frame Field

The left panel of [38.1] shows the definition of the ***cylindrical frame field***, derived from the illustrated cylindrical polar coordinate system,[8] (r, ϑ, z). From this figure, one easily deduces [exercise] that

$$\begin{aligned}
\mathbf{m}_1 &= \cos\vartheta\, \mathbf{e}_1 + \sin\vartheta\, \mathbf{e}_2 \\
\mathbf{m}_2 &= -\sin\vartheta\, \mathbf{e}_1 + \cos\vartheta\, \mathbf{e}_2 \\
\mathbf{m}_3 &= \mathbf{e}_3.
\end{aligned}$$



[38.1] *On the left, the definition of the* cylindrical frame field*, based on the illustrated cylindrical polar coordinates,* (r, ϑ, z). *Looking down the z-axis from above, the middle panel shows the* $\mathbf{m}_1$ *field, and the right panel shows the* $\mathbf{m}_2$ *field.*

It is desirable to get a feel for the cylindrical frame field throughout space, rather than at the single illustrated point. To that end, the middle panel of [38.1] shows a top view (looking down the z-axis) of the entire $\mathbf{m}_1$ vector field, which radiates outward (horizontally), away from the the z-axis. Likewise, the panel on the right shows the top view of the $\mathbf{m}_2$ vector field, which swirls around the z-axis. We have not drawn it, but the $\mathbf{m}_3$ vector field is simply a uniform upward flow, of unit speed, parallel to the z-axis.

From the formulas above, we easily read off the attitude matrix:

$$[A] = \begin{bmatrix} \cos\vartheta & \sin\vartheta & 0 \\ -\sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Therefore, (38.6) yields

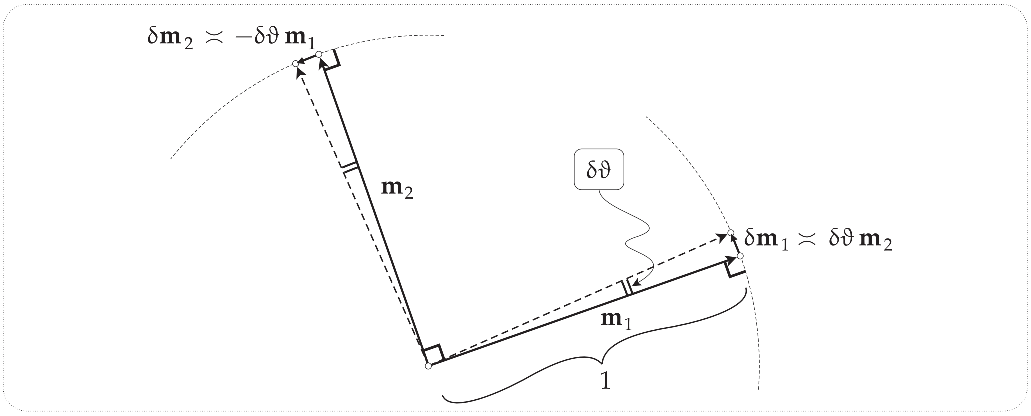
$$\begin{aligned} [\boldsymbol{\omega}] = \big(\mathbf{d}[A]\big)\,[A]^T \quad &= \begin{bmatrix} -\sin\vartheta\,\mathbf{d}\vartheta & \cos\vartheta\,\mathbf{d}\vartheta & 0 \\ -\cos\vartheta\,\mathbf{d}\vartheta & -\sin\vartheta\,\mathbf{d}\vartheta & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\vartheta & -\sin\vartheta & 0 \\ \sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & \mathbf{d}\vartheta & 0 \\ -\mathbf{d}\vartheta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Thus the connection equations (38.3) become

$$\begin{aligned} \nabla_{\mathbf{v}}\mathbf{m}_1 &= \mathbf{d}\vartheta(\mathbf{v})\,\mathbf{m}_2 = (\nabla_{\mathbf{v}}\vartheta)\,\mathbf{m}_2 \\ \nabla_{\mathbf{v}}\mathbf{m}_2 &= -\mathbf{d}\vartheta(\mathbf{v})\,\mathbf{m}_1 = -(\nabla_{\mathbf{v}}\vartheta)\,\mathbf{m}_1 \\ \nabla_{\mathbf{v}}\mathbf{m}_3 &= 0. \end{aligned}$$

We presented this example as very useful practice with the mechanics of the attitude matrix, the connection 1-forms, and the connection equations. However, we can in fact obtain these final three equations directly and geometrically, without any computation.

[8]Here we have switched from the traditional θ to ϑ, in order to avoid any possible confusion with the dual-basis 1-forms, $\boldsymbol{\theta}^j$.



[38.2] *Moving along a short (ultimately vanishing) vector* $\mathbf{v}$ *in space results in a small (ultimately vanishing) rotation* $\delta\vartheta$ *of the orthogonal pair* $(\mathbf{m}_1, \mathbf{m}_2)$. *The tip of* $\mathbf{m}_1$ *moves through distance* $\delta\vartheta$ *along the illustrated arc of the unit circle. Since the tip initially moves in the direction orthogonal to itself,* $\delta\mathbf{m}_1 \asymp \delta\vartheta\, \mathbf{m}_2$. *Likewise,* $\delta\mathbf{m}_2 \asymp -\delta\vartheta\, \mathbf{m}_1$.

The last equation is trivial: $\mathbf{m}_3 = \mathbf{e}_3$ is constant, so of course its derivative vanishes. The first two equations are much more interesting, and we now derive them geometrically.

First, we observe that if $\mathbf{v}$ is very short, and ultimately vanishing, then

$$\mathbf{d}\vartheta(\mathbf{v}) \asymp \delta\vartheta \equiv \text{change in } \vartheta \text{ from tail to tip of } \mathbf{v}.$$

Next, by the same token,

$$\begin{aligned} \boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_j &\asymp \delta\mathbf{m}_j \\ &\equiv \text{change in } \mathbf{m}_j \text{ from tail to tip of } \mathbf{v} \\ &= \text{change in } \mathbf{m}_j \text{ brought about by the change } \delta\vartheta, \\ & \text{resulting from the movement along } \mathbf{v}. \end{aligned}$$

Now consider [38.2], which shows the orthogonal pair $(\mathbf{m}_1, \mathbf{m}_2)$ rotating rigidly together through $\delta\vartheta$. Focus on the change $\delta\mathbf{m}_1$. The tip of $\mathbf{m}_1$ moves through distance $\delta\vartheta$ along the illustrated arc of the unit circle. Initially, the tip moves in the direction orthogonal to itself, in other words, in the direction of $\mathbf{m}_2$, as illustrated. Therefore, we have proved the first of the two connection equations, in the form

$$\delta\mathbf{m}_1 \asymp \delta\vartheta\, \mathbf{m}_2.$$

The same diagram simultaneously proves the second connection equation:

$$\delta\mathbf{m}_2 \asymp -\delta\vartheta\, \mathbf{m}_1.$$

38.4 Cartan's Two Structural Equations

38.4.1 The Duals $\boldsymbol{\theta}^i$ *of* $\mathbf{m}_i$ *in Terms of the Duals* $\mathbf{d}x^j$ *of* $\mathbf{e}_j$

Recall that (Cartan's) IDEA **5** is to replace the moving frame field $\mathbf{m}_i$ with its dual 1-form basis $\boldsymbol{\theta}^i$.

We know that (by definition) the attitude matrix $[a_{ij}]$ transforms the fixed Euclidean basis $\{\mathbf{e}_j\}$ into Cartan's moving frame field, according to

$$\mathbf{m}_i = \sum_k a_{ik}\,\mathbf{e}_k.$$

We now show that the *same* rotation matrix transforms the 1-form basis $\{\mathbf{d}x^j\}$ dual to $\{\mathbf{e}_j\}$ into the 1-form basis $\{\boldsymbol{\theta}^i\}$ dual to $\{\mathbf{m}_i\}$:

$$\boxed{\boxed{[\boldsymbol{\theta}] = [A]\,[\mathbf{d}x].}} \tag{38.8}$$

Expressed in terms of components,

$$\boxed{\boldsymbol{\theta}^i = \textstyle\sum_k a_{ij}\,\mathbf{d}x^j.} \tag{38.9}$$

To prove this, first recall from (32.14), page 357, that *any* 1-form φ can be decomposed into components in the $\{\mathbf{d}x^j\}$ basis by applying φ to the Euclidean basis vectors. Thus, in particular,

$$\boldsymbol{\theta}^i = \sum_j \boldsymbol{\theta}^i(\mathbf{e}_j)\,\mathbf{d}x^j.$$

But because $\boldsymbol{\theta}^i$ is dual to $\mathbf{m}_i$,

$$\boldsymbol{\theta}^i(\mathbf{e}_j) = \mathbf{m}_i \bullet \mathbf{e}_j = \left[\sum_k a_{ik}\,\mathbf{e}_k\right] \bullet \mathbf{e}_j = a_{ij},$$

completing the proof of (38.9).

38.4.2 Cartan's First Structural Equation

The connection equations (38.3) tell us the rotation of the moving frame field $\mathbf{m}_i$ as we move in an arbitrary direction $\mathbf{v}$. It is therefore natural to ask how the dual 1-forms $\boldsymbol{\theta}^i$ vary in space. Since the exterior derivative measures the rate of change in *all* directions (simultaneously!), it is $\mathbf{d}[\boldsymbol{\theta}]$ that we seek.

The remarkably elegant answer is ***Cartan's First Structural Equation***:

$$\boxed{\boxed{\boxed{\mathbf{d}[\boldsymbol{\theta}] = [\boldsymbol{\omega}] \wedge [\boldsymbol{\theta}],}}} \tag{38.10}$$

where

$$[\boldsymbol{\theta}] \equiv \begin{bmatrix} \boldsymbol{\theta}^1 \\ \boldsymbol{\theta}^2 \\ \boldsymbol{\theta}^3 \end{bmatrix}.$$

The proof is a simple computation. Writing

$$[\mathbf{d}x] \equiv \begin{bmatrix} \mathbf{d}x_1 \\ \mathbf{d}x_2 \\ \mathbf{d}x_3 \end{bmatrix},$$

(38.9) becomes

$$[\boldsymbol{\theta}] = [A][\mathbf{d}x].$$

Differentiating, and using first (38.5) and then (38.6),

$$\mathbf{d}[\boldsymbol{\theta}] = \mathbf{d}[A] \wedge [\mathbf{d}x] = \mathbf{d}[A][A]^T \wedge [A][\mathbf{d}x] = [\boldsymbol{\omega}] \wedge [\boldsymbol{\theta}].$$

Done!

When performing concrete calculations, it is sometimes useful to rephrase (38.10) in terms of its components:

$$\boxed{\mathbf{d}\boldsymbol{\theta}^i = \textstyle\sum_j \boldsymbol{\omega}_{ij} \wedge \boldsymbol{\theta}^j} \tag{38.11}$$

38.4.3 Cartan's Second Structural Equation

Now let us ask how the connection 1-forms vary in space. Again, the exterior derivative simultaneously encodes the variation in all possible directions, so it is $\mathbf{d}[\boldsymbol{\omega}]$ that we seek.

The remarkably elegant answer is ***Cartan's Second Structural Equation***:

$$\boxed{\mathbf{d}[\boldsymbol{\omega}] = [\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}].} \tag{38.12}$$

As we shall see in Section 38.12, the deeper geometric meaning of this equation is that characterized the lack of curvature, the *flatness* of the Euclidean space $\mathbb{R}^3$ within which we have been operating up till now. When we move on to *curved* manifolds, such as Einstein's gravitationally warped spacetime, the two sides of this equation are no longer equal, and in fact the *difference* between the two sides will be found to encode the curvature of the manifold.

To prove the equation, we require three simple lemmas. First, note that if f and g are functions, then

$$\mathbf{d}(\mathbf{d}f\, g) = \mathbf{d}(g\, \mathbf{d}f) = \mathbf{d}g \wedge \mathbf{d}f = -\mathbf{d}f \wedge \mathbf{d}g.$$

Next, recall that the geometrical interpretation of the transpose (given in Ex. 12, p. 221) implies that $(AB)^T = B^T A^T$. Finally, recall that $[\boldsymbol{\omega}]$ is skew-symmetric: $[\boldsymbol{\omega}]^T = -[\boldsymbol{\omega}]$.

Therefore, differentiating (38.6),

$$\begin{aligned}
\mathbf{d}[\boldsymbol{\omega}] &= \mathbf{d}\big[\big(\mathbf{d}[A]\big)\,[A]^T\big] \\
&= -\big(\mathbf{d}[A]\big) \wedge \mathbf{d}[A]^T \\
&= -\big(\mathbf{d}[A]\big)[A]^T \wedge [A]\,\mathbf{d}[A]^T \\
&= -\big(\mathbf{d}[A]\big)[A]^T \wedge \big[\big(\mathbf{d}[A]\big)[A]^T\big]^T \\
&= -[\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}]^T \\
&= [\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}].
\end{aligned}$$

Done!

When performing concrete calculations, it is sometimes useful to rephrase (38.12) in terms of its components:

$$\boxed{\mathbf{d}\boldsymbol{\omega}_{ij} = \textstyle\sum_k \boldsymbol{\omega}_{ik} \wedge \boldsymbol{\omega}_{kj}} \tag{38.13}$$

Finally, note that it is easy to remember *both* Structural Equations at once, by remembering just this:

$$\mathbf{d}[\mathbf{?}] = [\boldsymbol{\omega}] \wedge [\mathbf{?}].$$

38.4.4 *Example: The Spherical Frame Field*

In this section we shall find the dual basis and the connection Forms associated with the spherical polar coordinates[9] (r, ϕ, ϑ) of Section 35.4, and we shall use them to test/illustrate Cartan's Structural Equations.

In the following, we shall write $c_\vartheta \equiv \cos\vartheta$, $s_\phi \equiv \sin\phi$, etc., in part to save space, but also because we find this to be a clarifying notation in private, pencil-and-paper calculations.

From [38.3], we see that,

$$[x] = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \rho\, c_\vartheta \\ \rho\, s_\vartheta \\ r\, c_\phi \end{bmatrix} = \begin{bmatrix} r\, s_\phi\, c_\vartheta \\ r\, s_\phi\, s_\vartheta \\ r\, c_\phi \end{bmatrix}$$

From this same figure we also deduce [exercise] that the illustrated ***spherical frame field*** is given by

$$[\mathbf{m}] = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \end{bmatrix} = \begin{bmatrix} s_\phi \left[c_\vartheta \mathbf{e}_1 + s_\vartheta \mathbf{e}_2\right] + c_\phi \mathbf{e}_3 \\ c_\phi \left[c_\vartheta \mathbf{e}_1 + s_\vartheta \mathbf{e}_2\right] - s_\phi \mathbf{e}_3 \\ -s_\vartheta \mathbf{e}_1 + c_\vartheta \mathbf{e}_2 \end{bmatrix} = [A] \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix},$$

from which we immediately deduce the attitude matrix:

$$[A] = \begin{pmatrix} s_\phi\, c_\vartheta & s_\phi\, s_\vartheta & c_\phi \\ c_\phi\, c_\vartheta & c_\phi\, s_\vartheta & -s_\phi \\ -s_\vartheta & c_\vartheta & 0 \end{pmatrix}.$$

At this point we can calculate the connection Forms using (38.6):

$$[\boldsymbol{\omega}] = \big(\mathbf{d}[A]\big)\, [A]^T.$$

Clearly, this will require *lengthy* and arduous computations, which we hereby assign as a (cruel)[10] exercise! However, let us at least offer some concrete guidance on *how* to carry out such computations.

First, we certainly *don't* need to multiply the *entire* matrix $\mathbf{d}[A]$ by the *entire* matrix $[A]$. For example, calculating a diagonal entry is a waste of time: we know it must be zero! In fact we *only* need to calculate the three entries in the top-right corner of the product: $\boldsymbol{\omega}_{12}$, $\boldsymbol{\omega}_{13}$, $\boldsymbol{\omega}_{23}$.

In order to target just these three entries, it is simpler to use the component form of the matrix product, namely, (38.7):

$$\boldsymbol{\omega}_{ij} = \sum_k \big(\mathbf{d}a_{ik}\big)\, a_{jk}.$$

Thus we never need to actually form the transpose of $[A]$. Instead, we interpret this formula as telling us that

> *To obtain $\boldsymbol{\omega}_{ij}$ we take $\mathbf{d}$ of the i-th row of $[A]$, then multiply each element by the corresponding one in the j-th row of $[A]$.*

---

[9]O'Neill (2006, pp. 86, 97) instead measures ϕ from the equator, so our ϕ is obtained by subtracting his from $(\pi/2)$. This has the effect of interchanging $\sin\phi$ and $\cos\phi$. Also, in order to make the frame right-handed, the *order* of O'Neill's coordinates is (r, ϑ, ϕ), and his spherical frame field, written $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$, is therefore related to ours by $(\mathbf{m}_1 = \mathbf{F}_1, \mathbf{m}_2 = -\mathbf{F}_3, \mathbf{m}_3 = \mathbf{F}_2)$.

[10]When forced to assign such a cruel computation to my USF students, I simultaneously assign them a viewing of Tarentino's *Kill Bill, Vol. 2, Chapter 8: The Cruel Tutelage of Pai Mei*, for it demonstrates how such "cruelty" ultimately *saved* Beatrix Kiddo from the grave!

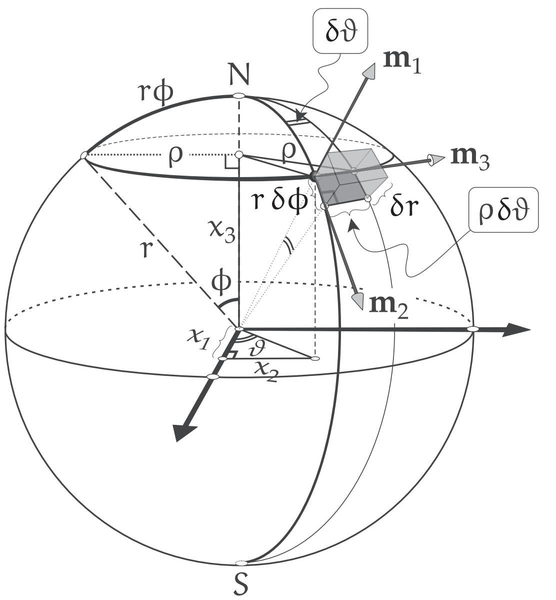
For example, to find $\boldsymbol{\omega}_{12}$, we must take $\mathbf{d}$ of the first row, and multiply it by the second row:

$$\boldsymbol{\omega}_{12} = \left[\mathbf{d}(s_\phi\, c_\vartheta)\right] c_\phi\, c_\vartheta + \left[\mathbf{d}(s_\phi\, s_\vartheta)\right] c_\phi\, s_\vartheta - \left[\mathbf{d}(c_\phi)\right] s_\phi.$$

Even if you don't feel up to calculating all three connection Forms, we urge you to at least do this one. After taking the derivatives, you will be faced with five terms, each of which is the product of four trigonometric functions. But then, "miraculously," everything *cancels or simplifies*, and we are left with a beautifully simple answer: $\boldsymbol{\omega}_{12} = \mathbf{d}\phi$!

Equally "miraculous" cancellations and simplifications occur in the calculation of the other two connection Forms, and we obtain the following simple formulas:

$$\boxed{\begin{aligned} \boldsymbol{\omega}_{12} &= \mathbf{d}\phi \\ \boldsymbol{\omega}_{13} &= \sin\phi\, \mathbf{d}\vartheta \\ \boldsymbol{\omega}_{23} &= \cos\phi\, \mathbf{d}\vartheta. \end{aligned}} \tag{38.14}$$



[38.3] *Geometric proof that* $\boldsymbol{\theta}_1 = \mathbf{d}r$, $\boldsymbol{\theta}_2 = r\,\mathbf{d}\phi$, *and* $\boldsymbol{\theta}_3 = r\, \sin\phi\, \mathbf{d}\vartheta$.

Recalling what was said in the Prologue, we have here three examples of what we call "false miracles": if all those terms cancelled, they should never have been there in the first place—we must be looking at the mathematics in the *wrong way!*

By this point in our drama it can surely come as no surprise to learn that the *correct* way to look at this is *geometrically*. *All* of the painful calculations you just performed are thereby avoided, and the results are obtained directly and intuitively! We shall provide this Newtonian, geometrical explanation at the end of this section, but for now we press on in our quest to test Cartan's Structural Equations.

To that end, let us find the dual-basis 1-forms, $\boldsymbol{\theta}^i$. Again, we can certainly obtain these by blind computation: we know [A] and we know [x], so we can calculate $\mathbf{d}[x] = [\mathbf{d}x]$, and hence we can use (38.8):

$$[\boldsymbol{\theta}] = [A]\,[\mathbf{d}x].$$

But, *again*, this will clearly require a *lengthy* computation, which we hereby assign as *another* cruel exercise! (You will appreciate what comes next so much more if you actually *carry out* these calculations, which are filled with many cancellations and simplifications. Having successfully hacked your way through this dense thicket of symbols, you should emerge into a clearing, within which you will find these simple formulas: (38.16).)

Here, too, Newtonian geometry allows us leap through the air, flying over the lengthy and opaque calculations beneath us, landing directly at the answers!

Consider the shaded box in [38.3] in the limit that it shrinks and ultimately vanishes. Its edges are ultimately equal to

$$\delta r\, \mathbf{m}_1, \qquad r\,\delta\phi\, \mathbf{m}_2, \qquad \rho\,\delta\vartheta\, \mathbf{m}_3 = r\, s_\phi\, \delta\vartheta\, \mathbf{m}_3,$$

as illustrated. Therefore, the diagonal of the box, which points in an *arbitrary* direction, is the short (ultimately vanishing) vector **v**, given by

$$\boxed{\mathbf{v} \asymp \delta r\,\mathbf{m}_1 + r\,\delta\phi\,\mathbf{m}_2 + r\,s_\phi\,\delta\vartheta\,\mathbf{m}_3.} \tag{38.15}$$

By definition of the dual basis, $\boldsymbol{\theta}^i(\mathbf{m}_j) = \delta^i_j$, so,

$$\boldsymbol{\theta}^1(\mathbf{v}) \asymp \delta r \asymp \mathbf{d}r(\mathbf{v}) \qquad \Longrightarrow \qquad \boldsymbol{\theta}^1 = \mathbf{d}r.$$

This follows from the fact that **v** is *arbitrary*, and therefore can be abstracted away.

We find $\boldsymbol{\theta}^2$ and $\boldsymbol{\theta}^3$ in exactly the same way. Therefore, the geometry of [38.3] immediately yields

$$\boxed{[\boldsymbol{\theta}] = \begin{bmatrix} \boldsymbol{\theta}^1 \\ \boldsymbol{\theta}^2 \\ \boldsymbol{\theta}^3 \end{bmatrix} = \begin{bmatrix} \mathbf{d}r \\ r\,\mathbf{d}\phi \\ r\,s_\phi\,\mathbf{d}\vartheta \end{bmatrix}.} \tag{38.16}$$

Contrast this geometry with the calculations you just performed!

Now that we know the dual basis and the connection Forms, let us check Cartan's *First* Structural Equation, (38.10),

$$\mathbf{d}[\boldsymbol{\theta}] = [\boldsymbol{\omega}] \wedge [\boldsymbol{\theta}].$$

On the left-hand side, we find

$$\mathbf{d}[\boldsymbol{\theta}] = \mathbf{d}\begin{bmatrix} \mathbf{d}r \\ r\,\mathbf{d}\phi \\ r\,s_\phi\,\mathbf{d}\vartheta \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{d}r \wedge \mathbf{d}\phi \\ s_\phi\,\mathbf{d}r \wedge \mathbf{d}\vartheta + r\,c_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta \end{bmatrix}.$$

On the right-hand side, we find

$$\begin{aligned}
[\boldsymbol{\omega}] \wedge [\boldsymbol{\theta}] &= \begin{bmatrix} 0 & \boldsymbol{\omega}_{12} & \boldsymbol{\omega}_{13} \\ -\boldsymbol{\omega}_{12} & 0 & \boldsymbol{\omega}_{23} \\ -\boldsymbol{\omega}_{13} & -\boldsymbol{\omega}_{23} & 0 \end{bmatrix} \wedge \begin{bmatrix} \boldsymbol{\theta}^1 \\ \boldsymbol{\theta}^2 \\ \boldsymbol{\theta}^3 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \mathbf{d}\phi & s_\phi\,\mathbf{d}\vartheta \\ -\mathbf{d}\phi & 0 & c_\phi\,\mathbf{d}\vartheta \\ -s_\phi\,\mathbf{d}\vartheta & -c_\phi\,\mathbf{d}\vartheta & 0 \end{bmatrix} \wedge \begin{bmatrix} \mathbf{d}r \\ r\,\mathbf{d}\phi \\ r\,s_\phi\,\mathbf{d}\vartheta \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ \mathbf{d}r \wedge \mathbf{d}\phi \\ s_\phi\,\mathbf{d}r \wedge \mathbf{d}\vartheta + r\,c_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta \end{bmatrix},
\end{aligned}$$

thereby verifying Cartan's First Structural Equation for the spherical frame field.

