Discrete Differential Geometry

Computer Science and Engineering
University of California San Diego

Albert Chern

Last update: November 19, 2020

Contents

Curves and Surfaces

1. Plane Curves

Curves are one-dimensional geometric objects which are straight or curved within a higher-dimensional ambient space. They are widely used to represent thin physical objects such as rods and wires, as well as to describe the trails of moving objects. Less common but no less interesting examples of curves include singular features under physical processes with concentration mechanisms; *e.g.* rivers in an eroded terrain or tornadoes in a fluid.

Mathematically, one views curves

- **explicitly**—a curve is a family of points that can be continuously parametrized by a single variable.
- **implicitly**—a curve in the plane $\mathbb{R}^2$ is the level set $\{(x,y) \in \mathbb{R}^2 \,|\, f(x,y) = 0\}$ of a continuous scalar function $f : \mathbb{R}^2 \to \mathbb{R}$; a curve in $\mathbb{R}^3$ is the intersection of two level set surfaces $f(x,y,z) = 0, g(x,y,z) = 0$.

These two representations are locally equivalent in generic cases. Throughout this lecture we study the explicit viewpoint of curves.

In this chapter, we focus on **plane curves**, which are curves lying in a plane. We take plane curves as elementary examples in the broader context of differential geometry. In particular, we will take a glimpse into a few recurring concepts including

- **What are geometric measurements?**—When a piece of geometry is described by an artificial parametrization or under a coordinate system, a meaningful formula for a geometric measurement should yield values that do not depend on the choice of those artificial structures.
- **Local and global geometric measurements**—Curvatures measure the how much a curve locally deviates from being straight. The total length of a curve is a global geometric measurement. Can you think of other examples?
- **Rigid relationships between geometric measurements**—Can you deform a closed curve in the shape of "8" to a closed curve in the shape of "0" without creating infinite curvature?

1.1 Parametrized and Unparametrized Curves

A **parametrized plane curve** is a continuous function $\gamma\colon I \to \mathbb{R}^2$, where the domain $I \subseteq \mathbb{R}$ is some (connected) interval on the real line.

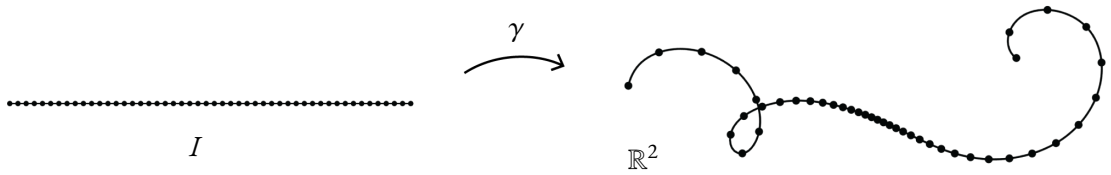


Figure 1.1 A parametrized plane curve is a map from a real interval to the plane. The speed of the parametrization is visualized by the markers along the curve.

Think of a parametrized curve $\gamma\colon I \to \mathbb{R}^2$ as the function that describes the position $\gamma(t) \in \mathbb{R}^2$ of a moving car in the plane at each time $t \in I$. Such a function describes not only the path trailed by the car, but also encodes the kinematic information, *e.g.* the acceleration experienced by the passenger. Suppose we are interested only in the geometry of the path trailed by the moving car. We call the path trailed by the car γ the **unparametrized curve**, or simply the **curve**, underneath the parametrized curve γ.

Definition 1.1 An **(unparametrized) plane curve** is a (multi)set of points in $\mathbb{R}^2$ that admits a parametrization by one variable from an interval $I \subseteq \mathbb{R}$. Two parametrizations $\gamma_1\colon I_1 \to \mathbb{R}^2$ and $\gamma_2\colon I_2 \to \mathbb{R}^2$ are said to yield the same curve if there exists a continuous and continuously invertible function $\phi\colon I_1 \to I_2$, called a **reparametrization**, such that $\gamma_1(t) = \gamma_2(\phi(t))$ for all $t \in I_1$; in short $\gamma_1 = \gamma_2 \circ \phi$.

R A curve is not associated with any specific parametrization. A parametrized curve can be thought of as a curve decorated with a distinguished parametrization as an additional structure.

▪ **Example 1.1** Following a similar philosophy, one also introduces the notion of an **oriented curve**. It is a curve with an additional structure called orientation. More precisely, two parametrized curves $\gamma_1\colon I_1 \to \mathbb{R}^2$ and $\gamma_2\colon I_2 \to \mathbb{R}^2$ are said to yield the same oriented curve if there exists a continuous and *increasing* (*i.e.* orientation-preserving) function $\phi\colon I_1 \to I_2$ such that $\gamma_1 = \gamma_2 \circ \phi$. In other words, two parametrizations of the same curve are of different orientations if there is no increasing reparametrization between them. ▪

A **closed** parametrized curve is a continuous function $\gamma\colon \mathbb{T}^1 \to \mathbb{R}^2$, where the domain is now a periodic domain $\mathbb{T}^1 = (\mathbb{R} \mod 2\pi)$; that is, a closed parametrized curve is a parametrized curve $\gamma\colon \mathbb{R} \to \mathbb{R}^2$ with the periodicity condition $\gamma(t + 2\pi) = \gamma(t)$ for all $t \in \mathbb{R}$. Equivalently, a closed parametrized curve is a map $\gamma\colon \mathbb{S}^1 \to \mathbb{R}^2$ where $\mathbb{S}^1$ denotes the unit circle.