Now let us check Cartan's *Second* Structural Equation, (38.12),

$$\mathbf{d}[\boldsymbol{\omega}] = [\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}].$$

On the left-hand side, we find

$$\mathbf{d}\begin{bmatrix} 0 & \mathbf{d}\phi & s_\phi\,\mathbf{d}\vartheta \\ -\mathbf{d}\phi & 0 & c_\phi\,\mathbf{d}\vartheta \\ -s_\phi\,\mathbf{d}\vartheta & -c_\phi\,\mathbf{d}\vartheta & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & c_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta \\ 0 & 0 & -s_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta \\ -c_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta & s_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta & 0 \end{bmatrix}.$$

On the right-hand side, we find

$$
\begin{aligned}
[\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}] &= \begin{bmatrix} 0 & \mathbf{d}\phi & s_\phi\,\mathbf{d}\vartheta \\ -\mathbf{d}\phi & 0 & c_\phi\,\mathbf{d}\vartheta \\ -s_\phi\,\mathbf{d}\vartheta & -c_\phi\,\mathbf{d}\vartheta & 0 \end{bmatrix} \wedge \begin{bmatrix} 0 & \mathbf{d}\phi & s_\phi\,\mathbf{d}\vartheta \\ -\mathbf{d}\phi & 0 & c_\phi\,\mathbf{d}\vartheta \\ -s_\phi\,\mathbf{d}\vartheta & -c_\phi\,\mathbf{d}\vartheta & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & c_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta \\ 0 & 0 & -s_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta \\ -c_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta & s_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta & 0 \end{bmatrix},
\end{aligned}
$$

thereby verifying Cartan's Second Structural Equation for the spherical frame field.

As promised, we end this section by returning to the connection Forms, showing that the lengthy calculations that led to (38.14) can be replaced with simple, clear, Newtonian geometry.

To do so, we shall apply Newtonian ultimate equalities *directly* to the definition (38.2):

$\boldsymbol{\omega}_{ij}(\mathbf{v}) \equiv (\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i)\boldsymbol{\cdot}\mathbf{m}_j$ *is the initial rate at which* $\mathbf{m}_i$ *tips towards* $\mathbf{m}_j$ *as the frame moves along* $\mathbf{v}$.

Here we shall take $\mathbf{v}$ to be the arbitrary, short (ultimately vanishing) vector given by (38.15):

$$\boxed{\mathbf{v} \asymp \delta r\,\mathbf{m}_1 + r\,\delta\phi\,\mathbf{m}_2 + r\,s_\phi\,\delta\vartheta\,\mathbf{m}_3.}$$

The key tool in the following analysis is the fact that, since $\mathbf{v}$ is short, and ultimately vanishing,

$$\boxed{\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_i \asymp \text{change in } \mathbf{m}_i \text{ from tail to tip of } \mathbf{v} \equiv \delta\mathbf{m}_i.}$$

Note that in order to easily visualize the changes $\delta\mathbf{m}_i$, *we shall draw all the vectors* $\mathbf{m}_i$ *as emanating from a common point; then* $\delta\mathbf{m}_i$ *is simply the movement of the tip of* $\mathbf{m}_i$.

We begin with $\boldsymbol{\omega}_{12}(\mathbf{v}) \equiv (\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_1)\boldsymbol{\cdot}\mathbf{m}_2$. Consider the changes $\delta\mathbf{m}_1$ in $\mathbf{m}_1$ that result from moving along *each* of the three components of $\mathbf{v}$, separately:

• Moving radially outward along $\delta r\,\mathbf{m}_1$ keeps $\mathbf{m}_1$ *constant*: it does not tip towards $\mathbf{m}_2$ (or anything else!).

• Moving along $r\,s_\phi\,\delta\vartheta\,\mathbf{m}_3$ (due East, along the circle of latitude) causes $\mathbf{m}_1$ to rotate around the cone shown in the top half of [38.5], of semivertical angle ϕ and slant height 1, and therefore base radius s_ϕ. Evidently, $\delta\mathbf{m}_1$ is orthogonal to $\mathbf{m}_2$, so this too makes zero contribution to $\boldsymbol{\omega}_{12}$.

In greater detail, we note for future use that the figure demonstrates that

$$\delta\mathbf{m}_1 = s_\phi\,\delta\vartheta\,\mathbf{m}_3. \tag{38.17}$$

• As an aside, for future use, we note that this rotation causes $\mathbf{m}_2$ to rotate around the cone shown in the bottom half of [38.5], of semivertical angle $\left(\frac{\pi}{2} - \phi\right)$ and slant height 1, and therefore base radius c_ϕ. Thus, as illustrated,

$$\delta\mathbf{m}_2 = c_\phi\,\delta\vartheta\,\mathbf{m}_3. \tag{38.18}$$

• Returning our focus to $\mathbf{m}_1$, the *only* part of $\mathbf{v}$ that *does* cause $\mathbf{m}_1$ to tip towards $\mathbf{m}_2$ is the southward movement $r\,\delta\phi\,\mathbf{m}_2$ towards the equator, along the meridian. Figure [38.4] is drawn in the vertical plane of this meridian, containing $\mathbf{m}_1$ and $\mathbf{m}_2$. As $\mathbf{m}_1$ rotates through angle $\delta\phi$, its tip ultimately moves along $\delta\mathbf{m}_1 \asymp \delta\phi\,\mathbf{m}_2$, as illustrated. Therefore,

$$\boldsymbol{\omega}_{12}(\mathbf{v}) \asymp \delta\mathbf{m}_1\boldsymbol{\cdot}\mathbf{m}_2 \asymp (\delta\phi\,\mathbf{m}_2)\boldsymbol{\cdot}\mathbf{m}_2 = \delta\phi \asymp \mathbf{d}\phi(\mathbf{v}) \quad\Longrightarrow\quad \boldsymbol{\omega}_{12} = \mathbf{d}\phi.$$

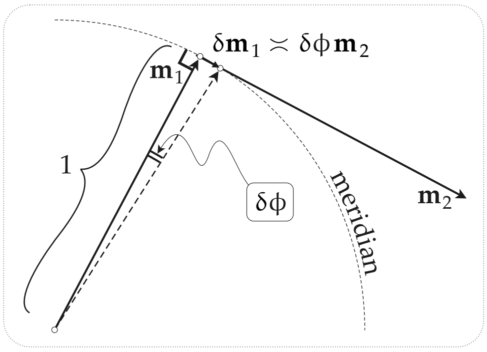
Next, consider

$$\omega_{13}(\mathbf{v}) \equiv (\nabla_{\mathbf{v}}\mathbf{m}_1) \cdot \mathbf{m}_3.$$

Here we are considering the *same* changes $\delta\mathbf{m}_1$ in $\mathbf{m}_1$ as above, but now we want to know how much $\delta\mathbf{m}_1$ points in the direction of $\mathbf{m}_3$ (instead of $\mathbf{m}_2$).

As we have seen, the only part of $\mathbf{v}$ that produces a $\delta\mathbf{m}_1$ along $\mathbf{m}_3$ is the movement along $r\, s_\phi\, \delta\vartheta\, \mathbf{m}_3$. From (38.17) we immediately deduce that

$$\omega_{13}(\mathbf{v}) \asymp \delta\mathbf{m}_1 \cdot \mathbf{m}_3 \asymp (s_\phi\, \delta\vartheta\, \mathbf{m}_3) \cdot \mathbf{m}_3$$
$$= s_\phi\, \delta\vartheta \asymp s_\phi\, \mathbf{d}\vartheta(\mathbf{v}) \quad \Longrightarrow \quad \omega_{13} = s_\phi\, \mathbf{d}\vartheta.$$

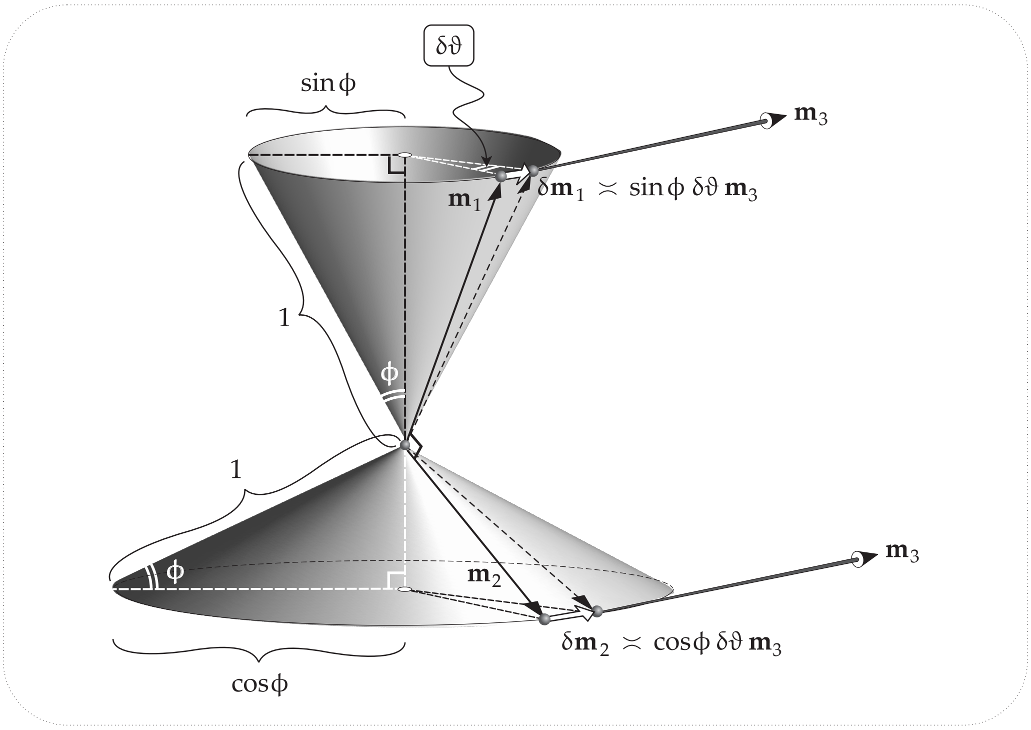


[38.4] *Geometric proof that* $\omega_{12} = \mathbf{d}\phi$.

Finally, consider $\omega_{23}(\mathbf{v}) \equiv (\nabla_{\mathbf{v}}\mathbf{m}_2) \cdot \mathbf{m}_3$. It is easy to see [exercise] that the only component of $\mathbf{v}$ that causes $\mathbf{m}_2$ to tip in the direction of $\mathbf{m}_3$ is the movement along $r\, s_\phi\, \delta\vartheta\, \mathbf{m}_3$ (due East, along the circle of latitude). Thus, using (38.18), illustrated in the bottom cone of [38.5], we deduce that

$$\omega_{23}(\mathbf{v}) \asymp \delta\mathbf{m}_2 \cdot \mathbf{m}_3 \asymp (c_\phi\, \delta\vartheta\, \mathbf{m}_3) \cdot \mathbf{m}_3 = c_\phi\, \delta\vartheta \asymp c_\phi\, \mathbf{d}\vartheta(\mathbf{v}) \quad \Longrightarrow \quad \omega_{23} = c_\phi\, \mathbf{d}\vartheta.$$

This completes our geometrical proof of (38.14). Again, contrast the illuminating simplicity and directness of this Newtonian geometrical reasoning with the three long, unilluminating calculations you were originally forced to perform in order to arrive at these three formulas!



[38.5] *Geometric proof of* $\omega_{13} = \sin\phi\, \mathbf{d}\vartheta$ *(top cone), and of* $\omega_{23} = \cos\phi\, \mathbf{d}\vartheta$ *(bottom cone).*

38.5 The Six Fundamental Form Equations of a Surface

38.5.1 Adapting Cartan's Moving Frame to a Surface: The Shape Operator and the Extrinsic Curvature

We now implement Darboux's **IDEA 3** (adapt the frame to a surface) in combination with Cartan's **IDEA 6**: use Forms, specializing Cartan's two Structural Equations to this adapted frame field.

Given a surface $\mathcal{S}$ with normal $\mathbf{n}$, we *choose* $\mathbf{m}_3 = \mathbf{n}$; this is called the ***adapted frame field***. Then $(\mathbf{m}_1, \mathbf{m}_2)$ *becomes an intrinsic (i.e., tangent) basis for the vectors* $\mathbf{v}$ *"within"* $\mathcal{S}$.

Henceforth, we shall only *consider such vectors* $\mathbf{v}$ *that are tangent to* $\mathcal{S}$. When applied to such a tangent vector, the connection equations (38.4) yield the following simple expression for the Shape Operator S ((15.4), p. 151) in terms of the connection Forms:

$$\boxed{\mathsf{S}(\mathbf{v}) = \boldsymbol{\omega}_{13}(\mathbf{v})\,\mathbf{m}_1 + \boldsymbol{\omega}_{23}(\mathbf{v})\,\mathbf{m}_2.} \tag{38.19}$$

The proof goes as follows:

$$\mathsf{S}(\mathbf{v}) = -\nabla_{\mathbf{v}}\mathbf{n} = -\nabla_{\mathbf{v}}\mathbf{m}_3 = -\left[\nabla_{\mathbf{v}}\mathbf{m}_3 \bullet \mathbf{m}_1\right]\mathbf{m}_1 - \left[\nabla_{\mathbf{v}}\mathbf{m}_3 \bullet \mathbf{m}_2\right]\mathbf{m}_2.$$

The last step is to insert the definitions of the connection 1-forms, (38.2), taking note of their skew-symmetry.

The first column of the *matrix* [S] representing the Shape Operator is the image under S of the first basis vector, namely, $\mathsf{S}(\mathbf{m}_1)$. Likewise, the second column is the image of the second basis vector. Therefore,

$$\boxed{[\mathsf{S}] = \begin{bmatrix} \boldsymbol{\omega}_{13}(\mathbf{m}_1) & \boldsymbol{\omega}_{13}(\mathbf{m}_2) \\ \boldsymbol{\omega}_{23}(\mathbf{m}_1) & \boldsymbol{\omega}_{23}(\mathbf{m}_2) \end{bmatrix}.} \tag{38.20}$$

Let us now use this to express the extrinsic curvature $\mathcal{K}_{\text{ext}} = \kappa_1 \kappa_2$ in terms of the connection Forms. Since we intend to *reprove* fundamental results from the first four Acts, we must avoiding making *assumptions* based on this prior knowledge. Thus we must revert to distinguishing between extrinsic and intrinsic measurements of curvature. Only then can we prove (and appreciate!) Gauss's *Theorema Egregium*, which reveals that these two measures of curvature are actually the *same*!

On the other hand, it is certainly *not* our intention to use Forms to rebuild the entire edifice of Differential Geometry from scratch! So long as we are explicit about which former results are being invoked, and circular reasoning is scrupulously avoided, no harm can arise from this hybrid approach.

Thus, now, recall from (15.8), page 153, that the *extrinsic* curvature $\mathcal{K}_{\text{ext}}$ of $\mathcal{S}$ measures the *spreading* of the normal vectors $\mathbf{m}_3$ over a small patch of surface. More precisely, it is the local area expansion factor of S. In other words, it is the *determinant* of the matrix [S] representing S:

$$\mathcal{K}_{\text{ext}} = \det[\mathsf{S}] = \boldsymbol{\omega}_{13}(\mathbf{m}_1)\,\boldsymbol{\omega}_{23}(\mathbf{m}_2) - \boldsymbol{\omega}_{13}(\mathbf{m}_2)\,\boldsymbol{\omega}_{23}(\mathbf{m}_1).$$

We have arrived at an expression for the extrinsic curvature that will soon prove its importance:

$$\boxed{\mathcal{K}_{\text{ext}} = \left(\boldsymbol{\omega}_{13} \wedge \boldsymbol{\omega}_{23}\right)(\mathbf{m}_1, \mathbf{m}_2).} \tag{38.21}$$

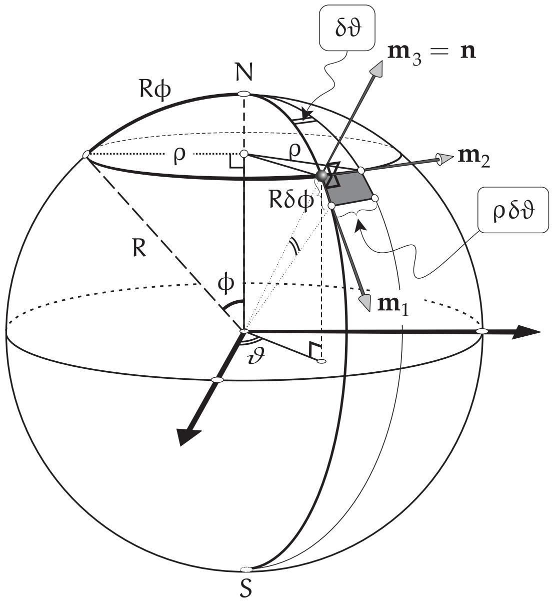
Finally, consider the 1-form basis $\boldsymbol{\theta}^{i}$ dual to the adapted frame field $(\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3)$. These give the coordinates $\boldsymbol{\theta}^{i}(\mathbf{v}) = \mathbf{v} \bullet \mathbf{m}_i$ of any tangent vector $\mathbf{v}$ with respect to the adapted frame field. As with the connection Forms, these dual-basis Forms will be applied *only* to tangent vectors to $\mathcal{S}$, so that they become Forms *on* $\mathcal{S}$. O'Neill (2006, p. 266) states the implication perfectly: "this restriction is fatal to $\boldsymbol{\theta}^3$, for if $\mathbf{v}$ is tangent to $\mathcal{S}$, it is orthogonal to $\mathbf{m}_3$, so $\boldsymbol{\theta}^3(\mathbf{v}) = \mathbf{v} \bullet \mathbf{m}_3 = 0$. Thus $\boldsymbol{\theta}^3$ is identically zero on $\mathcal{S}$."

38.5.2 Example: The Sphere

In order to adapt the spherical frame field to a particular sphere, $r = R$, we must relabel the vectors shown in [38.3]. The radial vector (which *was* $\mathbf{m}_1$) is the normal to the sphere, and therefore (according to our convention) it must be relabelled as $\mathbf{m}_3$. The remaining two members of the original frame field, tangent to the sphere, must likewise be relabelled, as shown in [38.6], subject to the ***right-hand rule***: $\mathbf{m}_1$ points South, and $\mathbf{m}_2$ points East.

It follows that the previously obtained 1-form basis $\boldsymbol{\theta}^{i}$ and the connection Forms $\boldsymbol{\omega}_{ij}$ are *now* given by,[11]

$$\begin{aligned} \boldsymbol{\theta}^1 &= R\,\mathbf{d}\phi & \boldsymbol{\omega}_{12} &= c_\phi\,\mathbf{d}\vartheta \\ \boldsymbol{\theta}^2 &= R\,s_\phi\,\mathbf{d}\vartheta & \boldsymbol{\omega}_{13} &= -\mathbf{d}\phi \\ & & \boldsymbol{\omega}_{23} &= -s_\phi\,\mathbf{d}\vartheta \end{aligned}$$



[38.6] *The* **adapted frame field** *has (by definition)* $\mathbf{m}_3 = \mathbf{n}$. *The resulting changes of names imply that, now,* $\boldsymbol{\theta}^1 = R\,\mathbf{d}\phi$, *and* $\boldsymbol{\theta}^2 = \rho\,\mathbf{d}\vartheta = R\,\sin\phi\,\mathbf{d}\vartheta$.

As valuable practice with the geometrical methods we have just explained, try checking each of these five results, simply by *inspecting* [38.6]—no calculations required!

Finally, note that the shaded element of area is $\mathcal{A} = (R\,\delta\phi)(\rho\,\delta\vartheta) = R^2 \sin\vartheta\,\delta\phi\,\delta\vartheta$. Thus, the ***area 2-form of the sphere***—or, indeed, of *any* surface—is

$$\boxed{\mathcal{A} = \boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2.} \tag{38.22}$$

38.5.3 Uniqueness of Basis Decompositions

Let us briefly review the uniqueness of the decomposition of 1-forms and 2-forms relative to a basis; both these facts will be needed in short order.

For 1-forms within $\mathcal{S}$, our basis is $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$. Recall that *two 1-forms $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are equal if and only if*

$$\boldsymbol{\varphi}(\mathbf{m}_1) = \boldsymbol{\psi}(\mathbf{m}_1) \qquad \text{and} \qquad \boldsymbol{\varphi}(\mathbf{m}_2) = \boldsymbol{\psi}(\mathbf{m}_2).$$

It follows immediately [exercise] that $\boldsymbol{\varphi}$ can be *uniquely* decomposed as

$$\boxed{\boldsymbol{\varphi} = \boldsymbol{\varphi}(\mathbf{m}_1)\,\boldsymbol{\theta}^1 + \boldsymbol{\varphi}(\mathbf{m}_2)\,\boldsymbol{\theta}^2.} \tag{38.23}$$

[11] Again, these do *not* match the results you will find in O'Neill (2006, p. 267), because his different method of defining ϕ changes *everything*.

Likewise, *two 2-forms* $\boldsymbol{\Phi}$ *and* $\boldsymbol{\Psi}$ *are equal if and only if*

$$\boldsymbol{\Phi}(\mathbf{m}_1, \mathbf{m}_2) = \boldsymbol{\Psi}(\mathbf{m}_1, \mathbf{m}_2).$$

It follows immediately [exercise] that $\boldsymbol{\Phi}$ can be *uniquely* decomposed as

$$\boxed{\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{m}_1, \mathbf{m}_2)\,\mathcal{A} = \boldsymbol{\Phi}(\mathbf{m}_1, \mathbf{m}_2)\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2.} \tag{38.24}$$

In particular, applying this result to (38.21), we find that

$$\boxed{\boldsymbol{\omega}_{13} \wedge \boldsymbol{\omega}_{23} = \big(\boldsymbol{\omega}_{13} \wedge \boldsymbol{\omega}_{23}\big)(\mathbf{m}_1, \mathbf{m}_2)\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 = \mathcal{K}_{\text{ext}}\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2.} \tag{38.25}$$

38.5.4 The Six Fundamental Form Equations of a Surface

Let us immediately state all six of these equations, and also state their names,[12] in one fell swoop. We will then prove them by calculation, and finally, gradually, extract their true geometrical meanings.

$$\boxed{\begin{array}{ll}
\begin{array}{l}\mathbf{d}\boldsymbol{\theta}^1 = \boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2 \\ \mathbf{d}\boldsymbol{\theta}^2 = \boldsymbol{\omega}_{21} \wedge \boldsymbol{\theta}^1\end{array} & \textbf{\textit{First Structural Equations}} \\[2ex]
\boldsymbol{\omega}_{31} \wedge \boldsymbol{\theta}^1 + \boldsymbol{\omega}_{32} \wedge \boldsymbol{\theta}^2 = 0 & \textbf{\textit{Symmetry Equation}} \\[2ex]
\mathbf{d}\boldsymbol{\omega}_{12} = \boldsymbol{\omega}_{13} \wedge \boldsymbol{\omega}_{32} & \textbf{\textit{Gauss Equation}} \\[2ex]
\begin{array}{l}\mathbf{d}\boldsymbol{\omega}_{13} = \boldsymbol{\omega}_{12} \wedge \boldsymbol{\omega}_{23} \\ \mathbf{d}\boldsymbol{\omega}_{23} = \boldsymbol{\omega}_{21} \wedge \boldsymbol{\omega}_{13}\end{array} & \textbf{\textit{Peterson–Mainardi–Codazzi Equations}}
\end{array}} \tag{38.26}$$

In our description of the surface $\mathcal{S}$, only $\mathbf{m}_3$ is uniquely, geometrically determined by the surface itself, as its normal. Meanwhile, we may arbitrarily twirl $(\mathbf{m}_1, \mathbf{m}_2)$ within the surface, about the axis $\mathbf{m}_3$. All these arbitrarily different frame fields yield *different* descriptions of the surface—*different* dual-basis 1-forms, and *different* connection 1-forms. Nevertheless, despite these freedoms, the six equations above embody everything that can be known or said about the surface.

It is therefore remarkable how quickly and easily all six equations follow from Cartan's Structural Equations—all we need do is set $\boldsymbol{\theta}^3 = 0$.

Because $\boldsymbol{\omega}_{11} = 0$, the First Structural Equation, (38.11), immediately yields

$$\mathbf{d}\boldsymbol{\theta}^1 = \sum_j \boldsymbol{\omega}_{1j} \wedge \boldsymbol{\theta}^j = \boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2 + \boldsymbol{\omega}_{13} \wedge \boldsymbol{\theta}^3 = \boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2.$$

The calculation of $\mathbf{d}\boldsymbol{\theta}^2$ is essentially identical.

---

[12]Here we largely adopt the names employed by O'Neill (2006, p. 267), but the reader should be aware that, while widely used, these names are not *universal*. In particular, what O'Neill calls the Codazzi Equations are also commonly known as the *Mainardi–Codazzi Equations*, but we have attempted to give credit where credit is due, calling them instead the ***Peterson–Mainardi–Codazzi Equations***. For although Mainardi discovered the equations in 1856, and, independently, Codazzi rediscovered them in 1860, both were anticipated in 1853 by a young Latvian student, Karl M. Peterson. Peterson's work (and priority) remained unknown until his dissertation (supervised by Minding!) was finally translated into Russian, and published in 1952. See Phillips (1979).

The Symmetry Equation follows in the same way:

$$0 = \mathbf{d}\boldsymbol{\theta}^3 = \sum_j \boldsymbol{\omega}_{3j} \wedge \boldsymbol{\theta}^j = \boldsymbol{\omega}_{31} \wedge \boldsymbol{\theta}^1 + \boldsymbol{\omega}_{32} \wedge \boldsymbol{\theta}^2.$$

Next, since $\boldsymbol{\omega}_{11} = 0 = \boldsymbol{\omega}_{22}$, the Gauss Equation follows immediately from the Second Structural Equation, (38.13):

$$\mathbf{d}\boldsymbol{\omega}_{12} = \sum_k \boldsymbol{\omega}_{1k} \wedge \boldsymbol{\omega}_{k2} = \boldsymbol{\omega}_{13} \wedge \boldsymbol{\omega}_{32}.$$

The final two Peterson–Mainardi–Codazzi Equations follow from the Second Structural Equation in exactly the same way.

38.6 Geometrical Meanings of the Symmetry Equation and the Peterson–Mainardi–Codazzi Equations

But what do these fundamental form equations actually *mean*?

Let us begin with the Symmetry Equation. If we apply this vanishing 2-form to the basis vectors, we obtain

$$\begin{aligned} 0 &= \left[\boldsymbol{\omega}_{31} \wedge \boldsymbol{\theta}^1 + \boldsymbol{\omega}_{32} \wedge \boldsymbol{\theta}^2\right](\mathbf{m}_1, \mathbf{m}_2) \\ &= \boldsymbol{\omega}_{31}(\mathbf{m}_1)\boldsymbol{\theta}^1(\mathbf{m}_2) - \boldsymbol{\omega}_{31}(\mathbf{m}_2)\boldsymbol{\theta}^1(\mathbf{m}_1) + \boldsymbol{\omega}_{32}(\mathbf{m}_1)\boldsymbol{\theta}^2(\mathbf{m}_2) - \boldsymbol{\omega}_{32}(\mathbf{m}_2)\boldsymbol{\theta}^2(\mathbf{m}_1) \\ &= \boldsymbol{\omega}_{32}(\mathbf{m}_1) - \boldsymbol{\omega}_{31}(\mathbf{m}_2). \end{aligned}$$

Thus, invoking the skew-symmetry of the connection Forms,

$$\boxed{\boldsymbol{\omega}_{23}(\mathbf{m}_1) = \boldsymbol{\omega}_{13}(\mathbf{m}_2).}$$

Therefore, referring back to (38.20), we deduce that

> *The Symmetry Equation is equivalent to the fact the Shape Operator is symmetric:* $S^T = S$.

The geometrical interpretation of this symmetry was previously derived in (15.13), on page 156: *The eigenvectors of* S *are orthogonal.*

If we allow ourselves the former knowledge that these eigenvectors are in fact the *principal directions* of maximum and minimum normal curvature, these curvatures being the associated eigenvalues, κ_1 and κ_2, then we have reproved a major part of Euler's discovery: these maximum and minimum curvatures occur in *orthogonal* directions.

Let us break new ground, and refine our adapted frame field so as to make $(\mathbf{m}_1, \mathbf{m}_2)$ align with these principal directions at each point of $\mathcal{S}$; this is called the ***principal frame field***. In this case, $S(\mathbf{m}_1) = \kappa_1 \mathbf{m}_1$ and $S(\mathbf{m}_2) = \kappa_2 \mathbf{m}_2$, so we deduce that (38.20) takes the form

$$[S] = \begin{bmatrix} \boldsymbol{\omega}_{13}(\mathbf{m}_1) & \boldsymbol{\omega}_{13}(\mathbf{m}_2) \\ \boldsymbol{\omega}_{23}(\mathbf{m}_1) & \boldsymbol{\omega}_{23}(\mathbf{m}_2) \end{bmatrix} = \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix}.$$

Thus, by virtue of (38.23),

$$\boxed{\boldsymbol{\omega}_{13} = \kappa_1\, \boldsymbol{\theta}^1 \qquad \text{and} \qquad \boldsymbol{\omega}_{23} = \kappa_2\, \boldsymbol{\theta}^2.} \tag{38.27}$$

We can now derive a geometrical interpretation of the Peterson–Mainardi–Codazzi Equations—this is genuinely new mathematics that we have not seen in the previous Acts. Inserting the equations (38.27) into the first Peterson–Mainardi–Codazzi Equation of (38.26), we find that

$$\mathbf{d}(\kappa_1\,\boldsymbol{\theta}^1) = \boldsymbol{\omega}_{12} \wedge \kappa_2\,\boldsymbol{\theta}^2.$$

Thus,

$$\mathbf{d}\kappa_1 \wedge \boldsymbol{\theta}^1 + \kappa_1\,\mathbf{d}\boldsymbol{\theta}^1 = \kappa_2\,\boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2.$$

Using the First Structural Equation from (38.26), namely, $\mathbf{d}\boldsymbol{\theta}^1 = \boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2$, we deduce that

$$\mathbf{d}\kappa_1 \wedge \boldsymbol{\theta}^1 + \kappa_1\,\boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2 = \kappa_2\,\boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2,$$

and therefore,

$$\mathbf{d}\kappa_1 \wedge \boldsymbol{\theta}^1 = (\kappa_2 - \kappa_1)\,\boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2.$$

Finally, applying these 2-forms (these are the left and right sides of this equation) to the basis vectors $(\mathbf{m}_1, \mathbf{m}_2)$, we obtain

$$0 - \mathbf{d}\kappa_1(\mathbf{m}_2) = (\kappa_2 - \kappa_1)\,\boldsymbol{\omega}_{12}(\mathbf{m}_1) - 0.$$

Hence, the ***Peterson–Mainardi–Codazzi Equation*** now takes the geometrically meaningful form,

$$\boxed{\boldsymbol{\nabla}_{\mathbf{m}_2}\kappa_1 = (\kappa_1 - \kappa_2)\,\boldsymbol{\omega}_{12}(\mathbf{m}_1).} \tag{38.28}$$

Repeating this calculation for the other Peterson–Mainardi–Codazzi Equation, we obtain

$$\boxed{\boldsymbol{\nabla}_{\mathbf{m}_1}\kappa_2 = (\kappa_1 - \kappa_2)\,\boldsymbol{\omega}_{12}(\mathbf{m}_2).} \tag{38.29}$$

In both these equations, $\boldsymbol{\omega}_{12}(\mathbf{v})$ tells us how fast the principal directions are rotating within $\mathcal{S}$ as we move along the surface in the direction $\mathbf{v}$. So, putting both equations into words, as we move off at right angles to a principal direction, the rate of change of that principal curvature is proportional to both the difference of the two principal curvatures, and to the rate at which the principal directions rotate.

38.7 Geometrical Form of the Gauss Equation

Of the six Form equations (38.26), the Gauss Equation is undoubtedly the most important, for it holds the key both to the intrinsic curvature, and to its relation to the extrinsic curvature. The connection to the latter is immediate from (38.25), for this allows us rewrite the Gauss Equation in a new, geometrically meaningful form.

$$\textit{Gauss Equation:} \qquad \boxed{\boxed{\mathbf{d}\boldsymbol{\omega}_{12} = -\mathcal{K}_{\text{ext}}\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2.}} \tag{38.30}$$

For example, consider the sphere of radius R, analyzed in Section 38.5.2:

$$\mathbf{d}\boldsymbol{\omega}_{12} = \mathbf{d}(c_\phi\,\mathbf{d}\vartheta) = -s_\phi\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta = -\frac{1}{R^2}[R\,\mathbf{d}\phi] \wedge [R\,s_\phi\,\mathbf{d}\vartheta] = -\frac{1}{R^2}\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2.$$

From this we immediately deduce (correctly!) that $\mathcal{K}_{\text{ext}} = +(1/R^2)$.

38.8 Proof of the Metric Curvature Formula and the *Theorema Egregium*

38.8.1 Lemma: Uniqueness of ω_{12}

The connection form $\boldsymbol{\omega}_{12} = -\boldsymbol{\omega}_{21}$ is the only *1-form that satisfies Cartan's First Structural Equations of (38.26):*

$$\mathbf{d}\boldsymbol{\theta}^1 = \boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^2 \qquad \textit{and} \qquad \mathbf{d}\boldsymbol{\theta}^2 = -\boldsymbol{\omega}_{12} \wedge \boldsymbol{\theta}^1. \tag{38.31}$$

To prove this, we apply these 2-forms to the pair of basis vectors, $(\mathbf{m}_1, \mathbf{m}_2)$, and find that [exercise]

$$\boldsymbol{\omega}_{12}(\mathbf{m}_1) = \mathbf{d}\boldsymbol{\theta}^1(\mathbf{m}_1, \mathbf{m}_2) \qquad \text{and} \qquad \boldsymbol{\omega}_{12}(\mathbf{m}_2) = \mathbf{d}\boldsymbol{\theta}^2(\mathbf{m}_1, \mathbf{m}_2).$$

It follows from (38.23) that $\boldsymbol{\omega}_{12}$ is uniquely and *explicitly* given by

$$\boxed{\boldsymbol{\omega}_{12} = \left[\mathbf{d}\boldsymbol{\theta}^1(\mathbf{m}_1, \mathbf{m}_2)\right] \boldsymbol{\theta}^1 + \left[\mathbf{d}\boldsymbol{\theta}^2(\mathbf{m}_1, \mathbf{m}_2)\right] \boldsymbol{\theta}^2,} \tag{38.32}$$

thereby completing the proof of (38.31).

38.8.2 Proof of the Metric Curvature Formula

Mathematically speaking, we are now living in the twenty-third century—the future from which the "*Star Trek* phaser" formula (4.10) was delivered to us, back through time, long ago on page 38.

Cartan's Forms now make possible a computational proof of this Metric Curvature Formula that is startling in its brevity and simplicity, on a par with our geometrical demonstration in [27.4].

From

$$ds^2 = (A\, du)^2 + (B\, dv)^2,$$

we immediately deduce that the adapted 1-form basis for the surface is

$$\boxed{\boldsymbol{\theta}^1 = A\, \mathbf{d}u \qquad \text{and} \qquad \boldsymbol{\theta}^2 = B\, \mathbf{d}v,} \tag{38.33}$$

and that the area 2-form of the surface is

$$\boxed{\mathcal{A} = \boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 = AB\, \mathbf{d}u \wedge \mathbf{d}v.}$$

We now calculate the exterior derivatives of the dual-basis 1-forms:

$$\begin{aligned}
\mathbf{d}\boldsymbol{\theta}^1 &= \mathbf{d}A \wedge \mathbf{d}u = \partial_v A\, \mathbf{d}v \wedge \mathbf{d}u = -\frac{\partial_v A}{B}\, \mathbf{d}u \wedge \mathbf{d}\boldsymbol{\theta}^2 \\
\mathbf{d}\boldsymbol{\theta}^2 &= \mathbf{d}B \wedge \mathbf{d}v = \partial_u B\, \mathbf{d}u \wedge \mathbf{d}v = -\frac{\partial_u B}{A}\, \mathbf{d}v \wedge \mathbf{d}\boldsymbol{\theta}^1
\end{aligned}$$

Comparing these equations with Cartan's First Structural Equations, (38.31), we see immediately that one possible solution for $\boldsymbol{\omega}_{12}$ is

$$\boxed{\boldsymbol{\omega}_{12} = -\frac{\partial_v A}{B}\,\mathbf{d}u + \frac{\partial_u B}{A}\,\mathbf{d}v.} \tag{38.34}$$

But lemma (38.31) tells us that if this is *a* solution, then it is the *only* solution—it is *the* solution!

Taking the exterior derivative,

$$\begin{aligned}
\mathbf{d}\boldsymbol{\omega}_{12} &= -\mathbf{d}\left[\frac{\partial_v A}{B}\right]\wedge\mathbf{d}u + \mathbf{d}\left[\frac{\partial_u B}{A}\right]\wedge\mathbf{d}v \\
&= -\partial_v\left[\frac{\partial_v A}{B}\right]\mathbf{d}v\wedge\mathbf{d}u + \partial_u\left[\frac{\partial_u B}{A}\right]\mathbf{d}u\wedge\mathbf{d}v \\
&= \frac{1}{AB}\left(\partial_v\left[\frac{\partial_v A}{B}\right] + \partial_u\left[\frac{\partial_u B}{A}\right]\right)\boldsymbol{\theta}^1\wedge\boldsymbol{\theta}^2.
\end{aligned}$$

Comparing this with the Gauss Equation, (38.30), we see that, in a mere fraction of a page, our work is *done!* We have (re)proved the very remarkable formula, (4.10):

$$\boxed{\mathcal{K}_{\text{ext}} = -\frac{1}{AB}\left(\partial_v\left[\frac{\partial_v A}{B}\right] + \partial_u\left[\frac{\partial_u B}{A}\right]\right).} \tag{38.35}$$

The extrinsic curvature $\mathcal{K}_{\text{ext}} = \kappa_1\kappa_2$ on the left-hand side of this equation arose from (38.21)—it measures the spreading of the normal vectors to the surface in space. But the expression on the right-hand side of this equation depends *only* on the intrinsic metric geometry of the surface. Therefore, this formula is a very explicit and very pure expression of (and proof of!) Gauss's *Theorema Egregium*.

There is a *reason* that the simplicity of this computational proof rivals that of our geometrical proof in [27.4]—it is, fundamentally, the *same* proof!

To begin to understand this, first observe that the ***intrinsic*** connection form $\boldsymbol{\omega}_{12}$ tells us about the rotation of $(\mathbf{m}_1, \mathbf{m}_2)$ within the surface, *relative to a parallel-transported vector.* Since this is the *opposite* of the rotation $\delta\mathcal{R}$ of a parallel-transported vector relative to the frame, this explains *why* (38.34) is simply the negative of our original, geometrically derived formula, (27.5), page 267.

The final step of the geometrical proof was to find the holonomy of a vector that was parallel-transported around a small closed loop. But, as we learned in (37.8), page 408, this integral of $\boldsymbol{\omega}_{12}$ around a small loop is *precisely* what is measured by $\mathbf{d}\boldsymbol{\omega}_{12}$.

38.9 A New Curvature Formula

We now derive a new curvature formula that will enable us to reach new results. With the *Theorema Egregium* (re)established, we may revert to denoting *the* curvature of the surface as $\mathcal{K}$, since the extrinsic and intrinsic curvatures are in fact one and the same.

We will now show that

$$\boxed{\mathcal{K} = \boldsymbol{\nabla}_{\mathbf{m}_2}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_1)\right] - \boldsymbol{\nabla}_{\mathbf{m}_1}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_2)\right] - \left[\boldsymbol{\omega}_{12}(\mathbf{m}_1)\right]^2 - \left[\boldsymbol{\omega}_{12}(\mathbf{m}_2)\right]^2.} \tag{38.36}$$

To avoid clutter, let us write the components of $\boldsymbol{\omega}_{12}$ as f_1 and f_2, so that (38.32) takes the form

$$\boldsymbol{\omega}_{12} = f_1\,\boldsymbol{\theta}^1 + f_2\,\boldsymbol{\theta}^2,$$

where

$$f_1 = \boldsymbol{\omega}_{12}(\mathbf{m}_1) = \mathbf{d}\boldsymbol{\theta}^1(\mathbf{m}_1, \mathbf{m}_2) \qquad \text{and} \qquad f_2 = \boldsymbol{\omega}_{12}(\mathbf{m}_2) = \mathbf{d}\boldsymbol{\theta}^2(\mathbf{m}_1, \mathbf{m}_2).$$

If we apply the 2-form $\mathbf{d}\boldsymbol{\omega}_{12}$ to the pair of basis vectors, $(\mathbf{m}_1, \mathbf{m}_2)$, then the Gauss Equation (38.30) yields,

$$\begin{aligned}
\mathcal{K} &= -\mathbf{d}\boldsymbol{\omega}_{12}(\mathbf{m}_1, \mathbf{m}_2) \\
&= -\left[\mathbf{d}f_1 \wedge \boldsymbol{\theta}^1 + f_1\,\mathbf{d}\boldsymbol{\theta}^1 + \mathbf{d}f_2 \wedge \boldsymbol{\theta}^2 + f_2\,\mathbf{d}\boldsymbol{\theta}^2\right](\mathbf{m}_1, \mathbf{m}_2) \\
&= \mathbf{d}f_1(\mathbf{m}_2) - \mathbf{d}f_2(\mathbf{m}_1) - f_1^2 - f_2^2 \\
&= \boldsymbol{\nabla}_{\mathbf{m}_2} f_1 - \boldsymbol{\nabla}_{\mathbf{m}_1} f_2 - f_1^2 - f_2^2.
\end{aligned}$$

which is, indeed, the desired formula, (38.36).

38.10 Hilbert's Lemma

Hilbert showed that if a point p on a surface has the following three properties, then the curvature of the surface *cannot be positive* at p:

$$\boxed{\left\{\begin{array}{l} \kappa_1 \text{ has local maximum at } p \\ \kappa_2 \text{ has local minimum at } p \\ \kappa_1(p) > \kappa_2(p) \end{array}\right\} \quad \Longrightarrow \quad \mathcal{K}(p) \leqslant 0.} \tag{38.37}$$

For example, these conditions are realized at each point on the inner equator of a torus, and the curvature at these points is indeed negative, in accord with Hilbert's Lemma.

Since p is a critical point for κ_1 and for κ_2, at p we have

$$\boldsymbol{\nabla}_{\mathbf{m}_2} \kappa_1 = 0 = \boldsymbol{\nabla}_{\mathbf{m}_1} \kappa_2. \tag{38.38}$$

Also, since κ_1 achieves a local *maximum* at p, and κ_2 achieves a local *minimum* at p,

$$\boldsymbol{\nabla}_{\mathbf{m}_2} \boldsymbol{\nabla}_{\mathbf{m}_2} \kappa_1 \leqslant 0 \qquad \text{and} \qquad \boldsymbol{\nabla}_{\mathbf{m}_1} \boldsymbol{\nabla}_{\mathbf{m}_1} \kappa_2 \geqslant 0. \tag{38.39}$$

Since $(\kappa_1 - \kappa_2) > 0$, inserting (38.38) into the Peterson–Mainardi–Codazzi Equations (38.28) and (38.29), we find that, at p,

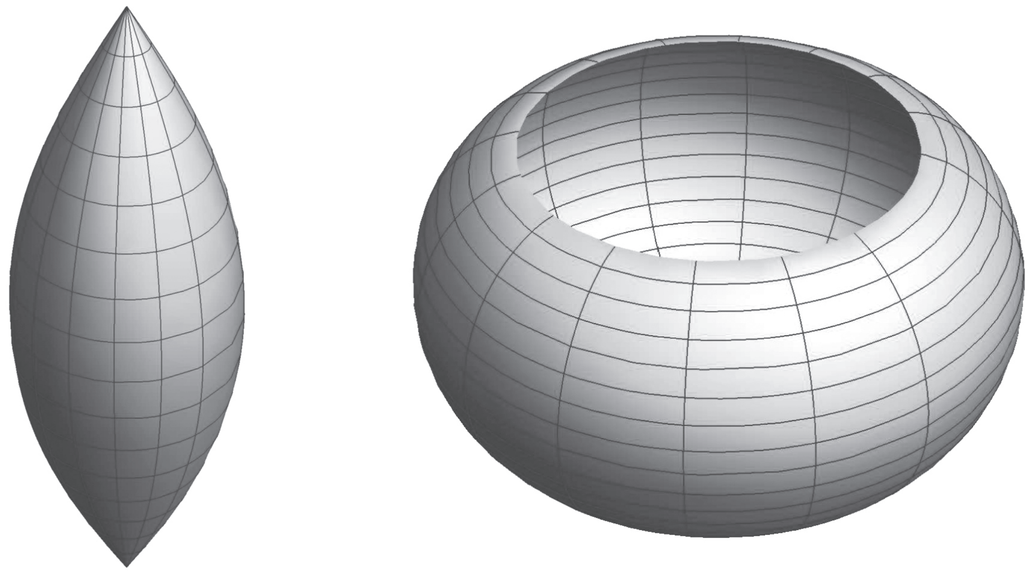
$$\boldsymbol{\omega}_{12}(\mathbf{m}_1) = 0 = \boldsymbol{\omega}_{12}(\mathbf{m}_2). \tag{38.40}$$

Thus, the curvature formula (38.36) reduces to

$$\mathcal{K} = \boldsymbol{\nabla}_{\mathbf{m}_2}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_1)\right] - \boldsymbol{\nabla}_{\mathbf{m}_1}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_2)\right]. \tag{38.41}$$

Differentiating (38.28) and inserting (38.40),

$$\begin{aligned}
\boldsymbol{\nabla}_{\mathbf{m}_2} \boldsymbol{\nabla}_{\mathbf{m}_2} \kappa_1 &= \left[\boldsymbol{\nabla}_{\mathbf{m}_2}(\kappa_1 - \kappa_2)\right] \boldsymbol{\omega}_{12}(\mathbf{m}_1) + (\kappa_1 - \kappa_2)\,\boldsymbol{\nabla}_{\mathbf{m}_2}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_1)\right] \\
&= (\kappa_1 - \kappa_2)\,\boldsymbol{\nabla}_{\mathbf{m}_2}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_1)\right].
\end{aligned}$$



[38.7] *These two kinds of surfaces have constant positive curvature, and the same* intrinsic *geometry as a sphere, but they are clearly* not *spherical!*

Finally, since $(\kappa_1 - \kappa_2) > 0$, (38.39) implies

$$\nabla_{\mathbf{m}_2}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_1)\right] \leqslant 0.$$

Applying the same logic to the other Peterson–Mainardi–Codazzi Equation, (38.29), we find

$$\nabla_{\mathbf{m}_1}\left[\boldsymbol{\omega}_{12}(\mathbf{m}_2)\right] \geqslant 0.$$

Inserting the previous two inequalities into the curvature formula (38.41), the lemma is proved.

38.11 Liebmann's Rigid Sphere Theorem

Recall that Minding's Theorem of 1839 (proved in Ex. 7, p. 336) implies that if a surface $\mathcal{S}$ has *constant* positive curvature $\mathcal{K}$, then its *intrinsic* metric geometry must be identical to that of a sphere of radius $1/\sqrt{\mathcal{K}}$, at least locally.

It is tempting to go further, speculating that $\mathcal{S}$ must actually *be* part of a sphere, manifesting the completely symmetrical *extrinsic* geometry of a sphere, in which every point is an *umbilic* point: $\kappa_1 = \kappa_2$. In fact, in the following, we shall take it as given that this condition does indeed *characterize* the sphere.

However, such a guess would be incorrect: [38.7] depicts two kinds of surfaces that are very definitely *not* spherical, and yet which *both have constant positive curvature, and the same intrinsic geometry as a sphere!* (For the equations of these surfaces, see Ex. 7, p. 336.)

We can also see this experimentally. Take a ping pong ball and cut it in half; now gently squeeze together antipodal points on the equator. The surface will flex without stretching, creating a new, nonspherical surface of constant positive curvature, and the same intrinsic geometry as the original sphere.

But what if we try to flex a *complete* ping pong ball? It would appear that we *cannot!* Minding himself had first speculated about this issue in 1839, but it took another 60 years before Heinrich

Liebmann was finally able to *prove* (in 1899) that a *closed* (bounded) surface of constant curvature embedded in $\mathbb{R}^3$ can *only* be a sphere.

Thus, if the geometry is intrinsically uniform (in the sense of constant $\mathcal{K}$), then the extrinsic geometry can also *only* be the extrinsically uniform geometry of a sphere, and therefore the sphere *cannot be deformed* into any other extrinsic shapes—a complete sphere is *rigid*.

The best known proof of Liebmann's Theorem was discovered, not long afterwards, by David Hilbert,[13] and the crucial ingredient of his proof was (38.37), hence its name—"Hilbert's Lemma."

We begin by observing that since $\mathcal{S}$ is bounded, its constant curvature must be *positive* curvature. By definition, being bounded means that $\mathcal{S}$ can be enclosed within a sufficiently large sphere. Now shrink the sphere till it first touches $\mathcal{S}$, at c, say. Then the little piece of $\mathcal{S}$ surrounding c all lies on one side of the common tangent plane to $\mathcal{S}$ and the sphere at c. Thus c must be an *elliptic* point of *positive* curvature, and since the curvature is assumed to be constant, it is positive everywhere on $\mathcal{S}$.

Let us assume that on a closed (bounded) surface the function κ_1 must achieve a maximum. (A proof can be found in O'Neill (2006, pp. 185–186).) Since $\mathcal{K} = \kappa_1 \kappa_2$ is assumed to be constant, κ_2 must achieve a minimum at a point at which κ_1 achieves a maximum. But this means that it is *impossible* that $\kappa_1 > \kappa_2$. For, if that were the case, Hilbert's Lemma would demand that $\mathcal{K} \leqslant 0$, contradicting the fact that the surface has (constant) *positive* curvature.

Thus $\kappa_1(p) = \kappa_2(p)$. And since $\kappa_1(p)$ is the *maximum* of κ_1 on $\mathcal{S}$, and $\kappa_2(p)$ is the *minimum* of κ_2 on $\mathcal{S}$, this means that $\kappa_1 = \kappa_2$ at *every* point of $\mathcal{S}$, and therefore $\mathcal{S}$ is a sphere.

38.12 The Curvature 2-Forms of an n-Manifold

38.12.1 Introduction and Summary

Recall the scene as the curtain rose on Act I of our drama. The *flatness* of the Euclidean plane $\mathbb{R}^2$ was *characterized* by the following property of of an arbitrary triangle Δ constructed within it:

$$\boxed{\text{Angle sum of } \Delta = \pi.} \tag{38.42}$$

If this equation is *not* satisfied, it means that Δ lives within a surface that is *curved*.

The situation is analogous to the juxtaposition of Newton's First and Second Laws of Motion: if we observe an object that is *not* moving at constant speed in a straight line, there must be a force acting upon it. And the magnitude of the force manifests itself as the acceleration, that is, as the *difference* between the actual motion and force-free, constant-speed, straight-line motion.

Likewise, the amount of curvature within the surface is quantified by the *difference* between the two sides of the equation above. This is the *angular excess*, $\mathcal{E}$:

$$\boxed{\mathcal{E}(\Delta) \equiv (\text{Angle sum of } \Delta) - \pi.} \tag{38.43}$$

When applied to a small (ultimately vanishing) triangle, this allows us to *quantify* the curvature $\mathcal{K}$ *at* the triangle Δ, of area $\mathcal{A}(\Delta)$, by means of this fundamental formula:

$$\boxed{\boxed{\mathcal{E}(\Delta) \asymp \mathcal{K}\,\mathcal{A}(\Delta).}} \tag{38.44}$$

[13] Liebmann completed his doctoral work at Göttingen, but *not* as Hilbert's student. Yet one senses, perhaps, some jealousy on the part of Hilbert, for he writes that Liebmann proved the theorem "at my urging" (Hilbert 1902, p. 197). In any event, it is Hilbert's *proof* (given here) of Liebmann's theorem that lives on!

Now let us move up a dimension, to *3-dimensional* Euclidean space, $\mathbb{R}^3$. As we will prove, the analogue of (38.42) turns out to be none other than *Cartan's Second Structural Equation*, (38.12)—*this equation characterizes the flatness of the space:*

$$\boxed{\mathbf{d}[\boldsymbol{\omega}] = [\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}].}$$

In a curved n-manifold, this equation does *not* hold, and the *difference* between the two sides of the equation once again quantifies the curvature of the space. The analogue of (38.43) is now *a matrix* $[\boldsymbol{\Omega}]$ *of 2-forms*, called the ***curvature matrix***:

$$\boxed{[\boldsymbol{\Omega}] \equiv \mathbf{d}[\boldsymbol{\omega}] - [\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}].} \tag{38.45}$$

The individual entries within this matrix are called[14] the

Curvature 2-forms:
$$\boxed{\boldsymbol{\Omega}_{ij} = \mathbf{d}\boldsymbol{\omega}_{ij} - \textstyle\sum_k \boldsymbol{\omega}_{ik} \wedge \boldsymbol{\omega}_{kj}.} \tag{38.46}$$

It is *this* (fully general) equation that is usually referred to (in more advanced works) as ***Cartan's Second Structural Equation***; in $\mathbb{R}^3$ it reduces to our original version, (38.13), for in that case $\boldsymbol{\Omega}_{ij} = 0$.

IMPORTANT NOTATIONAL NOTE: Two different notations exist in the literature for both the connection 1-forms and the curvature 2-forms. In the more common notation, which we too shall adopt shortly, both these objects have one raised index and one lowered index. As we shall explain, the relation between the notations is, $\boldsymbol{\omega}^i{}_j \equiv \boldsymbol{\omega}_{ji}$ and $\boldsymbol{\Omega}^i{}_j \equiv \boldsymbol{\Omega}_{ji}$. This reversal of the order of the indices has the effect that *in both (38.45) and (38.46) the "$-$" becomes a "$+$".*

We will prove that *these curvature 2-forms contain exactly the same information as the Riemann tensor,* but packaged in a more compact and elegant form [*sic*].

For your convenience, we repeat here the definitions of the Riemann tensor $\mathbf{R}$ and of the Riemann curvature operator $\mathcal{R}$, both of which are stated in (29.8), page 290:

$$\boxed{\mathbf{R}(\mathbf{u}, \mathbf{v};\, \mathbf{w}) = \mathcal{R}(\mathbf{u}, \mathbf{v})\, \mathbf{w} = \left\{ [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u}, \mathbf{v}]} \right\} \mathbf{w}.} \tag{38.47}$$

Recall that in Chapter 29 we never actually worked out the generalization of the "*Star Trek* phaser" formula (27.1) for an n-manifold. That is, we never found a formula for the Riemann tensor explicitly expressed directly in terms of the metric coefficients. There *is* such a formula—it's ugly and it's complicated, involving things called ***Christoffel symbols***, and it can be found in *every* standard textbook on Riemannian Geometry—but we chose not to derive that formula for a good reason: *we can do better!*

[14] Multiple notations exist for the curvature 2-forms. Our $\boldsymbol{\Omega}_{ij}$ is probably the most common, and it was the notation originally used by Cartan (1927) himself, but Flanders (1989) and Chern et al. (1999) instead use $\boldsymbol{\Theta}_{ij}$, while Misner, Thorne, and Wheeler (1973) use $\mathcal{R}_{ij}$.

The curvature 2-forms allow us to calculate the Riemann tensor much more efficiently and cleanly than any frontal assault could ever hope to achieve. Formula (38.46) allows us to calculate $\boldsymbol{\Omega}_{ij}$, and then we simply *read off* the components of the Riemann tensor, for *we shall prove that they are simply the components of the curvature 2-forms!* That is,

$$\boxed{\boxed{\boldsymbol{\Omega}_{ij} = R_{ijkl}\,\boldsymbol{\theta}^k \wedge \boldsymbol{\theta}^l.}} \tag{38.48}$$

NOTE: Here, and in the following, we have reverted to the Einstein summation convention: summation over the matching k-indices and l-indices is *understood*. However, we assume that the sum is over *distinct* basis 2-forms, or else we would require a factor of $(1/2)$.

38.12.2 The Generalized Exterior Derivative

Calculating the Riemann tensor via the curvature 2-forms, as defined in (38.46), only requires routine exterior differentiation of the connection 1-forms. However, in order to *reach* the formula (38.48), we shall need to generalize the action of **d** beyond Forms.

First, though, let us introduce an alternative notation for the connection 1-forms in which one index is raised and one is lowered: $\boldsymbol{\omega}^i{}_j$. This notation will not only allow us to reinstate the use of the Einstein summation convention, but, more importantly, it will also enable you to make an easy transition to studying such works as Misner, Thorne, and Wheeler (1973), Frankel (2012), Dray (2015), and Tu (2017), *all* of whom employ this notation.

These $\boldsymbol{\omega}^i{}_j$ are defined so that (38.4) now takes the form

$$\boxed{\boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_j = \textstyle\sum_i \boldsymbol{\omega}_{ji}(\mathbf{v})\,\mathbf{m}_i \equiv \sum_i \boldsymbol{\omega}^i{}_j(\mathbf{v})\,\mathbf{m}_i = \boldsymbol{\omega}^i{}_j(\mathbf{v})\,\mathbf{m}_i,} \tag{38.49}$$

in which the final equality signals the return to the Einstein summation convention—the sum over the matching pair or i-indices (one up and one down) is *understood*.

Note the *reversal* of the order of the indices![15]

$$\boxed{\boldsymbol{\omega}^i{}_j \equiv \boldsymbol{\omega}_{ji}.} \tag{38.50}$$

Now let us try to *generalize* the exterior derivative **d**, defining its action upon the moving frame field *vectors*, $\mathbf{m}_i$. By analogy with

$$\mathbf{d}f(\mathbf{v}) = \boldsymbol{\nabla}_{\mathbf{v}} f,$$

let us define,

$$\mathbf{dm}_j(\mathbf{v}) \equiv \boldsymbol{\nabla}_{\mathbf{v}}\mathbf{m}_j = \boldsymbol{\omega}^i{}_j(\mathbf{v})\,\mathbf{m}_i.$$

Abstracting away the arbitrary vector $\mathbf{v}$,

$$\boxed{\mathbf{dm}_j = \boldsymbol{\omega}^i{}_j\,\mathbf{m}_i.}$$

[15] I do not understand the origin of this choice, and I do not welcome it, but it *is* the standard notation, so we shall abide by it!

This now allows us to apply $\mathbf{d}$ to a general vector field, $\mathbf{w} = w^j\,\mathbf{m}_j$. Renaming indices as we go, we find that

$$\mathbf{d}\mathbf{w} = \mathbf{d}\big[w^j\,\mathbf{m}_j\big] = \big[\mathbf{d}w^j\big]\mathbf{m}_j + w^j\,\mathbf{d}\mathbf{m}_j = \mathbf{m}_i\,\big(\mathbf{d}w^i + \boldsymbol{\omega}^i{}_j\,w^j\big).$$

Now let us differentiate *again*.

Wait!—isn't $\mathbf{d}^2 = 0$?! No, $\mathbf{d}^2$ vanishes when it is applied to *Forms*, but not now! Renaming indices as we go, we find,

$$\begin{aligned}
\mathbf{d}^2\mathbf{w} &= \mathbf{d}\mathbf{m}_k \wedge \big(\mathbf{d}w^k + \boldsymbol{\omega}^k{}_j\,w^j\big) + \mathbf{m}_i\,\big(\mathbf{d}^2 w^i + w^j\,\mathbf{d}\boldsymbol{\omega}^i{}_j - \boldsymbol{\omega}^i{}_j \wedge \mathbf{d}w^j\big) \\
&= \boldsymbol{\omega}^i{}_k\,\mathbf{m}_i \wedge \big(\mathbf{d}w^k + \boldsymbol{\omega}^k{}_i\,w^i\big) + \mathbf{m}_i\,\big(w^j\,\mathbf{d}\boldsymbol{\omega}^i{}_j - \boldsymbol{\omega}^i{}_k \wedge \mathbf{d}w^k\big) \\
&= \mathbf{m}_i\,\big(\mathbf{d}\boldsymbol{\omega}^i{}_j + \boldsymbol{\omega}^i{}_k \wedge \boldsymbol{\omega}^k{}_j\big)\,w^j.
\end{aligned}$$

Thus,

$$\boxed{\mathbf{d}^2\mathbf{w} = \mathbf{m}_i\,\boldsymbol{\Omega}^i{}_j\,w^j,} \tag{38.51}$$

where $\boldsymbol{\Omega}^i{}_j$ are, again, the

Curvature 2-Forms:
$$\boxed{\boldsymbol{\Omega}^i{}_j = \mathbf{d}\boldsymbol{\omega}^i{}_j + \boldsymbol{\omega}^i{}_k \wedge \boldsymbol{\omega}^k{}_j.} \tag{38.52}$$

IMPORTANT NOTATIONAL NOTE: Observe that there is an unfortunate sign discrepancy between this formula for the 2-forms and our original definition, (38.46). This discrepancy arises from (38.50):

$$\boxed{\boldsymbol{\Omega}_{ij} = \mathbf{d}\boldsymbol{\omega}_{ij} - \textstyle\sum_k \boldsymbol{\omega}_{ik} \wedge \boldsymbol{\omega}_{kj} = \mathbf{d}\boldsymbol{\omega}^j{}_i - \sum_k \boldsymbol{\omega}^k{}_i \wedge \boldsymbol{\omega}^j{}_k = \mathbf{d}\boldsymbol{\omega}^j{}_i + \sum_k \boldsymbol{\omega}^j{}_k \wedge \boldsymbol{\omega}^k{}_i = \boldsymbol{\Omega}^j{}_i.}$$

Likewise, note that the sign also changes in the formula (38.45) for the *curvature matrix*:

$$\boxed{[\boldsymbol{\Omega}^i{}_j] = \mathbf{d}[\boldsymbol{\omega}] + [\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}].} \tag{38.53}$$

Likewise, the new notation flips the sign in ***Cartan's First Structural Equation***, so that (38.11) now becomes

$$\boxed{\mathbf{d}\boldsymbol{\theta}^i = -\boldsymbol{\omega}^i{}_j \wedge \boldsymbol{\theta}^j.} \tag{38.54}$$

Henceforth, we shall *only* employ $\boldsymbol{\Omega}^i{}_j$, given by (38.52), for this is the most common expression of the curvature 2-forms found in the literature.

38.12.3 *Extracting the Riemann Tensor from the Curvature 2-Forms*

Recall our very first formula (on p. 393) for the exterior derivative of a 1-form, (36.2):

$$\mathbf{d}\varphi(\mathbf{u},\mathbf{v}) = \nabla_{\mathbf{u}}\varphi(\mathbf{v}) - \nabla_{\mathbf{v}}\varphi(\mathbf{u}) - \varphi([\mathbf{u},\mathbf{v}]).$$

We now generalize this by substituting $\mathbf{dw}$ in place of φ, recalling that

$$\mathbf{dw}(\mathbf{v}) = \nabla_{\mathbf{v}}\mathbf{w}.$$

Thus,

$$\begin{aligned} \mathbf{d}^2\mathbf{w}(\mathbf{u},\mathbf{v}) &= \nabla_{\mathbf{u}}\mathbf{dw}(\mathbf{v}) - \nabla_{\mathbf{v}}\mathbf{dw}(\mathbf{u}) - \mathbf{dw}([\mathbf{u},\mathbf{v}]) \\ &= \nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w} - \nabla_{\mathbf{v}}\nabla_{\mathbf{u}}\mathbf{w} - \nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w} \\ &= \mathcal{R}(\mathbf{u},\mathbf{v})\,\mathbf{w}, \end{aligned}$$

by virtue of (38.47).

Combining this with (38.52), we have obtained a remarkable formula for the Riemann tensor in terms of the curvature 2-forms:

$$\boxed{R^i{}_{jkl}\,\boldsymbol{\theta}^k \wedge \boldsymbol{\theta}^l = \boldsymbol{\Omega}^i{}_j = \mathbf{d}\boldsymbol{\omega}^i{}_j + \boldsymbol{\omega}^i{}_m \wedge \boldsymbol{\omega}^m{}_j.} \tag{38.55}$$

We obtained this formula by climbing the ladder of the *generalized* exterior derivative, but, having *arrived* in the promised land, we may now toss the ladder aside! For this final formula (38.55) uses nothing more than the *ordinary* exterior derivative applied to the connection 1-forms.

Thus we are now in the possession of an elegant and extraordinarily powerful method of computing the Riemann tensor using curvature 2-forms. We shall *end* our drama by applying this method to a specific example of enormous physical importance.

38.12.4 *The Bianchi Identities Revisited*

Both the First (Algebraic) Bianchi Identity (29.15) and the Second (Differential) Bianchi Identity (29.17) can be very elegantly derived, and compactly expressed, using our shiny new curvature 2-forms.

Since the basis 1-forms $\boldsymbol{\theta}^i$ are ordinary 1-forms, $\mathbf{d}^2$ annihilates them, yielding a vanishing 3-form. Thus, differentiating Cartan's First Structural Equation, (38.54),

$$\begin{aligned} 0 &= -\mathbf{dd}\boldsymbol{\theta}^i \\ &= \mathbf{d}(\boldsymbol{\omega}^i{}_j \wedge \boldsymbol{\theta}^j) \\ &= \mathbf{d}\boldsymbol{\omega}^i{}_j \wedge \boldsymbol{\theta}^j - \boldsymbol{\omega}^i{}_k \wedge \mathbf{d}\boldsymbol{\theta}^k \\ &= \mathbf{d}\boldsymbol{\omega}^i{}_j \wedge \boldsymbol{\theta}^j + \boldsymbol{\omega}^i{}_k \wedge \boldsymbol{\omega}^k{}_j \wedge \boldsymbol{\theta}^j \\ &= \left(\mathbf{d}\boldsymbol{\omega}^i{}_j + \boldsymbol{\omega}^i{}_k \wedge \boldsymbol{\omega}^k{}_j\right) \wedge \boldsymbol{\theta}^j \\ &= \boldsymbol{\Omega}^i{}_j \wedge \boldsymbol{\theta}^j. \end{aligned}$$

Thus, employing the more compact matrix notation, we have obtained the

$$\boxed{\textbf{\textit{(First) Algebraic Bianchi Identity:}} \qquad [\boldsymbol{\Omega}] \wedge [\boldsymbol{\theta}] = 0.} \tag{38.56}$$

In its present form, this is certainly *unrecognizable* as the original First (Algebraic) Bianchi Identity (29.15)! To see that these two identities are, in fact, one and the same, we invoke (38.55):

$$0 = \boldsymbol{\Omega}^{i}{}_{j} \wedge \boldsymbol{\theta}^{j} = R^{i}{}_{jkl}\, \boldsymbol{\theta}^{j} \wedge \boldsymbol{\theta}^{k} \wedge \boldsymbol{\theta}^{l},$$

and this does indeed imply [exercise] the equivalent [exercise] of (29.15):

$$R^{i}{}_{jkl} + R^{i}{}_{klj} + R^{i}{}_{ljk} = 0.$$

The Second (Differential) Bianchi Identity (29.17) can be obtained similarly, by applying $\mathbf{d}^2$ to the connection 1-forms. Thus, differentiating Cartan's Second Structural Equation, (38.52),

$$\begin{aligned}
0 &= \mathbf{d}\mathbf{d}\boldsymbol{\omega}^{i}{}_{j} \\
&= \mathbf{d}(\boldsymbol{\Omega}^{i}{}_{j} - \boldsymbol{\omega}^{i}{}_{k} \wedge \boldsymbol{\omega}^{k}{}_{j}) \\
&= \mathbf{d}\boldsymbol{\Omega}^{i}{}_{j} - \mathbf{d}\boldsymbol{\omega}^{i}{}_{k} \wedge \boldsymbol{\omega}^{k}{}_{j} + \boldsymbol{\omega}^{i}{}_{k} \wedge \mathbf{d}\boldsymbol{\omega}^{k}{}_{j} \\
&= \mathbf{d}\boldsymbol{\Omega}^{i}{}_{j} - (\boldsymbol{\Omega}^{i}{}_{k} - \boldsymbol{\omega}^{i}{}_{m} \wedge \boldsymbol{\omega}^{m}{}_{k}) \wedge \boldsymbol{\omega}^{k}{}_{j} + \boldsymbol{\omega}^{i}{}_{k} \wedge (\boldsymbol{\Omega}^{k}{}_{j} - \boldsymbol{\omega}^{k}{}_{m} \wedge \boldsymbol{\omega}^{m}{}_{j}) \\
&= \mathbf{d}\boldsymbol{\Omega}^{i}{}_{j} - \boldsymbol{\Omega}^{i}{}_{k} \wedge \boldsymbol{\omega}^{k}{}_{j} + \boldsymbol{\omega}^{i}{}_{k} \wedge \boldsymbol{\Omega}^{k}{}_{j}.
\end{aligned}$$

Thus, employing the more compact matrix notation, we have obtained the

$$\boxed{\textbf{\textit{(Second) Differential Bianchi Identity:}} \qquad \mathbf{d}[\boldsymbol{\Omega}] = [\boldsymbol{\Omega}] \wedge [\boldsymbol{\omega}] - [\boldsymbol{\omega}] \wedge [\boldsymbol{\Omega}].} \tag{38.57}$$

Again, (38.55) provides the explicit link to the Riemann tensor that enables one to see that this is indeed equivalent [exercise] to our original version of the Differential Bianchi Identity, (29.17).

38.13 The Curvature of the Schwarzschild Black Hole

This section lowers the curtain on Act V, bringing our long drama to its end. It does so by applying the foregoing ideas to a mathematical and physical object of cosmic importance: the spacetime curvature of a black hole!

Specifically, we shall endeavour to verify that the Schwarzschild black hole geometry, given by[16] (38.58), which we repeat here,

[16]As we explained in the context of Maxwell's Equations, we have now reversed the sign of the metric coefficients, in order to bring our work into line with the majority of General Relativity texts that we recommend in the *Further Reading* section at the end of this book.

Schwarzschild Black Hole

$$ds^2 = -\left(1 - \frac{2GM}{r}\right)\,dt^2 + \frac{dr^2}{\left(1 - \frac{2GM}{r}\right)} + r^2(d\phi^2 + \sin^2\phi\,d\vartheta^2), \tag{38.58}$$

is *indeed* a solution of the Einstein Vacuum Field Equation, (30.11), which we also repeat here:

$$\boxed{\textit{\textbf{Einstein Vacuum Field Equation:}}\qquad \mathbf{R}\mathrm{icci} = \mathbf{0} \quad\Longleftrightarrow\quad R_{ik} = 0.} \tag{38.59}$$

Let us define[17]

$$\boxed{f(r) = 1 - \frac{2GM}{r},}$$

so that (for later use),

$$\boxed{f' = \frac{2GM}{r^2}, \qquad \text{and} \qquad f'' = -\frac{4GM}{r^3}, \qquad \text{and} \qquad 1 - f = \frac{2GM}{r}.} \tag{38.60}$$

Let us label each basis 1-form with its associated spacetime coordinate. Then the Schwarzschild metric (38.58) may be written:

$$\mathbf{g} = g_{ij}(\boldsymbol{\theta}^i \otimes \boldsymbol{\theta}^j) = -\boldsymbol{\theta}^t \otimes \boldsymbol{\theta}^t + \boldsymbol{\theta}^r \otimes \boldsymbol{\theta}^r + \boldsymbol{\theta}^\phi \otimes \boldsymbol{\theta}^\phi + \boldsymbol{\theta}^\vartheta \otimes \boldsymbol{\theta}^\vartheta,$$

where

$$\boxed{\boldsymbol{\theta}^t = \sqrt{f}\,\mathbf{d}t\,, \qquad \boldsymbol{\theta}^r = \frac{\mathbf{d}r}{\sqrt{f}}\,, \qquad \boldsymbol{\theta}^\phi = r\,\mathbf{d}\phi\,, \qquad \boldsymbol{\theta}^\vartheta = r\sin\phi\,\mathbf{d}\vartheta.}$$

Writing the First Structural Equation on the left of each equation—with Einstein summation on m—then actually evaluating each derivative on the right, finally rewriting each result in terms of the basis 1-forms, we find that

$$\begin{aligned}
\boldsymbol{\omega}^t{}_m \wedge \boldsymbol{\theta}^m &= -\mathbf{d}\boldsymbol{\theta}^t = -\frac{f'}{2\sqrt{f}}\,\mathbf{d}r \wedge \mathbf{d}t = \frac{f'}{2\sqrt{f}}\,\sqrt{f}\,\mathbf{d}t \wedge \frac{\mathbf{d}r}{\sqrt{f}} = \frac{f'}{2\sqrt{f}}\,\boldsymbol{\theta}^t \wedge \boldsymbol{\theta}^r, \\
\boldsymbol{\omega}^r{}_m \wedge \boldsymbol{\theta}^m &= -\mathbf{d}\boldsymbol{\theta}^r = -\mathbf{d}\left[\frac{1}{\sqrt{f}}\right] \wedge \mathbf{d}r = -\partial_r\left[\frac{1}{\sqrt{f}}\right]\,\mathbf{d}r \wedge \mathbf{d}r = 0,
\end{aligned}$$

[17]In the following, we largely follow the notation of Dray (2015, pp. 259–261), although he writes the basis 1-forms as $\boldsymbol{\sigma}^i$, whereas we shall continue to write them as $\boldsymbol{\theta}^i$. Also note that our ϕ and ϑ are the reverse of his! See the footnote on page 388.

$$\boldsymbol{\omega}^{\phi}{}_{m} \wedge \boldsymbol{\theta}^{m} = -\mathbf{d}\boldsymbol{\theta}^{\phi} = -\mathbf{d}r \wedge \mathbf{d}\phi = \frac{\sqrt{f}}{r}\, r\,\mathbf{d}\phi \wedge \frac{\mathbf{d}r}{\sqrt{f}} = \frac{\sqrt{f}}{r}\,\boldsymbol{\theta}^{\phi} \wedge \boldsymbol{\theta}^{r},$$

$$\boldsymbol{\omega}^{\vartheta}{}_{m} \wedge \boldsymbol{\theta}^{m} = -\mathbf{d}\boldsymbol{\theta}^{\vartheta} = -s_{\phi}\,\mathbf{d}r \wedge \mathbf{d}\vartheta - r\,c_{\phi}\,\mathbf{d}\phi \wedge \mathbf{d}\vartheta = \frac{\sqrt{f}}{r}\boldsymbol{\theta}^{\vartheta} \wedge \boldsymbol{\theta}^{r} + \frac{\cot\phi}{r}\boldsymbol{\theta}^{\vartheta} \wedge \boldsymbol{\theta}^{\phi}.$$

In place of the more famous *Occam's Razor*, let us invoke (and simultaneously christen[18]) the logically indefensible, yet remarkably effective[19]

> ***Cartan's Razor***: *If a term is not explicitly prohibited from vanishing, it does vanish.*

Inspecting the equations above, and writing down *only* the nonvanishing terms that are *explicitly* forced upon us, and *assuming that all other terms vanish* (as dictated by *Cartan's Razor*!), we find,

$$\boxed{\boldsymbol{\omega}^{t}{}_{r} = \frac{f'}{2\sqrt{f}}\,\boldsymbol{\theta}^{t}, \qquad \boldsymbol{\omega}^{\phi}{}_{r} = \frac{\sqrt{f}}{r}\,\boldsymbol{\theta}^{\phi}, \qquad \boldsymbol{\omega}^{\vartheta}{}_{r} = \frac{\sqrt{f}}{r}\,\boldsymbol{\theta}^{\vartheta}, \qquad \boldsymbol{\omega}^{\vartheta}{}_{\phi} = \frac{\cot\phi}{r}\,\boldsymbol{\theta}^{\vartheta}.}$$

The Uniqueness Lemma (38.31) easily generalizes, so these *are* the solutions!

Next, we must differentiate these connection 1-forms to obtain the curvature 2-forms, bearing in mind that their antisymmetry, and the metric coefficient $g_{tt} = -1$, together imply:

$$\boldsymbol{\omega}^{r}{}_{t} = \boldsymbol{\omega}^{t}{}_{r}, \qquad \boldsymbol{\omega}^{r}{}_{\phi} = -\boldsymbol{\omega}^{\phi}{}_{r}, \qquad \boldsymbol{\omega}^{r}{}_{\vartheta} = -\boldsymbol{\omega}^{\vartheta}{}_{r}, \qquad \boldsymbol{\omega}^{\phi}{}_{\vartheta} = -\boldsymbol{\omega}^{\vartheta}{}_{\phi},$$

and

$$\boldsymbol{\omega}^{t}{}_{t} = \boldsymbol{\omega}^{r}{}_{r} = \boldsymbol{\omega}^{\phi}{}_{\phi} = \boldsymbol{\omega}^{\vartheta}{}_{\vartheta} = 0.$$

There are six independent curvature 2-forms, and we now spell out (in detail) the calculation of the first of these, leaving it to you to verify the remaining five.

The first step in calculating $\mathbf{d}\boldsymbol{\omega}^{i}{}_{j}$ is to revert to expressing $\boldsymbol{\omega}^{i}{}_{j}$ in terms of the $\mathbf{d}x^{i}$; this makes it easy to calculate the exterior derivative. The final result is then expressed in terms of $\boldsymbol{\theta}^{i} \wedge \boldsymbol{\theta}^{j}$. From (38.52),

$$\begin{aligned}
\boldsymbol{\Omega}^{t}{}_{r} &= \mathbf{d}\boldsymbol{\omega}^{t}{}_{r} + \boldsymbol{\omega}^{t}{}_{m} \wedge \boldsymbol{\omega}^{m}{}_{r} \\
&= \mathbf{d}\left[\frac{f'}{2}\right] \wedge \mathbf{d}t + \boldsymbol{\omega}^{t}{}_{t} \wedge \boldsymbol{\omega}^{t}{}_{r} + \boldsymbol{\omega}^{t}{}_{r} \wedge \boldsymbol{\omega}^{r}{}_{r} + \boldsymbol{\omega}^{t}{}_{\phi} \wedge \boldsymbol{\omega}^{\phi}{}_{r} + \boldsymbol{\omega}^{t}{}_{\vartheta} \wedge \boldsymbol{\omega}^{\vartheta}{}_{r} \\
&= \frac{f''}{2}\,\mathbf{d}r \wedge \mathbf{d}t + 0 + 0 + 0 + 0 \\
&= -\frac{f''}{2}\,\boldsymbol{\theta}^{t} \wedge \boldsymbol{\theta}^{r}.
\end{aligned}$$

We now list all six curvature 2-forms, starting with the one we just calculated:

$$\begin{aligned}
\boldsymbol{\Omega}^{t}{}_{r} &= -\frac{f''}{2}\,\boldsymbol{\theta}^{t} \wedge \boldsymbol{\theta}^{r}, \\
\boldsymbol{\Omega}^{t}{}_{\phi} &= -\frac{f'}{2r}\,\boldsymbol{\theta}^{t} \wedge \boldsymbol{\theta}^{\phi},
\end{aligned}$$

[18] To be clear, Cartan himself never said any such thing (so far as I know)—nevertheless, I choose to *name* this principle after him!

[19] No doubt a counterexample exists, but I personally do not recall ever having seen one.

$$\begin{aligned}
\boldsymbol{\Omega}^{t}{}_{\vartheta} &= -\frac{f'}{2r}\,\boldsymbol{\theta}^{t}\wedge\boldsymbol{\theta}^{\vartheta},\\
\boldsymbol{\Omega}^{\phi}{}_{r} &= -\frac{f'}{2r}\,\boldsymbol{\theta}^{\phi}\wedge\boldsymbol{\theta}^{r},\\
\boldsymbol{\Omega}^{\vartheta}{}_{r} &= -\frac{f'}{2r}\,\boldsymbol{\theta}^{\vartheta}\wedge\boldsymbol{\theta}^{r},\\
\boldsymbol{\Omega}^{\vartheta}{}_{\phi} &= \left[\frac{1-f}{r^2}\right]\boldsymbol{\theta}^{\vartheta}\wedge\boldsymbol{\theta}^{\phi}.
\end{aligned}$$

Again, the curvature 2-forms that are purely spatial are antisymmetric, but those with a temporal component are symmetric, because $g_{tt}=-1$:

$$\boldsymbol{\Omega}^{\vartheta}{}_{r}=-\boldsymbol{\Omega}^{r}{}_{\vartheta} \qquad \boldsymbol{\Omega}^{\phi}{}_{r}=-\boldsymbol{\Omega}^{r}{}_{\phi} \qquad \boldsymbol{\Omega}^{\vartheta}{}_{\phi}=-\boldsymbol{\Omega}^{\phi}{}_{\vartheta},$$

and

$$\boldsymbol{\Omega}^{t}{}_{r}=\boldsymbol{\Omega}^{r}{}_{t} \qquad \boldsymbol{\Omega}^{\phi}{}_{t}=\boldsymbol{\Omega}^{t}{}_{\phi} \qquad \boldsymbol{\Omega}^{\vartheta}{}_{t}=\boldsymbol{\Omega}^{t}{}_{\vartheta}.$$

Comparing the formulas above for $\boldsymbol{\Omega}^{i}{}_{j}$ with (38.55), i.e., $\boldsymbol{\Omega}^{i}{}_{j}=R^{i}{}_{jkl}\,\boldsymbol{\theta}^{k}\wedge\boldsymbol{\theta}^{l}$, and appealing to (38.60), we can immediately read off the components of the Riemann tensor:

$$\boxed{\begin{aligned}
R^{t}{}_{rtr}=-\frac{f''}{2} &= +\frac{2GM}{r^3}\\
R^{t}{}_{\phi t\phi}=R^{t}{}_{\vartheta t\vartheta}=R^{\phi}{}_{r\phi r}=R^{\vartheta}{}_{r\vartheta r}=-\frac{f'}{2r} &= -\frac{GM}{r^3}\\
R^{\vartheta}{}_{\phi\vartheta\phi}=\left[\frac{1-f}{r^2}\right] &= +\frac{2GM}{r^3}.
\end{aligned}}$$

While the meaning of the r coordinate is no longer so simple as "radial distance," observe that these formulas are in perfect accord with the formulas (30.2) and (30.3) (p. 312) that we deduced from the tidal forces arising from Newton's Inverse-Square Law of Gravitation!

We have arrived at a moment of high drama: *Is this putative black hole spacetime geometry* physically possible—*does it satisfy the Einstein Vacuum Field Equation?* Let us find out!

Summing the relevant Riemann tensor components to find the Ricci tensor, we easily see that the off-diagonal components of the Ricci tensor vanish, and, using (38.60), we now calculate its diagonal components, as follows:

$$\begin{aligned}
R_{tt} &= R^{r}{}_{trt}+R^{\phi}{}_{t\phi t}+R^{\vartheta}{}_{t\vartheta t}=+\frac{f''}{2}+\frac{f'}{2r}+\frac{f'}{2r}=-\frac{2GM}{r^3}+\frac{GM}{r^3}+\frac{GM}{r^3}=0,\\
R_{rr} &= R^{t}{}_{rtr}+R^{\phi}{}_{r\phi r}+R^{\vartheta}{}_{r\vartheta r}=-\frac{f''}{2}-\frac{f'}{2r}-\frac{f'}{2r}=+\frac{2GM}{r^3}-\frac{GM}{r^3}-\frac{GM}{r^3}=0,\\
R_{\phi\phi} &= R^{t}{}_{\phi t\phi}+R^{r}{}_{\phi r\phi}+R^{\vartheta}{}_{\phi\vartheta\phi}=-\frac{f'}{2r}-\frac{f'}{2r}+\frac{1-f}{r^2}=-\frac{GM}{r^3}-\frac{GM}{r^3}+\frac{2GM}{r^3}=0,\\
R_{\vartheta\vartheta} &= R^{t}{}_{\vartheta t\vartheta}+R^{r}{}_{\vartheta r\vartheta}+R^{\phi}{}_{\vartheta\phi\vartheta}=-\frac{f'}{2r}-\frac{f'}{2r}+\frac{1-f}{r^2}=-\frac{GM}{r^3}-\frac{GM}{r^3}+\frac{2GM}{r^3}=0.
\end{aligned}$$

Thus the black hole geometry discovered by Leftenant Schwarzschild in December of 1915, written down in his World War I trench on the Russian front, is *indeed* an exact solution of the equation that Einstein had discovered one month earlier!

To be clear, Schwarzschild had no idea at the time that his discovery represented a black hole. For him, and for Einstein, it *only* represented the spacetime geometry surrounding ordinary celestial bodies, such as the Earth or the Sun. It would take another *half-century* before physicists finally grasped that this solution, in its entirety, *also* presents the pure vacuum gravitational field of a black hole.

In January 1965, ten years after Einstein's death, Roger Penrose proved for the first time that Einstein's theory mathematically implies that the collapse of a sufficiently massive body must *necessarily* result in the formation of a black hole.[20] For this discovery, Penrose was awarded half of the 2020 Nobel Prize for Physics; the other half was divided equally between Reinhard Genzel and Andrea Ghez, for their independent experimental discovery of the supermassive black hole (Sagittarius A*) at the centre of our own Milky Way galaxy.

Before our long *Mathematical Drama* finally comes to its end, and the curtain falls, we have one last trick up our sleeve. Why did we insist on expressing the curvature in terms of f, f', and f'' till the bitter end, despite the fact that their values were known to us from the very beginning, on full display in (38.60)?

Instead of "merely" *verifying* that Schwarzschild's geometry solves Einstein's equation, we can instead *deduce* it *from* Einstein's equation!

Einstein's Field Equation implies

$$\boxed{R_{\phi\phi} = R_{\vartheta\vartheta} = -\frac{f'}{r} + \frac{1-f}{r^2} = 0 \quad \Longrightarrow \quad \frac{df}{dr} = \frac{1-f}{r} \quad \Longrightarrow \quad f(r) = 1 - \frac{C}{r},}$$

where C is a constant.

But then the remaining diagonal components of the Ricci tensor automatically vanish, too, thereby confirming that we have successfully solved Einstein's equation:

$$\boxed{R_{tt} = -R_{rr} = \frac{f''}{2} + \frac{f'}{r} = -\frac{C}{r^3} + \frac{C}{r^3} = 0.}$$

It remains to understand why the constant C is tied to the mass[21] of the black hole by $C = 2GM$. But to explain this would require that we delve much deeper into the fascinating world of gravitational physics than this *Mathematical Drama* will permit us to venture.

---

[20]In fact what Penrose proved was that the collapse must result in the formation of a spacetime singularity. But, in principle, this singularity could be something far more terrifying than a black hole—it could be a so-called ***naked singularity***, visible to the outside world! In 1969, Penrose conjectured that this could never happen, and that a "Cosmic Censor" would insist that every singularity be respectably ***clothed*** inside an event horizon, thereby forming a black hole! Many theoretical arguments lend credence to this ***Cosmic Censorship Hypothesis***, but it remains unproven in 2020.

[21]This connection can be established by a variety of arguments, and it is treated in *all* of the General Relativity texts we recommend in the *Further Reading* section at the end of this book. However, perhaps the single best, most *physical* treatment of the topic can be found in Schutz (2003, §§18, 21).

Chapter 39

Exercises for Act V

1-Forms

1. Dirac Delta Function as 1-Form. (This example is adapted from Schutz 1980.) Consider the vector space $C[-1,+1]$ of infinitely differentiable, real-valued functions $f(x)$ defined on the interval $-1 \leqslant x \leqslant +1$. The dual space of 1-forms is called the ***distributions***. An example of such a distribution is the ***Dirac Delta*** *Function* $\delta(x)$, which is the 1-form whose action on the "vector" $f(x)$ produces its value at $x=0$:

$$\langle \delta(x), f(x) \rangle = f(0).$$

As we now explain, this is not at all how Paul Dirac (1930) himself originally introduced his delta function into quantum mechanics.

(i) For any function $g(x)$ in $C[-1,+1]$, define a corresponding 1-form $\widetilde{g}(x)$ such that its action on the vector $f(x)$ is given by

$$\langle \widetilde{g}(x), f(x) \rangle \equiv \int_{-1}^{+1} g(x)\, f(x)\, dx.$$

Prove that $\widetilde{g}$ is indeed a 1-form.

(ii) Thus Dirac actually defined his delta function by insisting that

$$\int_{-1}^{+1} \delta(x)\, f(x)\, dx = f(0). \tag{39.1}$$

No ordinary function $\delta(x)$ can do this, but consider a smooth, bell-shaped curve centred at 0, and imagine it growing higher and higher as it becomes narrower and narrower, all the while obeying (39.1), with $f(x)=1$:

$$\int_{-1}^{+1} \delta(x)\, 1\, dx = 1.$$

Explain why, in the limit that the width of the bell curve vanishes, we recover (39.1).

2. In $\mathbb{R}^3$, give an explicit *geometrical* interpretation of the action $\boldsymbol{\omega}(\mathbf{v})$ of the 1-form $\boldsymbol{\omega} = 2\,\mathbf{d}x + \mathbf{d}y + 2\,\mathbf{d}z$ on the vector $\mathbf{v}$.

3. Explanation of *Covariant* and *Contravariant*. Let R_θ denote a rotation of $\mathbb{R}^2$ by θ. Then, as explained in [15.3], page 153, its matrix is

$$[R_\theta] = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}, \qquad \text{with inverse} \qquad [R_\theta]^{-1} = [R_{-\theta}] = [R_\theta]^T = \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

where $c \equiv \cos\theta$ and $s \equiv \sin\theta$. Let $\{\mathbf{e}_1, \mathbf{e}_2\}$ be the standard Cartesian basis, and let $\{\widetilde{\mathbf{e}}_1, \widetilde{\mathbf{e}}_2\}$ be the new basis that results from rotating the old basis by θ, so that $\widetilde{\mathbf{e}}_j = R_\theta \mathbf{e}_j$.

(i) Draw both bases in the same picture, and also draw a general vector $\mathbf{v} = \nu^j \mathbf{e}_j$, and draw its components $\widetilde{\nu}^j$ in the rotated basis: $\mathbf{v} = \widetilde{\nu}^j \widetilde{\mathbf{e}}_j$. Use your picture to explain why the new components are obtained by applying the *opposite* rotation matrix $[R_\theta]^{-1}$ to the original components:

$$\begin{bmatrix} \widetilde{\nu}^1 \\ \widetilde{\nu}^2 \end{bmatrix} = [R_\theta]^{-1} \begin{bmatrix} \nu^1 \\ \nu^2 \end{bmatrix}.$$

Because the components vary in the *opposite* ("contra-") way to the basis vectors, in the older literature a vector is called ***contravariant***.

(ii) The dual basis of $\{\mathbf{e}_1, \mathbf{e}_2\}$ is $\{\mathbf{d}x^1, \mathbf{d}x^2\}$. Let $\boldsymbol{\varphi} = \varphi_j\, \mathbf{d}x^j$ be a general 1-form, and let $\{\widetilde{\boldsymbol{\omega}}^1, \widetilde{\boldsymbol{\omega}}^2\}$ be the basis dual to $\{\widetilde{\mathbf{e}}_1, \widetilde{\mathbf{e}}_2\}$, so that the new components are given by $\boldsymbol{\varphi} = \widetilde{\varphi}_j\, \widetilde{\boldsymbol{\omega}}^j$. Show that the components of the 1-form transform in the *same* ("co-") way as the basis vectors:

$$\begin{bmatrix} \widetilde{\varphi}_1 \\ \widetilde{\varphi}_2 \end{bmatrix} = [R_\theta] \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}.$$

For this reason, 1-forms used to be called ***covariant vectors***, or ***covectors***. (NOTE: Modern Differential Geometry focuses on (coordinate-independent) geometric objects, which are unchanged by basis transformations such as this one. For this reason, the terms ***contravariant*** and ***covariant*** have largely disappeared from the modern literature. For an excellent, fully general discussion of this topic, see Schutz (1980, §2.26).)

Tensors

4. **Matrix multiplication.** From Schutz (1980). Show that in matrix algebra the action of one matrix on a second matrix, to produce their matrix product, is a tensor of valence $\left\{ \begin{smallmatrix} 2 \\ 2 \end{smallmatrix} \right\}$. (*Hint:* Recall that each matrix is itself a tensor of valence $\left\{ \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right\}$; see Section 33.2.)

5. **Tensor Contractions.** (From Schutz (1980).) How many different $\left\{ \begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \right\}$ tensors can be made by contraction on pairs of indices of the $\left\{ \begin{smallmatrix} 3 \\ 2 \end{smallmatrix} \right\}$ tensor $Q^{ijk}{}_{lm}$? How many $\left\{ \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right\}$ tensors (i.e., vectors) can be made by a second contraction?

6. **Tensor View of Matrix Multiplication as Contraction.** Referring back to Section 33.2 and Section 33.6, let **A** be a linear transformation, represented as a vector-valued tensor of valence $\left\{ \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right\}$: $\mathbf{v} \to \mathbf{A}(\mathbf{v})$. Let **B** be a second linear transformation, and let **C** be their composition: $\mathbf{C}(\mathbf{v}) \equiv \mathbf{B}\big[\mathbf{A}(\mathbf{v})\big]$. Show that the components of the $\left\{ \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right\}$ tensor **C** are given by contraction (aka matrix multiplication): $C^i{}_j = B^i{}_k A^k{}_j$.

2-Forms

7. **Factorizes in $\mathbb{R}^3$.** In $\mathbb{R}^3$, let $\boldsymbol{\Psi}$ be a general 2-form. Show that $\boldsymbol{\Psi}$ can always be factorized as the wedge product of two 1-forms: $\boldsymbol{\Psi} = \boldsymbol{\alpha} \wedge \boldsymbol{\beta}$.

8. **Need not factorize in $\mathbb{R}^4$.** In $\mathbb{R}^4$, let $\boldsymbol{\Psi}$ be a general 2-form.

(i) Show that $\boldsymbol{\Psi}$ *cannot* always be factorized as the wedge product of two 1-forms. *Hint:* If it can be factorized, consider $\boldsymbol{\Psi} \wedge \boldsymbol{\Psi}$. But now look back at §35.7.

(ii) If $\boldsymbol{\Psi} \wedge \boldsymbol{\Psi} = 0$, prove that $\boldsymbol{\Psi}$ *can* be factorized as the wedge product of two 1-forms.

(iii) Show that $\boldsymbol{\Psi}$ can always be expressed as the *sum* of *two* wedge products:

$$\boldsymbol{\Psi} = \boldsymbol{\alpha} \wedge \boldsymbol{\beta} + \boldsymbol{\gamma} \wedge \boldsymbol{\delta}.$$

9. Area formula for surface f = const. Let $\mathcal{S}$ be the surface with equation $f(x, y, z) = \mathrm{const}$.

(i) Show that the unit normal $\mathbf{n}$ to $\mathcal{S}$ is

$$\mathbf{n} = \frac{1}{\sqrt{(\partial_x f)^2 + (\partial_y f)^2 + (\partial_z f)^2}} \begin{bmatrix} \partial_x f \\ \partial_y f \\ \partial_z f \end{bmatrix}.$$

(ii) By thinking of $\mathbf{n}$ as the velocity of fluid flowing at unit speed orthogonally across $\mathcal{S}$, deduce that the area 2-form on $\mathcal{S}$ is

$$\mathcal{A} = \frac{\partial_x f\, \mathbf{d}y \wedge \mathbf{d}z + \partial_y f\, \mathbf{d}z \wedge \mathbf{d}x + \partial_z f\, \mathbf{d}x \wedge \mathbf{d}y}{\sqrt{(\partial_x f)^2 + (\partial_y f)^2 + (\partial_z f)^2}}.$$

3-Forms

10. Factorizes in $\mathbb{R}^4$. Show that any 3-form in $\mathbb{R}^4$ can be expressed as the wedge product of three 1-forms.

Differentiation

11. Closed and Exact Forms. In $\mathbb{R}^3$, let

$$\begin{aligned} \varphi &= 2x\,\mathbf{d}x + 2y\,\mathbf{d}y + 2z\,\mathbf{d}z, \\ \chi &= xy\,\mathbf{d}z, \\ \boldsymbol{\Psi} &= x\,\mathbf{d}y \wedge \mathbf{d}z + y\,\mathbf{d}x \wedge \mathbf{d}z. \end{aligned}$$

(i) Verify that φ is closed, and explain this by showing that it is *exact*: $\varphi = \mathbf{d}r^2$, where r is the distance from the origin.

(ii) Verify that $\boldsymbol{\Psi}$ is closed, and explain this by showing that it is *exact*: $\boldsymbol{\Psi} = \mathbf{d}\chi$.

12. Closed and Exact Forms. Continuing with the Forms of Exercise 11, compute each of the following exterior derivatives, (A) directly; (B) using the Leibniz product rule, (36.9).

(i) $\mathbf{d}(\chi \wedge \chi)$.

(ii) $\mathbf{d}(\varphi \wedge \chi)$.

(iii) $\mathbf{d}(\varphi \wedge \boldsymbol{\Psi})$.

(iv) $\mathbf{d}(\chi \wedge \boldsymbol{\Psi})$.

13. Closed and Exact Forms. If you did not do so earlier, prove each of the following facts:

(i) If $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Phi}$ are closed, then $\boldsymbol{\Upsilon} \wedge \boldsymbol{\Phi}$ is closed, too.

(ii) If $\boldsymbol{\Upsilon}$ is closed, then $\boldsymbol{\Upsilon} \wedge \mathbf{d}\boldsymbol{\Phi}$ is closed for all $\boldsymbol{\Phi}$.

(iii) If deg $\boldsymbol{\Phi}$ is *even*, then $\boldsymbol{\Phi} \wedge \mathbf{d}\boldsymbol{\Phi}$ is closed . *Hint:* See (35.4).

14. Vector Calculus Identities via Forms. Use Forms to derive the following identities of Vector Calculus:

(i)
$$\nabla \bullet \left[\underline{\varphi} \times \nabla f\right] = \left[\nabla \times \underline{\varphi}\right] \bullet \nabla f.$$

(ii)
$$\nabla \bullet \left[\underline{\varphi} \times \underline{\psi}\right] = \left[\nabla \times \underline{\varphi}\right] \bullet \underline{\psi} - \underline{\varphi} \bullet \left[\nabla \times \underline{\psi}\right].$$

(*Hint:* (36.9).) Note that (i) may now be derived as a special case of (ii).

15. Hodge Star Duality Operator ($\star$) (For a fully general discussion of this concept, see Schutz (1980), Baez and Muniain (1994), or Dray (2015). In particular, Baez and Muniain (1994, §§1, 5) contains an excellent discussion of the relation of Hodge duality to Maxwell's equations.)

(i) In n dimensions, show that the space of p-forms has the same dimension as the space of $(n-p)$-forms. These two spaces can therefore be put into 1-to-1 correspondence, and Hodge duality is a specific method of accomplishing this.

(ii) In $\mathbb{R}^3$, let $\mathcal{V}_3 = \mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}z$ be the volume 3-form. Given a basis p-form $\boldsymbol{\sigma}$, we define $\star$ to be the linear operator that fills in the "missing pieces" of $\mathcal{V}_3$:

$$\boxed{\boldsymbol{\sigma} \wedge \star\boldsymbol{\sigma} = \mathcal{V}_3.}$$

Deduce that

$$\begin{aligned} \star\mathbf{d}x &= \mathbf{d}y \wedge \mathbf{d}z, \\ \star\mathbf{d}y &= \mathbf{d}z \wedge \mathbf{d}x, \\ \star\mathbf{d}z &= \mathbf{d}x \wedge \mathbf{d}y, \end{aligned}$$

and, symmetrically, deduce that

$$\begin{aligned} \star(\mathbf{d}y \wedge \mathbf{d}z) &= \mathbf{d}x, \\ \star(\mathbf{d}z \wedge \mathbf{d}x) &= \mathbf{d}y, \\ \star(\mathbf{d}x \wedge \mathbf{d}y) &= \mathbf{d}z. \end{aligned}$$

Thus, in this case, two applications of the Hodge star operator yields the identity: $\star\star = 1$. Finally, in the interest of completeness, it is natural to define

$$\star 1 = \mathcal{V}_3 \qquad \text{and} \qquad \star\mathcal{V}_3 = 1.$$

(iii) Continuing in $\mathbb{R}^3$, we remind you of our earlier notation: $\underline{\varphi}$ is the vector corresponding ($\rightleftarrows$) to the 1-form φ. Show that Hodge duality yields two forms of the vector (cross) product:

$$\underline{\boldsymbol{\alpha}} \times \underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\gamma}} \quad \Longleftrightarrow \quad \boldsymbol{\alpha} \wedge \boldsymbol{\beta} = \star\boldsymbol{\gamma} \quad \Longleftrightarrow \quad \star(\boldsymbol{\alpha} \wedge \boldsymbol{\beta}) = \boldsymbol{\gamma}.$$

(iv) Still in $\mathbb{R}^3$, recall that every 2-form $\boldsymbol{\Psi}$ can be thought of as the flux associated with a flow of velocity $\underline{\boldsymbol{\Psi}}$, via (34.10):

$$\boldsymbol{\Psi} = \Psi^1\,(\mathbf{d}x^2 \wedge \mathbf{d}x^3) + \Psi^2\,(\mathbf{d}x^3 \wedge \mathbf{d}x^1) + \Psi^3\,(\mathbf{d}x^1 \wedge \mathbf{d}x^2) \quad \rightleftarrows \quad \underline{\boldsymbol{\Psi}} = \begin{bmatrix} \Psi^1 \\ \Psi^2 \\ \Psi^3 \end{bmatrix}.$$

If ψ is the 1-form corresponding to $\underline{\boldsymbol{\Psi}}$, deduce that the correspondence between $\boldsymbol{\Psi}$ and ψ is Hodge duality:

$$\star\boldsymbol{\Psi} = \psi \qquad \text{and} \qquad \boldsymbol{\Psi} = \star\psi.$$

(v) Now consider Minkowski spacetime, with metric

$$ds^2 = g_{ij}\, dx^i\, dx^j = -dt^2 + dx^2 + dy^2 + dz^2.$$

NOTE: Here, and henceforth, the metric coefficients g_{ij} are the *negatives* of those used earlier in Act II, for this is the convention favoured by the majority of more advanced physics texts we shall refer to, going forward.

Let the volume 4-form be $\mathcal{V}_4 = \mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}z \wedge \mathbf{d}t$, and define $\star\mathbf{d}x^i$ by

$$\boxed{\mathbf{d}x^i \wedge \star\mathbf{d}x^i = g_{ii}\, \mathcal{V}_4.}$$

Deduce that

$$\begin{aligned}
\star\mathbf{d}t &= \mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}z = \mathcal{V}_3,\\
\star\mathbf{d}x &= \mathbf{d}y \wedge \mathbf{d}z \wedge \mathbf{d}t,\\
\star\mathbf{d}y &= \mathbf{d}z \wedge \mathbf{d}x \wedge \mathbf{d}t,\\
\star\mathbf{d}z &= \mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}t.
\end{aligned}$$

(vi) Still in Minkowski spacetime, we define the dual $\star\boldsymbol{\Psi}$ of a basis 2-form $\boldsymbol{\Psi}$ by

$$\boxed{\boldsymbol{\Psi} \wedge \star\boldsymbol{\Psi} = \pm\mathcal{V}_4,}$$

in which we choose *minus* $(-)$ if $\boldsymbol{\Psi}$ contains $\mathbf{d}t$, and we choose *plus* $(+)$ otherwise. (Note that this rule encompasses the rule in the previous part as a special case.) Deduce that

$$\begin{aligned}
\star(\mathbf{d}x \wedge \mathbf{d}t) &= -\mathbf{d}y \wedge \mathbf{d}z, & \star(\mathbf{d}y \wedge \mathbf{d}z) &= \mathbf{d}x \wedge \mathbf{d}t,\\
\star(\mathbf{d}y \wedge \mathbf{d}t) &= -\mathbf{d}z \wedge \mathbf{d}x, & \star(\mathbf{d}z \wedge \mathbf{d}x) &= \mathbf{d}y \wedge \mathbf{d}t,\\
\star(\mathbf{d}z \wedge \mathbf{d}t) &= -\mathbf{d}x \wedge \mathbf{d}y, & \star(\mathbf{d}x \wedge \mathbf{d}y) &= \mathbf{d}z \wedge \mathbf{d}t.
\end{aligned}$$

(vii) Verify that for 2-forms in Minkowski spacetime, $\star\star = -1$. (This is in fact true of all p-forms.)

(viii) Recall that the Faraday 2-form is given by (34.22):

$$\mathbf{F} = \mathbf{Faraday} = \boldsymbol{\varepsilon} \wedge \mathbf{d}t + \mathbf{B}.$$

Use the previous parts to deduce that the Maxwell 2-form (34.26) is indeed the Hodge dual of the Faraday 2-form:

$$\star\mathbf{F} = \mathbf{Maxwell} = \boldsymbol{\beta} \wedge \mathbf{d}t - \mathbf{E}.$$

(ix) Likewise, verify that the dual of the spacetime current 1-form $\mathbf{J} = -\rho\, \mathbf{d}t + \mathbf{j}$, is indeed given by formula (36.18) for the current density 3-form:

$$\star\mathbf{J} = -\rho\, \mathcal{V}_3 + \left[\text{flux 2-form of } \underline{\mathbf{j}}\right] \wedge \mathbf{d}t$$

16. Conservation of Electric Charge

(i) Assuming that electric charge is ***conserved***, i.e., that it can neither be created nor destroyed, deduce that if V is the interior of a closed surface $\mathcal{S} = \partial V$, then

$$\frac{d}{dt}\iiint_V \rho \, d\mathcal{V} = -\oiint_{\mathcal{S}} \underline{\mathbf{j}} \cdot \mathbf{n} \, d\mathcal{A}.$$

(ii) Use Gauss's Theorem to deduce that the local expression of charge conservation is

$$\frac{d\rho}{dt} + \boldsymbol{\nabla} \cdot \underline{\mathbf{j}} = 0.$$

(iii) By taking the exterior derivative of the second pair of Maxwell's Equations, in the form

$$\mathbf{d} \star \mathbf{F} = 4\pi \star \mathbf{J},$$

deduce that charge conservation is a logical *consequence* of Maxwell's Equations, taking the form

$$\boxed{\mathbf{d} \star \mathbf{J} = 0.}$$

HISTORICAL NOTE: Faraday's 1831 experimental discovery of electromagnetic induction was ultimately crystalized by Maxwell into mathematical form: a changing magnetic field generates curl in the electric field—see (37.13). Thirty years later, *without any experimental evidence*, the purely theoretical considerations of this exercise forced Maxwell to conclude, in 1861, that, symmetrically, a changing electric field *must* generate curl in the magnetic field. Without this symmetry, electromagnetic waves could not exist—*Let there be light!*

17. Spinning Photons and Self-Duality. In Minkowski spacetime, a *complex* 2-form $\boldsymbol{\Psi}$ is called ***self-dual*** if

$$\star\boldsymbol{\Psi} = i\,\boldsymbol{\Psi},$$

and ***anti-self-dual*** if

$$\star\boldsymbol{\Psi} = -i\,\boldsymbol{\Psi},$$

where $\star$ is the Hodge duality operator introduced in Exercise 15.

(i) Let us verify that the Faraday 2-form $\mathbf{F}$ (or any 2-form, for that matter) can always be split into a self-dual part ${}^{+}\mathbf{F}$ and an anti-self-dual part ${}^{-}\mathbf{F}$, so that

$$\mathbf{F} = {}^{+}\mathbf{F} + {}^{-}\mathbf{F}.$$

Starting from this formula, and using the fact that ${}^{\pm}\mathbf{F} = \pm i\, {}^{\pm}\mathbf{F}$, deduce that

$${}^{+}\mathbf{F} = \frac{1}{2}\left[\mathbf{F} + i \star \mathbf{F}\right] \quad \text{and} \quad {}^{-}\mathbf{F} = \frac{1}{2}\left[\mathbf{F} - i \star \mathbf{F}\right].$$

(ii) It turns out that in quantum mechanics, these complex conjugate components ${}^{+}\mathbf{F}$ and ${}^{-}\mathbf{F}$ describe, respectively, *right*-spinning and *left*-spinning photons (the quanta of the electromagnetic field). Verify that *all four of Maxwell's Equations can be encapsulated into a single complex equation:*

$$\boxed{\mathbf{d}\,{}^{+}\mathbf{F} = -2\pi i \star \mathbf{J}.}$$

Integration

18. Closed and Exact 1-Forms. In the language of physics, $\varphi = \mathbf{d}f$ corresponds to a ***conservative*** force field, $\underline{\varphi} = \boldsymbol{\nabla} f$, and f is the ***potential energy***. Confirm that each of the following 1-forms is conservative by proving that each is *closed*: $\mathbf{d}\varphi = 0$. Then show that each is *exact*, by explicitly finding a matching potential energy function, f.

(i) $\varphi = x\,\mathbf{d}x + y^2\,\mathbf{d}y + z^3\,\mathbf{d}z$.

(ii) $\varphi = 3x^2y^2z\,\mathbf{d}x + 2x^3yz\,\mathbf{d}y + x^3y^2\,\mathbf{d}z$.

(iii) $\varphi = (2xy+z)\,\mathbf{d}x + (x^2+3y^2)\,\mathbf{d}y + x\,\mathbf{d}z$.

19. Closed and Exact 2-Forms. Verify that each of the following 2-forms is *closed*: $\mathbf{d}\boldsymbol{\Psi} = 0$. Then show that each is *exact*, by explicitly finding a 1-form potential φ such that $\boldsymbol{\Psi} = \mathbf{d}\varphi$.

(i) $\boldsymbol{\Psi} = 3x^2y^4\,\mathbf{d}x \wedge \mathbf{d}y$.

(ii) $\boldsymbol{\Psi} = 2xyz^3\,\mathbf{d}x \wedge \mathbf{d}y - 3x^2yz^2\,\mathbf{d}y \wedge \mathbf{d}z$.

20. Closed and Exact 3-Forms. In $\mathbb{R}^3$ every 3-form $\boldsymbol{\Upsilon}$ is proportional to the volume 3-form, $\mathcal{V} = \mathbf{d}x \wedge \mathbf{d}y \wedge \mathbf{d}z$, and, since there are no 4-forms, every 3-form is *closed*: $\mathbf{d}\boldsymbol{\Upsilon} = 0$. Show that each $\boldsymbol{\Upsilon}$ is *exact*, by explicitly finding a 2-form potential $\boldsymbol{\Psi}$ such that $\boldsymbol{\Upsilon} = \mathbf{d}\boldsymbol{\Psi}$.

(i) $\boldsymbol{\Upsilon} = yz\,\mathcal{V}$.

(ii) $\boldsymbol{\Upsilon} = 2(x+y+z)\,\mathcal{V}$.

21. Homogeneous Functions. (From do Carmo (1994).) A function $g : \mathbb{R}^3 \to \mathbb{R}$ is called ***homogeneous*** and of ***degree*** k if $g(tx, ty, tz) = t^k\, g(x, y, z)$.

(i) Show that such a g satisfies ***Euler's Relation***:

$$x\,\partial_x g + y\,\partial_y g + z\,\partial_z g = kg.$$

Hint: Differentiate the defining relation with respect to t.

(ii) If the 1-form,

$$\psi = a\,\mathbf{d}x + b\,\mathbf{d}y + c\,\mathbf{d}z$$

is such that a, b, c are homogeneous of degree k, and $\mathbf{d}\psi = 0$, then show that $\psi = \mathbf{d}f$, where

$$f = \frac{xa + yb + zc}{k+1}.$$

Hint: Write out $\mathbf{d}\psi = 0$ in component form, and then apply Euler's Relation.

(iii) Consider the flux 2-form corresponding to ψ:

$$\boldsymbol{\Psi} = a\,\mathbf{d}y \wedge \mathbf{d}z + b\,\mathbf{d}z \wedge \mathbf{d}x + c\,\mathbf{d}x \wedge \mathbf{d}y.$$

Show that if $\mathbf{d}\boldsymbol{\Psi} = 0$, then $\boldsymbol{\Psi} = \mathbf{d}\gamma$, where

$$\gamma = \frac{(zb - yc)\,\mathbf{d}x + (xc - za)\,\mathbf{d}y + (ya - xb)\,\mathbf{d}z}{k+2}.$$

(iv) If V is the interior of $\mathbb{S}^2$, so that $\partial V = \mathbb{S}^2$, use FTEC, in the form of (37.14), to show that

$$\iiint_V [\nabla^2 g]\; d\mathcal{V} = k \oiint_{\mathbb{S}^2} g\; d\mathcal{A}.$$

Hint: If $\mathbf{n}$ is the unit normal to $\mathbb{S}^2$, verify that Euler's Relation can be written $\mathbf{n} \cdot \nabla g = kg$.

Differential Geometry via Forms

22. Hairy Ball Theorem. (From do Carmo (1994).) As we discussed in connection with [19.8], the Poincaré–Hopf Theorem (19.6) implies that there cannot exist a singularity-free vector field on $\mathbb{S}^2$. This is often called the *Hairy Ball Theorem*: we cannot comb the hair on a coconut! Here is a second proof, using FTEC. Suppose that such a nonvanishing field $\mathbf{v}$ exists, and use it to construct the first vector $\mathbf{m}_1 = \mathbf{v}/|\mathbf{v}|$ of an orthonormal basis $\{\mathbf{m}_1, \mathbf{m}_2\}$. Then the Gauss Equation (38.30) yields

$$\mathbf{d}\boldsymbol{\omega}_{12} = -\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 = -\boldsymbol{\mathcal{A}}.$$

Now use FTEC to integrate this equation over $\mathbb{S}^2$, thereby obtaining a contradiction.

23. Conformal Curvature Formula. If the conformal metric of a surface is $ds^2 = (du^2 + dv^2)/\Omega^2$, show that

$$\boxed{\mathcal{K} = \Omega\,\nabla^2\Omega - \left[(\partial_u\Omega)^2 + (\partial_v\Omega)^2\right].}$$

As a simple check, try this on the Beltrami–Poincaré half-plane model of the hyperbolic plane, for which $\Omega = v$.

24. Vanishing Curvature *Characterizes* Euclidean Geometry. (From do Carmo (1994).) This exercise will establish that

> *The Gaussian curvature vanishes if and only if a 2-surface is locally Euclidean.*

By "locally Euclidean" we mean that around each point we can find (u, v)-coordinates such the metric is $ds^2 = du^2 + dv^2$.

(i) Choose an orthonormal frame $\{\mathbf{m}_1, \mathbf{m}_2\}$ for tangent vectors to the surface. By the Gauss Equation, (38.30),

$$\mathbf{d}\boldsymbol{\omega}_{12} = -\mathcal{K}\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 = 0.$$

Deduce that there exists a function ϕ—which we shall interpret as an angle— such that $\boldsymbol{\omega}_{12} = \mathbf{d}\phi$.

(ii) Consider a new orthonormal frame $\{\widetilde{\mathbf{m}}_1, \widetilde{\mathbf{m}}_2\}$, obtained by *rotating* $\{\mathbf{m}_1, \mathbf{m}_2\}$ *by angle* ϕ. Show that in the new frame, $\widetilde{\boldsymbol{\omega}}_{12} = 0$.

(iii) Deduce that

$$\mathbf{d}\widetilde{\boldsymbol{\theta}}^1 = 0 = \mathbf{d}\widetilde{\boldsymbol{\theta}}^2.$$

(iv) Deduce the existence of coordinates in which the metric is Euclidean.

25. Connection Forms of the Beltrami–Poincaré Half-Plane. The metric (5.8) of the Beltrami–Poincaré half-plane immediately yields the basis 1-forms:

$$ds^2 = \frac{dx^2 + dy^2}{y^2} \quad\Longrightarrow\quad \boldsymbol{\theta}^1 = \frac{1}{y}\,\mathbf{d}x \quad\text{and}\quad \boldsymbol{\theta}^2 = \frac{1}{y}\,\mathbf{d}y.$$

(i) Calculate $\mathbf{d}\boldsymbol{\theta}^1$ and $\mathbf{d}\boldsymbol{\theta}^2$.

(ii) Use Cartan's First Structural Equation,

$$\mathbf{d}\boldsymbol{\theta}^1 = -\boldsymbol{\omega}^1{}_2 \wedge \boldsymbol{\theta}^2, \qquad \mathbf{d}\boldsymbol{\theta}^2 = -\boldsymbol{\omega}^2{}_1 \wedge \boldsymbol{\theta}^1 = \boldsymbol{\omega}^1{}_2 \wedge \boldsymbol{\theta}^1,$$

to deduce that the (unique) solution is

$$\boldsymbol{\omega}^1{}_2 = -\frac{1}{y}\,\mathbf{d}x.$$

(iii) Calculate $\mathbf{d}\boldsymbol{\omega}^1{}_2$, and, by comparing the result with the Gauss Equation,

$$\mathbf{d}\boldsymbol{\omega}^1{}_2 = \mathcal{K}\,\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2,$$

deduce that the curvature of the hyperbolic plane is indeed $\mathcal{K} = -1$.

26. Curvature 2-Forms of a 2-Surface. Let $\{\mathbf{m}_1, \mathbf{m}_2\}$ be an orthonormal basis field for tangent vectors $\mathbf{v}$ to a 2-surface, so that

$$\begin{aligned} \nabla_{\mathbf{v}}\mathbf{m}_1 &= -\boldsymbol{\omega}^1{}_2(\mathbf{v})\,\mathbf{m}_2 \\ \nabla_{\mathbf{v}}\mathbf{m}_2 &= \boldsymbol{\omega}^1{}_2(\mathbf{v})\,\mathbf{m}_1. \end{aligned}$$

(i) Write down the connection matrix $[\boldsymbol{\omega}]$, and hence prove that

$$[\boldsymbol{\omega}] \wedge [\boldsymbol{\omega}] = 0.$$

(ii) Deduce that the curvature matrix (38.53) is

$$[\boldsymbol{\Omega}] = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{d}\boldsymbol{\omega}^1{}_2.$$

Thus, the curvature matrix is completely described by the single curvature 2-form $\boldsymbol{\Omega}^1{}_2$, governed by the Gauss Equation:

$$\boldsymbol{\Omega}^1{}_2 = \mathbf{d}\boldsymbol{\omega}^1{}_2 = \mathcal{K}\boldsymbol{\mathcal{A}}.$$

27. Curvature of Hyperbolic 3-Space, $\mathbb{H}^3$. Recall that the metric of $\mathbb{H}^3$ is given by (6.23), depicted in [6.6], page 80:

$$ds^2 = \frac{dx^2 + dy^2 + dz^2}{z^2} \qquad \Longrightarrow \qquad [\boldsymbol{\theta}] = \begin{bmatrix} \boldsymbol{\theta}^1 \\ \boldsymbol{\theta}^2 \\ \boldsymbol{\theta}^3 \end{bmatrix} = \frac{1}{z} \begin{bmatrix} \mathbf{d}x \\ \mathbf{d}y \\ \mathbf{d}z \end{bmatrix}.$$

(i) Recall that the switch from our original (O'Neill) notation, $\boldsymbol{\omega}_{ij}$, to our new, more standard notation, $\boldsymbol{\omega}^i{}_j$, induces a change of sign in Cartan's First Structural Equation, so that it now takes the form (see (38.54)):

$$\mathbf{d}[\boldsymbol{\theta}] = -[\boldsymbol{\omega}] \wedge [\boldsymbol{\theta}].$$

Deduce that the connection matrix is

$$[\boldsymbol{\omega}] = \frac{1}{z} \begin{bmatrix} 0 & 0 & -\mathbf{d}x \\ 0 & 0 & -\mathbf{d}y \\ \mathbf{d}x & \mathbf{d}y & 0 \end{bmatrix}$$

(ii) Use (38.53) to deduce that the curvature matrix is

$$[\boldsymbol{\Omega}] = \frac{1}{z^2} \begin{bmatrix} 0 & -\mathbf{d}x \wedge \mathbf{d}y & -\mathbf{d}x \wedge \mathbf{d}z \\ \mathbf{d}x \wedge \mathbf{d}y & 0 & -\mathbf{d}y \wedge \mathbf{d}z \\ \mathbf{d}x \wedge \mathbf{d}z & \mathbf{d}y \wedge \mathbf{d}z & 0 \end{bmatrix}.$$

(iii) Verify that $[\boldsymbol{\Omega}]$ may be rewritten as

$$[\boldsymbol{\Omega}] = \begin{bmatrix} 0 & -\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 & -\boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^3 \\ \boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 & 0 & -\boldsymbol{\theta}^2 \wedge \boldsymbol{\theta}^3 \\ \boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^3 & \boldsymbol{\theta}^2 \wedge \boldsymbol{\theta}^3 & 0 \end{bmatrix},$$

and use (38.55) to deduce that the Riemann tensor has components

$$R^1{}_{212} = R^1{}_{313} = R^2{}_{323} = -1.$$

(iv) Interpret these components of the Riemann tensor as sectional curvatures. What does the Sectional Jacobi Equation (29.21) therefore tell us about the behaviour of neighbouring geodesics that are initially launched in the same direction within this space? Conclude that it is meaningful to say that "$\mathbb{H}^3$ has constant curvature -1."

28. Cosmic Curvature. (The detailed solution to this exercise can be found in Dray (2015, §A.9) and in Sternberg (2012, §6.7).) The ***Friedmann–Lemaître–Robertson–Walker*** (FLRW) solutions of Einstein's Field Equations are now considered the ***standard model*** for the large-scale geometry of our expanding Universe. They were first discovered by Alexander Friedmann in 1922, rediscovered by the Jesuit priest, Georges Lemaître in 1927, and in 1935 Robertson and Walker jointly proved their uniqueness as *the* spatially homogeneous and isotropic geometries that are possible for the Universe.

Friedmann sought and found solutions of the *original* Einstein Equation, whereas we now know that the Cosmological Einstein Equation (30.25) is in fact correct, with cosmological constant $\Lambda > 0$. Fortunately, it is possible to adapt Friedmann's discovery to this new reality.

The FLRW metrics are given by

$$ds^2 = -dt^2 + R^2(t)\left[\frac{dr^2}{\left(1 - Kr^2\right)} + r^2(d\phi^2 + \sin^2\phi\, d\vartheta^2)\right],$$

where $R(t)$ measures the size of the expanding Universe at time t after the big bang, and K is the constant spatial curvature of the Universe for all time. See Penrose (2005, §§27.11, 27.12) for an excellent discussion (and lovely drawings) of the three cases, $K > 0$, $K = 0$, $K < 0$, including the effect of $\Lambda > 0$.

By following in the footsteps of our calculation of the curvature of the Schwarzschild black hole, write down the basis 1-forms, $\boldsymbol{\theta}^i$, take their exterior derivatives to find the connection 1-forms, $\boldsymbol{\omega}^i{}_j$, then calculate $\mathbf{d}\boldsymbol{\omega}^i{}_j$, and finally deduce the curvature 2-forms $\boldsymbol{\Omega}^i{}_j$ of such a Universe.

Further Reading

NOTES: In the following, I merely list title and author; full details can be found in the bibliography. Some of the following works are included simply because they are highly relevant, and I therefore believe they deserve to be brought to your attention (even if they are not quite my cup of tea). Most, however, are included because I consider them to be gems, and I strongly recommend them to you. Many other excellent books sit on my bookshelves, and I seek their counsel often, and yet they are not included here simply in order to cut down this (already long) list to a manageable size—I apologize to the authors of all those excellent works for failing to highlight them here. Finally, I also apologize to the authors of the wonderful works I have yet to discover!

Global Recommendations

First, let me set the stage with six works that I hold to be *invaluable*, the content of each of which spans multiple Acts of this book.

- *The Road to Reality*, by Roger Penrose.

  An extraordinary panorama of almost all of physics, and most of mathematics, by a master of both. Many of the insights can only be found here, and they are brought to life by Penrose's remarkable (and beautiful!) hand drawings.

- *Gravitation*, by Misner, Thorne, and Wheeler.

  Almost 50 years after its original publication in 1973, this classic remains one of the very best introductions to Einstein's geometrical theory of gravity (General Relativity) and to the Differential Geometry upon which it rests. It also contains one of the best, most geometrical introductions to Forms, including the curvature 2-forms that allow one to calculate the Riemann tensor efficiently. The new (2017) edition from Princeton University Press is beautifully done, and contains a new introduction by Charles Misner and Kip Thorne, discussing the exciting developments in the field since the book's original publication.

- *Differential Geometry in the Large*, by Heinz Hopf.

  Hopf was not only one of the towering figures of twentieth-century mathematics, he was also a master of exposition. Here, ideas of Differential Geometry and Topology (many of which are due to Hopf himself) come together in a beautiful way, explained with remarkable clarity and simplicity. Every time I return to this *Meisterwerk*, I feel that some beneficent magician has inserted more wonderful ideas into its pages, for I swear that *this* beautiful idea wasn't on the page the last time I looked!

- *Elementary Differential Geometry* (revised 2nd edition), by Barrett O'Neill.

  First published in 1966, this trail-blazing text pioneered the use of Forms at the undergraduate level. Today, more than a half-century later, O'Neill's work remains, in my view, the single most clear-eyed, elegant, and (ironically) *modern* treatment of the subject available—present company excepted!—at the undergraduate level.

- *Geometrical Methods of Mathematical Physics*, by Bernard Schutz.

  This work—now 40 years old!—is a timeless treasure trove, covering manifolds, tensors, Lie derivatives and Lie groups, Forms, Riemannian Geometry, gauge theories, and a host of other applications to physics. To achieve this, Professor Schutz channels *Star Trek*'s Mr. Spock. His *Vulcan* half enables him to erect a logical structure of crystalline perfection, in which everything is concisely and rigorously proven, and—unlike my (I hope delicious) cheeseburger approach in this book—Kashrut is strictly observed: concepts that depend on the existence of a metric are scrupulously and explicitly separated from those that do not. But, in tandem with this, Schutz is able to harness his *human* half to provide a wealth of intuition that reveals the underlying *geometric* reality.

- *Mathematics and Its History* (3rd edition), by John Stillwell.

  A remarkable panorama of all of mathematics through the lens of history. But make no mistake, this is not primarily a book about history, rather it is fundamentally a work about the interconnectedness and meaning of mathematics itself, all explained in a rather concise style (relative to mine!), with deep insight and lucidity.

Geometry in General

The following works are concerned with geometry in general, but especially with Hyperbolic Geometry. (*Differential* Geometry has its own category.)

- *Geometry and the Imagination*, by David Hilbert and S. Cohn-Vossen.

  A magnificent, deeply insightful survey of geometry, focusing on intuitive understanding, by one of the greatest mathematicians of the 20th century. The diagrams (drawn by K. H. Naumann and H. Bödeker) are *astonishingly* beautiful, to the point of causing me envy!

- *Experiencing Geometry*, by David W. Henderson and Daina Taimina.

  A highly unusual approach, philosophically akin to mine (but using the Moore method), focused on intuitive, *experimental* investigations of geometry. It contains significant discussion of parallel transport and holonomy. The overlap of their approach with mine is made clear by this quotation from the preface: "This book is based on a view of proof as a *convincing communication that answers—Why?*" (Their italics.)

- *Introduction to Geometry* (2nd edition), by H.S.M. Coxeter.

  A wonderfully clear survey by a modern master.

- *Geometry*, by Brannan, Esplen, and Gray.

  An excellent modern survey of geometry, based on Klein's vision of groups of transformations.

- *Euclidean and Non-Euclidean Geometries: Development and History*,
  by Marvin J. Greenberg.

  A valuable, detailed history of the development of Hyperbolic Geometry, including lengthy quotations from critical, private letters of Gauss, Bolyai, and many others.

- *The Poincaré Half-Plane*, by Saul Stahl.

  The title says it all.

- *Geometry Revealed*, by Marcel Berger.

  A much more advanced survey of geometry, with a focus on conceptual proofs and unsolved problems, by one of the great geometers of the 20th century.

Topology

- *Intuitive Topology*, by V. V. Prasolov.

  Less than 100 pages long, and filled with diagrams, this super-friendly introduction lives up to its title!

- *Euler's Gem*, by David S. Richeson.

  A masterful, mathematically accurate, yet riveting account of Euler's polyhedral formula, its history and the connected mathematical ideas.

- *Surface Topology*, by P. A. Firby and C. F. Gardiner.

  A very gentle, nicely illustrated introduction to the fundamental *geometric* ideas of topology.

- *First Concepts of Topology*, by W. G. Chinn and N. E. Steenrod.

  Another very gentle, nicely illustrated introduction to the fundamental *geometric* ideas of topology.

- *Topology: A Very Short Introduction*, by Richard Earl

  This remarkable little book lives up to its title, covering a huge range of fundamental ideas in just 140 pages, and it does so in a very clear, elementary, informal style. This is my new favourite introduction to the subject.

- *The Shape of Space* (3rd edition), by Jeffrey R. Weeks.

  A wonderfully lucid, engaging, elementary treatment of the topology of two and three dimensional spaces. The last of the four parts of the book deals with the possibility of detecting the topology of the Universe! An appendix contains John Horton Conway's famous ZIP Proof of the Classification Theorem for surfaces, beautifully illustrated by George K. Francis.

- *Three-Dimensional Geometry and Topology*, by William P. Thurston.

  Thurston won the Fields Medal for discovering that 3-manifolds are fundamentally built out of Hyperbolic Geometry. In this book you will hear Thurston's discoveries in his own distinctive voice, and although the difficulty of the *topology* accelerates rapidly, the first 100 pages provide a relatively elementary, highly original introduction to Hyperbolic Geometry that should not be missed.

Hopf's Line Fields and the Poincaré–Hopf Theorem in Physics

In this book I have sought to draw attention to *line fields* and Hopf's beautiful result that the Poincaré–Hopf Theorem applies to them, too, ((19.9), p. 213). These ideas have all but disappeared from modern *mathematics* textbooks, and I strongly believe that it is past time for a revival. That said, *physicists* never lost sight of the value of these ideas, and they have sustained them with wonderful new discoveries.

Here I am forced to list research papers rather than expository textbooks. The *only* expository mathematical treatments I know of are the lectures by Hopf (1956) himself (see previously), and the book by his student, Stoker (1969). Although the latest contributions to optics explicitly cite Hopf's work—going so far as to call fractional indices "Hopf indices"—these ideas were pioneered by J. F. Nye, and later, Sir Michael Berry, neither of whom seem to have been aware of Hopf (1956). (NOTE: In reading these papers, it is important to understand that what mathematicians call *holonomy*, physicists sometimes call *anholonomy*!)

- *The Topology of Ridge Systems*, by Roger Penrose.

  A completely elementary introduction (for nonmathematicians) to this circle of ideas, using fingerprints and palm prints as exemplars, for these are indeed line fields!

- *The Fingerprint of the Weyl Tensor*, Spinors and Space-Time, Vol. 2, pp. 229–233, by Roger Penrose.

  Penrose's "fingerprint" description of the Weyl curvature tensor reveals line fields having singularities possessing fractional indices.

- *I. Liquid Crystals. On the Theory of Liquid Crystals*, by F. C. Frank.

  One of the earliest appearances of line fields in physics. Note that here they are called ***nematic fields*** and their singularities are called ***topological defects***.

- *The Poincaré–Hopf Theorem for Line Fields Revisited*, by Diarmuid Crowley and Mark Grant.

  A nice review of the history and applications of the theorem; it also includes a general proof, in all dimensions.

- *Generic Singularities of Line Fields on 2D Manifolds*, by Ugo Boscain, Ludovic Sacchelliy, and Mario Sigalotti.

  A very nice review of the many appearances of line fields in nature, and their mathematical classification, related to umbilic points.

- *Natural Focusing and Fine Structure of Light: Caustics and Wave Dislocations*, by J. F. Nye

  Nye pioneered the introduction of line fields (polarization fields) into optics: the generic singularities are circularly polarized points in 2D or lines in 3D. Two other sorts of singularities are analyzed: those of intensity (caustics) and phase (vortices), which were pioneered by Nye and Sir Michael Berry.

- *The Quantum Phase, Five Years After*, by M. V. Berry, in Shapere and Wilczek (1989).

  A review by the discoverer of the quantum phase himself, explicitly relating the discovery to Differential Geometry, discussing the lines of curvature surrounding umbilic points, and making the crucial observation that a circuit around such a point results in a rotation of $\pm\pi$, corresponding to an index of $\pm\frac{1}{2}$, as illustrated in [19.14 a & A]. For a lovely illustration of this phenomenon, see Hilbert (1952, p. 189).

- *Geometry of Phase and Polarization Singularities, Illustrated by Edge Diffraction and the Tides*, by M. V. Berry.

  Edge diffraction and the tides also yield singular points with fractional indices.

- *Index Formulae for Singular Lines of Polarization*, by M. V. Berry.

  Formulas are obtained for the indices of lines in space on which the polarization of a monochromatic light field is purely circular (C) or purely linear (L). The indices ($\pm\frac{1}{2}$ for C lines and ± 1 for L lines) involve the electric or magnetic field and its derivatives on the line.

- *Polarization Singularities in the Clear Sky*, by M. V. Berry, M. R. Dennis, and R. L. Lee.

  An account of the pattern of polarization directions in daylight. The singularities (two near the Sun and two near the anti-Sun) are points in the sky where the polarization line field pattern has index $+\frac{1}{2}$ and the intensity of polarization is zero.

- *A Half-Century of Physical Asymptotics and Other Diversions: Selected Works by Michael Berry*, by M. V. Berry

  Sir Michael Berry's seminal contributions to optics, focusing on revealing ubiquitous manifestations of a variety of optical *singularities* in both everyday life and in cutting-edge, fundamental photonic research.

- *Singularities and Poincaré Indices of Electromagnetic Multipoles*, by Weijin Chen, Yuntian Chen, and Wei Liu.

  The authors map all the singularities of multipolar radiations of different orders, identify their indices, and show explicitly that the index sum over the entire momentum sphere is always 2, consistent with the Poincaré–Hopf Theorem.

- *Global Mie Scattering: Polarization Morphologies and the Underlying Topological Invariant*, by Weijin Chen, Qingdong Yang, Yuntian Chen, and Wei Liu.

  The Poincaré–Hopf Theorem (with fractional indices) is used to show that if you shine coherently polarized light of any form on any particle or particle clusters, there must be one direction where the scattering is either zero or circularly polarized.

- *Line Singularities and Hopf Indices of Electromagnetic Multipoles*, by Weijin Chen, Yuntian Chen, and Wei Liu

  The paper investigates the connections between half-integer Hopf indices, electromagnetic multipoles, and Bloch modes.

Differential Geometry

- *Differential Geometry: A Geometric Introduction*, by David W. Henderson.

  Of all the Differential Geometry books I will list, this one is the closest in spirit to my work, and that is why I have placed it first. I only discovered it late in the writing of VDGF, and was immediately convinced that Professor Henderson was a kindred spirit, a mathematical brother-in-arms. For the sad story of my attempt to make contact with him, see the footnote on page 252.

- *150 Years After Gauss' "Disquisitiones generales circa superficies curvas,"* by Peter Dombrowski.

  Using Gauss's letters to friends, private notebooks, and unpublished draft manuscripts, this uniquely valuable and extremely insightful work traces the evolution of Gauss's discoveries in Differential Geometry, culminating in his revolutionary, *General Investigations of Curved Surfaces*, which announced the *Theorema Egregium* to the world. As an added bonus, Dombrowski includes Gauss's 1827 masterpiece in its entirety, in the original Latin, with the English translation on facing pages.

- *A Comprehensive Introduction to Differential Geometry* (Vol. 2), by Michael Spivak

  Spivak's magnum opus actually comprises *five* volumes, but, for my purposes, Volume 2 is the most relevant, for it includes detailed, insightful analysis of the pioneering investigations of Gauss and of Riemann. This volume also includes Cartan's method of moving frames, and his two Structural Equations.

- *Differential Geometry of Curves and Surfaces* (2nd edition), by Thomas Banchoff and Stephen Lovett.

  While the mathematical machinery employed here is the standard one, featuring Christoffel symbols, and the like, it contains several highly original geometrical insights (due to Banchoff) that you will not find elsewhere. Furthermore, the accompanying website contains a wonderful array of Java applets for interactively exploring the concepts.

- *Differential Geometry of Curves and Surfaces*, by Kristopher Tapp.

  One of the best, most *geometrical* introductions of recent years. Although the mathematical machinery is the standard one—ugly equations filled with Christoffel symbols!—there are nevertheless many *conceptual* points of contact with my approach. It has many, nicely done colour diagrams, *many* excellent and interesting applications, and (like this book) it emphasizes the importance of parallel transport and holonomy. Highly recommended! (Note: It only treats 2-surfaces, so there is no discussion of the Riemann tensor or of General Relativity.)

- *Curved Spaces*, by P.M.H. Wilson.

  The first two-thirds of the book is a concise, modern, elegant treatment of Euclidean, Spherical, and Hyperbolic Geometry, employing the unifying power of Möbius transformations (exactly as I do). The last third is an introduction to Differential Geometry.

- *A First Course in Differential Geometry: Surfaces in Euclidean Space*, by L. M. Woodward and J. Bolton.

  The long final chapter of this work is the best introduction I know of (at the undergraduate level) to an important and beautiful topic that I have chosen to neglect entirely in this book: *minimal surfaces*. Recall that these are surfaces of vanishing mean curvature, realized by soap films: $\mathcal{H} \equiv \frac{\kappa_1+\kappa_2}{2} = 0$. The discussion includes the 1982 breakthrough discovery by Celso Costa of new minimal surfaces (beyond planes, catenoids, and helicoids) that are closed and without self-intersections. Furthermore, unusually, it also features a thorough discussion of the generalization to the case where $\mathcal{H} = \text{const.} \neq 0$, called ***surfaces of constant mean curvature*** (CMC surfaces).

- *Elementary Differential Geomety* (2nd edition), by Andrew Pressley.

  An excellent, elementary introduction, and, like the previous book, it contains a very nice introduction to minimal surfaces.

- *A Course of Differential Geometry and Topology*,
  by A. Mishchenko and A. Fomenko.

  An excellent, somewhat advanced, wide-ranging introduction that includes Riemannian Geometry and Homology Theory. It is distinguished by its large number of remarkable illustrations in the unmistakable, manga-style of Fomenko.

- *Differential Geometry: Connections, Curvature, and Characteristic Classes*,
  by Loring W. Tu

  This is the best book I know of for making the transition from VDGF to the much more advanced works that deal with Chern's characteristic classes, which have become important both in pure mathematics and in physics. The book does assume that the reader is already familiar with Forms, but my Act V provides more than enough preparation. Tu also assumes the reader is familiar with de Rham cohomology, and here my "primer" in Act V will likely not be sufficient, though I hope my physical and geometrical treatment—which is hard to find elsewhere—will greatly ease further study of the subject.

- *Lectures on Differential Geometry*, by S. S. Chern, W. H. Chen, and K. S. Lam.

  This book is 356 pages long, and contains fewer than 10 diagrams. I include this advanced work for one reason, and one reason only: Chern (upon whose lectures this is based) was one of the greatest geometers of the 20th century, and therefore *anything* he has to say, we should all listen to intently. In particular, Chern includes a lengthy treatment of Finsler Geometry, a vast generalization of Riemannian Geometry, pioneered by Chern himself. The only point of contact with this book is that Chern employs Forms as his principal tool—no surprise, given that Chern actually studied under Cartan himself!

Riemann's Curvature

- *A Panoramic View of Riemannian Geometry*, by Marcel Berger.

  An invaluable, panoramic view of the subject from one of the great geometers of the 20th century. It focuses on results and concepts, clearly and intuitively explained, and it omits all but the most essential details.

- *Riemannian Geometry: A Beginner's Guide* (2nd edition), by Frank Morgan.

  This nicely written and well-illustrated introduction packs a lot into only 129 pages. Despite its subtitle, I suspect that a true beginner would have a much easier time tackling this work *after* reading mine!

- *Riemannian Geometry*, by Manfredo P. do Carmo.

  An excellent (but advanced) treatment of many deep theorems, some of which are hard to find discussed clearly anywhere else.

- *Semi-Riemannian Geometry: With Applications to Relativity*, by Barrett O'Neill

  Semi-Riemannian Geometry is the geometry of spacetime, and this is simply the best mathematical work on the subject that I am aware of.

Einstein's Curved Spacetime (General Relativity)

- *Gravity from the ground up*, by Bernard Schutz.

  A *second* masterpiece! Whereas Schutz channelled Mr. Spock in his first masterpiece (see above), here he channels Captain Kirk! He employs true, Feynmanesque, *physical*, intuitive reasoning—using only high-school mathematics—to treat *all* aspects of gravity. Simultaneously, he succeeds in painting a vivid picture of the historical, *human* context of our evolving understanding of this fundamental force of Nature.

- *A Journey into Gravity and Spacetime*, by John Archibald Wheeler.

  A deeply insightful, largely nonmathematical, *conceptual* explication of Einstein's vision, by one of the greatest physicists of the 20th century, and one of the key figures in the rebirth of General Relativity in the 1960s: for example, it was Wheeler who coined the term ***black hole***. I have a personal connection with this work: having read it, I discovered and sent to Wheeler a Newtonian, purely geometrical proof of his crucial *Law of Double Curvature* (p. 142). As explained in the Prologue, Wheeler (1911–2008), whom I was fortunate to meet several times as Penrose's student, was a pivotal figure in my life, so I was *thrilled* when he wrote back that my proof had delighted him!

- *A First Course in General Relativity* (3rd edition), by Bernard Schutz.

  As of this writing, the third edition does not exist, but thanks to my correspondence with its author, I know it is coming. The second edition was already excellent, so the third edition will only be better! In particular, I predict that the new edition will greatly expand upon the treatment of gravitational waves, for these had not yet even been detected at the time of the second edition, and Schutz has been a key player in the field for more than 30 years. In 2019 Schutz received the Eddington Medal for his prescient theoretical work (Schutz 1986) that has now provided an entirely new method of calculating the Hubble constant from gravitational wave data.

- *Gravity: An Introduction to Einstein's Gravity*, by James B. Hartle.

  One of the very best, most *physical* (less mathematical) introductions to Einstein's theory, written by a master practitioner of the theory.

- *General Relativity*, by Robert M. Wald.

  I view this superb work as a significantly updated successor to Misner, Thorne, and Wheeler's *Gravitation*. It gracefully balances physical intuition with advanced mathematics, including topics that are rarely covered in a first course, such as Penrose's pioneering work on 2-spinors, and the singularity theorems of Penrose and Hawking.

- *General Relativity: A Geometric Approach*, by Malcolm Ludvigsen.

  If you are in a hurry to master General Relativity, you can hardly do better than this. It is a remarkably clean, elegant, geometrical, and *concise* presentation of the theory, but this is achieved (miraculously) without feeling rushed.

- *Einstein Gravity in a Nutshell*, by A. Zee.

  An excellent introduction to General Relativity, covering many unusual topics, written in a lively, informal, engaging, and highly opinionated style. (Are we related?)

Forms (in Mathematics)

- *A Geometric Approach to Differential Forms*, by David Bachman.

  One of a handful of brave books that seek to bring Forms to an *undergraduate* audience. And while it does include a very brief application to Differential Geometry, it does not include any discussion of the Riemann tensor or of General Relativity.

- *A Visual Introduction to Differential Forms and Calculus on Manifolds*, by Jon Pierre Fortney.

  The most recent (and longest) book to attempt to bring Forms to an *undergraduate* audience. While it does include a discussion of electromagnetism, it does not include any Differential Geometry, and thus does not include any discussion of the Riemann tensor or of General Relativity.

- *Differential Forms: A Heuristic Introduction*, by M. Schreiber.

  One of the earliest attempts to bring Forms to the masses. It contains some nice insights, but hardly any pictures! It does not include any discussion of the Riemann tensor or of General Relativity.

- *Differential Forms: A Complement to Vector Calculus*, by Steven H. Weintraub.

  As the title indicates, this is an attempt to bring Forms into the advanced undergraduate multivariable calculus classroom. While perfectly clear, it is not geometric in flavour, and has very few pictures.

- *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach* (4th edition), by John H. Hubbard and Barbara Burke Hubbard

  As with the previous book, this is an attempt to bring Forms into the advanced undergraduate multivariable calculus classroom, and here the approach is much more geometrical, rubbing shoulders with mine. However, it must be noted that Forms only appear for the first time on page 558.

- *Differential Forms and Applications*, by Manfredo P. do Carmo.

  At only 118 pages, this slim volume (by a very well-respected student of Chern) packs in a wealth of ideas and information, and some lovely exercises, some of which I have appropriated here (with attribution!).

Forms (in Physics)

- *Mathematical Methods of Classical Mechanics* (2nd edition), by V. I. Arnol'd.

  A masterpiece by one of the most brilliant mathematicians of the twentieth century. Chapter 7 is a self-contained, elegant, and insightful introduction to Forms, which are then applied to Hamiltonian mechanics.

- *Gauge Fields, Knots and Gravity*, by John Baez and Javier P. Muniain.

  A wide-ranging, highly original, modern introduction to electromagnetism, gauge fields, and gravity, most of which is expressed in the language of Forms. Amongst its *many* excellent features is an insightful discussion of the role of Hodge duality in Maxwell's Equations. Do not overlook the Notes that conclude each of the three parts of the book: they contain annotated recommendations for further study (very much like this section!), and they also contain fascinating historical vignettes and pithy quotations. It is immensely refreshing and helpful that the authors speak to the reader directly, in a friendly, informal way, as if addressing an intelligent friend sitting right next to them, instead of proclaiming dry theorems into the void. (Of course this is precisely what I have attempted to do in this book, too!)

- *Differential Forms and the Geometry of General Relativity*, by Tevian Dray.

  The only introduction to General Relativity to exclusively employ Forms as its mathematical engine. Of the many features that recommend it, I would note that it includes a much more complete discussion of Hodge duality than I have provided.

- *The Geometry of Physics* (3rd edition), by Theodore Frankel.

  An impressive tome that treats almost *every* topic covered by *all* of the books above, combined, *and* has more beyond that! Forms are the primary language of the book. Of particular note, however, is its unique treatment of the *geometry* of the Einstein tensor.

- *Curvature in Mathematics and Physics*, by Shlomo Sternberg.

  Forms are the primary language of this book. As the title suggests, it contains *many* interesting applications to both mathematics and physics. In particular, it treats the following physical topics in depth: Hodge duality and electromagnetism, explicit calculations of the geometry and orbits of the Schwarzschild solution, and the geometry of the extremely important Kerr solution (representing a spinning black hole), though he stops shy of actually computing the curvature 2-forms. But this list hardly does justice to the wealth of material that is covered. WARNING: the author describes the book as suitable for advanced undergraduates—it is not! That said, if you have mastered my Act V, then you stand to learn a great deal from this book.

Bibliography

Aleksandrov, A. D.
1969. *Mathematics: Its Content, Methods, Meaning*, Vol. 2. Chapter 7: *Curves and Surfaces*, 57–117. A. D. Aleksandrov, A. N. Kolmogorov, M. A. Lavrent'ev, eds. Translated from the Russian by S. H. Gould. 2nd ed., paperback. Cambridge, Mass.: MIT.

Arnol'd, V. and V. Vasil'ev
1991. Newton's Principia read 300 years later. *Current Science*, 61(2).

Arnol'd, V. I.
1973. *Ordinary Differential Equations*. Cambridge, Mass.: MIT.

Arnol'd, V. I.
1989. *Mathematical Methods of Classical Mechanics*, Vol 60 of *Graduate Texts in Mathematics*, 2nd ed. New York: Springer.

Arnol'd, V. I.
1990. *Huygens and Barrow, Newton and Hooke*. Basel: Birkhäuser Verlag.
Pioneers in mathematical analysis and catastrophe theory from evolvents to quasicrystals, Translated from the Russian by Eric J. F. Primrose.

Bachman, D.
2012. *A Geometric Approach to Differential Forms*, 2nd ed. New York: Birkhäuser.

Baez, J. C. and J. P. Muniain
1994. *Gauge Fields, Knots, and Gravity*, Vol. 4. Singapore: World Scientific.

Banchoff, T.
1970. Critical points and curvature for embedded polyhedral surfaces. *American Mathematical Monthly*, 77(5):475–485.

Banchoff, T. and S. Lovett
2016. *Differential Geometry of Curves and Surfaces*, 2nd ed. Natick, Mass.: AK Peters.

Berger, M.
2003. *A Panoramic View of Riemannian Geometry*. Berlin: Springer.

Berger, M.
2010. *Geometry Revealed: A Jacob's Ladder to Modern Higher Geometry*. Berlin: Springer.

Berry, M.
1991. Bristol anholonomy calendar. In *Sir Charles Frank, OBE, FRS: An eightieth birthday tribute*. Bristol: A. Hilger.

Berry, M.
2017. *A Half-Century of Physical Asymptotics and Other Diversions: Selected Works by Michael Berry*. Singapore: World Scientific.

Berry, M. V.
1990. Anticipations of the geometric phase. *Physics Today*, 43(12):34–40.

Berry, M. V.
2001. Geometry of phase and polarization singularities illustrated by edge diffraction and the tides. In *Second International Conference on Singular Optics (Optical Vortices): Fundamentals and Applications*, M. S. Soskin and M. V. Vasnetsov, eds. 4403:(1–12). Bellingham, Wash.: International Society for Optics and Photonics.

Berry, M. V.
2004. Index formulae for singular lines of polarization. *Journal of Optics A: Pure and Applied Optics*, 6(7):675–678.

Berry, M. V., M. R. Dennis, and R. L. Lee
2004. Polarization singularities in the clear sky. *New Journal of Physics*, 6:162–162.

Blaschke, W.
1929. *Vorlesungen über Differentialgeometrie und geometrische Grundlagen von Einsteins Relativitätstheorie* [2 volumes in one]. Chelsea Publishing.

Bloye, N. and S. Huggett
2011. Newton, the Geometer. *Newsletter of the European Mathematical Society*, (82):19–27.

Boscain, U., L. Sacchelli, and M. Sigalotti
2016. Generic singularities of line fields on 2D manifolds. *arxiv.org: 1605.06295*.

Bottazzini, U.
1999. Ricci and Levi-Civita: From Differential Invariants to General Relativity. In *The Symbolic Universe: Geometry and Physics 1890-1930*, J. Gray, ed. Oxford: Oxford University.

Boy, W.
1903. Über die Curvatura integra und die Topologie geschlossener Flächen. *Mathematische Annalen*, 57(2):151–184.

Brackenridge, J. B. and M. Nauenberg
2002. Curvature in Newton's Dynamics. In *The Cambridge Companion to Newton*, I. B. Cohen and G. E. Smith, eds., Cambridge Companions to Philosophy, Chapter 3, pp. 85–137. Cambridge: Cambridge University.

Brannan, D. A., M. F. Esplen, and J. Gray
2012. *Geometry*, 2nd ed. Cambridge: Cambridge University.

Carathéodory, C.
1937. The most general transformations of plane regions which transform circles into circles. *Bulletin of the American Mathematical Society*, 43(8):573–579.

Cartan, E.
1927. *Riemannian Geometry in an Orthogonal Frame: from lectures delivered by Élie Cartan at the Sorbonne in 1926-1927*. [Translated into Russian in 1960

by S. P. Finikov, and from Russian into English by V. V. Golberg in 2001.] Singapore: World Scientific.

Cartan, E.
1928. *Geometry of Riemannian Spaces*, Volume 13. Brookline, Mass.: Math Sci Press. Translation of 2nd ed., 1951, published 1983.

Cartan, E.
1945. *Les Systèmes Differentiels Extérieure et leur Applications Géométriques*. Paris: Hermann.

Casey, J.
1996. *Exploring Curvature*. Braunschweig: F. Vieweg und Sohn.

Chandrasekhar, S.
1995. *Newton's Principia for the Common Reader*. Oxford: The Clarendon Press.

Chen, W., Y. Chen, and W. Liu
2019. Singularities and Poincaré indices of electromagnetic multipoles. *Phys. Rev. Lett.*, 122:153907.

Chen, W., Y. Chen, and W. Liu
2020a. Line singularities and Hopf indices of electromagnetic multipoles. *Laser & Photonics Reviews*, 2000049.

Chen, W., Q. Yang, Y. Chen, and W. Liu
2020b. Global Mie scattering: Polarization morphologies and the underlying topological invariant. *ACS Omega*, 5(23):14157–14163.

Chern, S.-S., W.-H. Chen, and K. S. Lam
1999. *Lectures on Differential Geometry*, Volume 1. Singapore: World Scientific.

Coxeter, H.S.M.
1967. The Lorentz group and the group of homographies. In *Proc. Internat. Conf. Theory of Groups*, L. G. Kovacs and B. H. Neumann, eds., New York: Gordon and Breach.

Coxeter, H.S.M.
1969. *Introduction to Geometry*, 2nd ed. New York: Wiley.

Crowe, M. J.
1985. *A History of Vector Analysis: the Evolution of the Idea of a Vectorial System*. New York: Dover.

Crowley, D. and M. Grant
2017. The Poincaré–Hopf theorem for line fields revisited. *Journal of Geometry and Physics*, 117:187–196.

Darling, R.W.R.
1994. *Differential Forms and Connections*. Cambridge: Cambridge University.

Darrigol, O.
2015. The Mystery of Riemann's Curvature. *Historia Mathematica*, 42(1):47–83.

de Gandt, F.
1995. *Force and Geometry in Newton's Principia*. Princeton, N.J.: Princeton University. [Translated from the French original and with an introduction by Curtis Wilson.]

Diacu, F. and P. Holmes
1996. *Celestial Encounters: the Origins of Chaos and Stability*. Princeton, N.J.: Princeton University.

Dirac, P.A.M.
1930. *The Principles of Quantum Mechanics* [The International Series Of Monographs On Physics]. Oxford: The Clarendon Press.

do Carmo, M. P.
1992. *Riemannian Geometry*. Basel: Birkhäuser.

do Carmo, M. P.
1994. *Differential Forms and Applications*, Universitext. Berlin: Springer.

Dombrowski, P.
1979. *150 Years after Gauss' "Disquisitiones generales circa superficies curvas*," Vol. 62 of *Astérisque*. Paris: Société Mathématique de France.
With the original text of Gauss.

Dray, T.
2015. *Differential Forms and the Geometry of General Relativity*. Boca Raton: CRC Press.

Earl, R.
2019. *Topology: A Very Short Introduction*. Oxford: Oxford University.

Feynman, R. P.
1985. *QED: the Strange Theory of Light and Matter*. Princeton, N.J.: Princeton University.

Feynman, R. P., R. B. Leighton, and M. L. Sands
1963. *The Feynman Lectures on Physics*. Reading, Mass.: Addison-Wesley.

Firby, P. and C. F. Gardiner
2001. *Surface Topology*, 3rd ed.: Cambridge: Woodhead.

Flanders, H.
1989. *Differential Forms with Applications to the Physical Sciences*. Mineola, N.Y.: Dover.

Forbes, N. and B. Mahon
2014. *Faraday, Maxwell, and the Electromagnetic Field: How Two Men Revolutionized Physics*. Amherst, N.Y.: Prometheus.

Fortney, J. P.
2018. *A Visual Introduction to Differential Forms and Calculus on Manifolds*. Basel: Birkhäuser.

Frank, F. C.
1958. I. Liquid crystals: On the theory of liquid crystals. *Discuss. Faraday Soc.*, 25:19–28.

Frankel, T.
2011. *Gravitational Curvature: An Introduction to Einstein's Theory*. Mineola, N.Y.: Dover.

Frankel, T.
2012. *The Geometry of Physics: An Introduction*, 3rd ed. Cambridge: Cambridge University.

Fricke, R.
1926. *Lehrbuch der Algebra*, Vol. 2. Braunschweig: F. Vieweg und Sohn.

Gauss, C. F.
1827. *General Investigations of Curved Surfaces*. [Translated from the Latin and German by Adam Hiltebeitel and James Morehead, 1965 edition.] Hewlett, N.Y.: Raven Press.

Gauß, C. F.
1973. *Werke. Band VIII*. Hildesheim: Georg Olms Verlag. [Reprint of the 1900 original.]

Gindikin, S.
2007. *Tales of Mathematicians and Physicists*, 2nd English ed. Berlin: Springer.
Goodstein, D. L. and J. R. Goodstein
1996. *Feynman's Lost Lecture: The Motion of Planets Around the Sun*, 1st ed. New York: Norton.
Goodstein, J. R.
2018. *Einstein's Italian Mathematicians: Ricci, Levi-Civita, and the Birth of General Relativity*. Providence, R.I.: American Mathematical Society.
Gray, J.
1989. *Ideas of Space: Euclidean, Non-Euclidean, and Relativistic*, Oxford: Clarendon Press.
Greenberg, M. J.
2008. *Euclidean and Non-Euclidean Geometries*, 4th ed. New York: W.H. Freeman.
Guicciardini, N.
1999. *Reading the Principia*. Cambridge: Cambridge University.
The debate on Newton's mathematical methods for natural philosophy from 1687 to 1736.
Guicciardini, N.
2009. *Isaac Newton on Mathematical Certainty and Method* [Transformations: Studies in the History of Science and Technology]. Cambridge, Mass.: MIT.
Guillemin, V. and A. Pollack
1974. *Differential Topology*. Englewood Cliffs, N.J.: Prentice Hall.
Hankins, T. L.
1970. *Jean d'Alembert: Science and the Enlightenment*. Oxford: Clarendon Press.
Hayes, D. F., T. Shubin, G. L. Alexanderson, and P. Ross
2004. *Mathematical Adventures for Students and Amateurs*. Washington, D.C.: Mathematical Association of America.
Henderson, D. W.
1998. *Differential Geometry: A Geometric Introduction*. Upper Saddle River, N.J.: Prentice Hall.
Henderson, D. W. and D. Taimina
2005. *Experiencing Geometry: Euclidean and non-Euclidean with History*, 3rd ed. Upper Saddle River, N.J.: Pearson Prentice Hall.
Hilbert, D.
1902. *The Foundations of Geometry*. [2nd English ed. of 10th German ed. Translated by Leo Unger. Revised and enlarged by Dr. Paul Bernays, published 1971.] Chicago: The Open Court.
Hilbert, D.
1952. *Geometry and the Imagination*. New York: Chelsea.
Hirsch, M. W.
1976. *Differential Topology*, Vol. 33. Berlin: Springer.
Hopf, H.
1935. Über die drehung der tangenten und sehnen ebener kurven. *Compositio Mathematica*, 2:50–62.
Hopf, H.
1946–1956. *Differential Geometry In The Large: Seminar Lectures, New York University, 1946 And Stanford University*, 1956, 2nd ed., 1983, Vol. 1000 of *Lecture Notes In Mathematics*. Berlin: Springer.
Hubbard, J. and B. Hubbard
2009. *Vector Calculus, Linear Algebra, and Differential Forms*, 4th ed. Ithaca, N.Y.: Matrix Editions.
James, I. M.
1999. *History of Topology*, 1st ed. Amsterdam: Elsevier.
Jones, B.
1870. *The Life and Letters of Faraday*. London: Longmans, Green and Co.
Katz, V.
1979. The History of Stokes's Theorem. *Mathematics Magazine*, 52(3):146–156.
Kerr, R. P.
2008. Discovering the Kerr and Kerr-Schild metrics. *arXiv:0706.1109v2*.
Klein, F.
1928. *Development of Mathematics in the 19th Century* [English translation by M. Ackerman of the German original, with an extensive appendix by R. Hermann, published 1979]. Vol. 9 of *Lie groups*. Brookline, Mass.: Math Sci Press.
Knoebel, A.
2007. *Mathematical Masterpieces: Further Chronicles by the Explorers*, Undergraduate texts mathematics. Readings in mathematics. Berlin: Springer.
Koenderink, J. J.
1990. *Solid Shape*, Artificial intelligence. Cambridge, Mass.: MIT.
Kolmogorov, A. N. and A. P. Yushkevich
1996. *Mathematics of the 19th Century: Geometry, Analytic Function Theory*. Basel: Birkhäuser Verlag.
Kühnel, W.
2015. *Differential Geometry: Curves - Surfaces - Manifolds*, Vol. 77, 3rd ed. Providence, R.I.: American Mathematical Society.
Lanczos, C.
1970. *The Variational Principles of Mechanics*, 4th ed. Mineola, N.Y.: Dover.
Laugwitz, D.
1999. *Bernhard Riemann, 1826-1866: Turning Points In The Conception Of Mathematics*. Boston: Birkhäuser.
Levi-Civita, T.
1917. Nozione di parallelismo in una varietà qualunque e conseguente specificazione geometrica della curvatura riemanniana. In *Opere Matematiche: Memorie e Note, Pubblicate a cura dell'Accademia Nazionale dei Lincei*, Vol. IV (1917 to 1928), published 1960, pp. 1–39. Bologna: N. Zanichelli.
Levi-Civita, T.
1926. *The Absolute Differential Calculus: (Calculus of Tensors)*. Mineola, N.Y.: Dover, 1977.

Levi-Civita, T.
1931. Calcolo differenzale assoluto. In *Opere Matematiche: Memorie e Note, Pubblicate a cura dell'Accademia Nazionale dei Lincei*, Vol. V (1929 to 1937), published 1970, pp. 219–229. Bologna: N. Zanichelli.

Lightman, A. P., W. H. Press, R. H. Price, and S. A. Teukolsky
1975. *Problem Book in Relativity and Gravitation*. Princeton, N.J.: Princeton University.

Ludvigsen, M.
1999. *General Relativity: A Geometric Approach*. Cambridge: Cambridge University.

Mahon, B.
2017. *The Forgotten Genius of Oliver Heaviside: A Maverick of Electrical Science*. Buffalo, N.Y.: Prometheus.

Mahoney, M. S.
1994. *The Mathematical Career of Pierre de Fermat, 1601–1665*, 2nd ed. Princeton, N.J.: Princeton University.

Maxwell, J. C.
2002. *The Scientific Letters and Papers of James Clerk Maxwell*. Cambridge: Cambridge University.

Maxwell, J. C.
2003. *On the Transformation of Surfaces by Bending. The Scientific Papers Of James Clerk Maxwell*, Vol. 1, pp. 80–114. Mineola, N.Y.: Dover.

Mishchenko, A. S. and A. T. Fomenko
1988. *A Course of Differential Geometry and Topology*. Moscow: Mir Publishers.

Misner, C. W., K. S. Thorne, and J. A. Wheeler
1973. *Gravitation*. San Francisco: W. H. Freeman.
A marvellous new hardback edtition was published by Princeton University Press in 2017. It contains a new introduction by Charles Misner and Kip Thorne, discussing exciting developments in the field since the book's original publication.

Monastyrskiĭ, M. I.
1999. *Riemann, Topology, and Physics*, 2nd ed. Boston: Birkhäuser.

Morgan, F.
1998. *Riemannian Geometry: A Beginner's Guide*, 2nd ed. Wellesley, Mass.: AK Peters.

Needham, T.
1993. Newton and the transmutation of force. *Amer. Math. Monthly*, 100(2):119–137.

Needham, T.
1997. *Visual Complex Analysis*. Oxford: The Clarendon Press.

Needham, T.
2014. Visual Differential Geometry and Beltrami's Hyperbolic Plane. In *The Art of Science: From Perspective Drawing to Quantum Randomness*, R. Lupacchini and A. Angelini, eds., pp. 71–99. Berlin: Springer.

Newton, I.
1687. *The Principia: Mathematical Principles of Natural Philosophy*, 1999 edition. Berkeley, Calif.: University of California.
A new translation by I. Bernard Cohen and Anne Whitman, assisted by Julia Budenz, Preceded by "A guide to Newton's *Principia*" by Cohen.

Nye, J.
1999. *Natural Focusing and Fine Structure of Light: Caustics and Wave Dislocations*. Boca Raton: Taylor & Francis.

O'Neill, B.
2006. *Elementary Differential Geometry*, Revised 2nd ed. Amsterdam: Elsevier.

Pais, A.
1982. *"Subtle is the Lord…": The Science and the Life of Albert Einstein*. Oxford: Oxford University.

Penrose, R.
1978. The geometry of the universe. In *Mathematics Today: Twelve Informal Essays*, L. A. Steen, ed., pp. 83–127. New York: Vintage Books.

Penrose, R.
1979. The topology of ridge systems. *Annals of Human Genetics*, 42(4):435–444. [Also reprinted in his *Collected Works*, Oxford University, 2010: Vol. 3, pp. 347–357.]

Penrose, R.
2005. *The Road to Reality*. New York: Alfred A. Knopf. A complete guide to the laws of the universe.

Penrose, R. and W. Rindler
1984. *Spinors and Space-Time*. Cambridge: Cambridge University.

Phillips, E. R.
1979. Karl M. Peterson: The earliest derivation of the Mainardi–Codazzi equations and the fundamental theorem of surface theory. *Historia Mathematica*, 6(2):137–163.

Poincaré, H.
1892, 1893, 1899. *New Methods of Celestial Mechanics*, Vol. 13. American Institute of Physics, published 1993.

Pólya, G.
1954. *Mathematics and Plausible Reasoning*. Princeton, N.J.: Princeton University.

Prasolov, V. V.
1995. *Intuitive Topology*, Vol. 4. Providence: American Mathematical Society.

Pressley, A.
2010. *Elementary Differential Geometry* [Springer Undergraduate Mathematics Series], 2nd ed. Berlin: Springer.

Richeson, D. S.
2008. *Euler's Gem: The Polyhedron Formula and the Birth of Topology*. Princeton, N.J.: Princeton University.

Rosenfeld, B. A.
1988. *A History of Non-Euclidean Geometry: Evolution of the Concept of a Geometric Space*, Vol 12. Berlin: Springer.

Schreiber, M.
1977. *Differential Forms: A Heuristic Introduction*, Universitext. Berlin: Springer.

Schutz, B. F.
1980. *Geometrical Methods of Mathematical Physics*. Cambridge: Cambridge University.

Schutz, B. F.
1986. Determining the Hubble constant from gravitational wave observations. *Nature*, 323(6086):310–311.

Schutz, B. F.
2003. *Gravity from the Ground Up*. Cambridge: Cambridge University.

Schutz, B. F.
2021. *A First Course in General Relativity*, 3rd ed. Cambridge: Cambridge University.

Segerman, H.
2016. *Visualizing Mathematics with 3D Printing*. Baltimore: Johns Hopkins.

Shapere, A. and F. Wilczek
1989. *Geometric Phases in Physics*, Vol. 5. Singapore: World Scientific.

Shenitzer, A. and J. Stillwell
2002. *Mathematical Evolutions*. Washington, D.C.: Mathematical Association of America.

Spivak, M.
1999. *A Comprehensive Introduction to Differential Geometry*, Vol. 2, 3rd ed. Houston: Publish or Perish.

Stahl, S.
1993. *The Poincaré Half-Plane: A Gateway to Modern Geometry*. Boston: Jones and Bartlett.

Sternberg, S.
2012. *Curvature in Mathematics and Physics*, Dover books on mathematics. Dover.

Stewart, G. W.
1993. On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566.

Stillwell, J.
1995. *Classical Topology and Combinatorial Group Theory*, Vol. 72, 2nd ed. Berlin: Springer.

Stillwell, J.
1996. *Sources of Hyperbolic Geometry*, Vol. 10 of *History of Mathematics*. Providence: American Mathematical Society.

Stillwell, J.
2005. *The Four Pillars of Geometry*. Berlin: Springer.

Stillwell, J.
2010. *Mathematics and its History*, 3d ed. Berlin: Springer.

Stoker, J. J.
1969. *Differential Geometry*, Vol. 20 of *Pure and Applied Mathematics*. New York: Wiley-Interscience.

Taylor, E. F. and J. A. Wheeler
1992. *Spacetime Physics: Introduction to Special Relativity*, 2nd ed. New York: W.H. Freeman.

Taylor, E. F. and J. A. Wheeler
2000. *Exploring Black Holes: Introduction to General Relativity*. San Francisco: Addison Wesley Longman.

Thorne, K. S. and R. D. Blandford
2017. *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*. Princeton, N.J.: Princeton University.

Thurston, W. P.
1997. *Three-Dimensional Geometry and Topology*, Vol. 35. Princeton, N.J.: Princeton University.

Tu, L. W.
2017. *Differential Geometry: Connections, Curvature, and Characteristic Classes*. Berlin: Springer.

Wald, R. M.
1984. *General Relativity*. Chicago: University of Chicago.

Watson, G. N.
1917. A problem of analysis situs. *Proceedings of the London Mathematical Society*, 2(1): 227–242.

Weeks, J. R.
2020. *The Shape of Space*, ed. Boca Raton: CRC Press.

Weintraub, S. H.
1997. *Differential Forms: A Complement to Vector Calculus*. San Diego: Academic Press.

Westfall, R. S.
1980. *Never at Rest: A biography of Isaac Newton*. Cambridge: Cambridge University.

Wilson, P.M.H.
2008. *Curved Spaces: From Classical Geometries to Elementary Differential Geometry*. Cambridge: Cambridge University.

Wiltshire, D. L., M. Visser, and S. Scott
2009. *The Kerr Spacetime: Rotating Black Holes in General Relativity*. Cambridge: Cambridge University.

Woodward, L. M. and J. Bolton
2019. *A First Course in Differential Geometry: Surfaces in Euclidean Space*. Cambridge: Cambridge University.

Index