Similar to Definition 1.1, unparametrized closed curves are closed parametrized curves modulo reparametrizations.

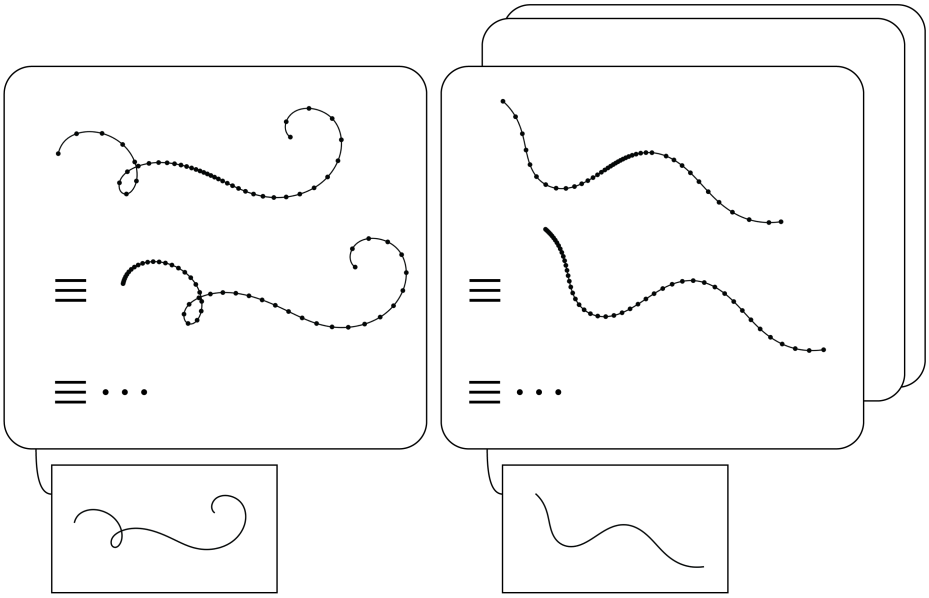


Figure 1.2 Two parametrized curves (curves with markers) are said to be equivalent if they are related merely by reparametrization. An unparametrized curve (curve without markers) is an equivalence of parametrized curves.

1.1.1 Regular Curves

In the above we only assume the parametrization γ is a continuous function. It can exhibit kinks, cusps, or even fractals along the curve. Many geometric measurements of a curve we will study later require taking derivatives. Therefore we should find a condition to exclude the non-smooth features.

Definition 1.2 A parametrized curve $\gamma : I \to \mathbb{R}^2$ is **regular** if γ has continuous first derivative, and has non-vanishing speed $|\gamma'(t)| \neq 0$ for all $t \in I$. A regular parametrized curve is k-th order smooth if γ has continuous k-th derivative.

Notation 1.1 We write $\gamma \in C^k(I;\mathbb{R}^2)$ if γ is a map from I to $\mathbb{R}^2$ and has continuous k-th derivative. If in addition $|\gamma'| \neq 0$, then γ is called a C^k regular parametrized plane curve. We also call a C^k regular parametrized curve a G^k curve. Here G stands for **geometric smoothness**.

The condition $|\gamma'| \neq 0$ is necessary for regularity in the resulting shape. The image of an $\mathbb{R}^2$-valued smooth function can still feature sharp corners and cusps when it is allowed to stop ($|\gamma'| = 0$). See Table 1.1.

Exercise 1.1 Provide an example of a parametrized curve γ that has continuous first derivative, but its image features sharp corners. ∎

A G^k *unparametrized* regular curve arises from identifying C^k regular parametrized curves that differ by a C^k*-smooth* (as opposed to just being continuous) bijective reparametrization function $\phi \in C^k(I_1;I_2)$ that has non-vanishing first derivatives $|\phi'| \neq 0, |\phi^{-1\prime}| \neq 0$. These smooth reparametrizations are called **diffeomorphisms**.

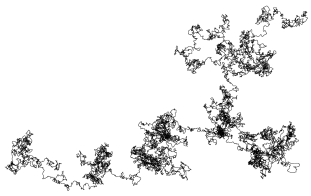
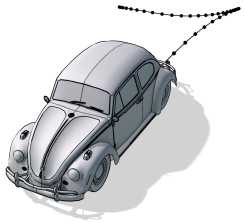
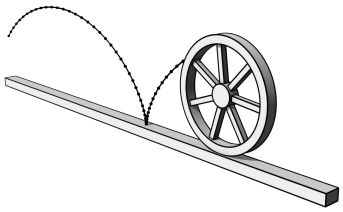
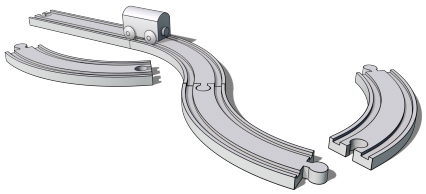
Class	Behavior	Archetypal example
$\gamma \in C^0(I;\mathbb{R}^2)$	The curve is continuous, but it may not have a well-defined velocity.	The trajectory of a Brownian motion. 
$\gamma \in C^1(I;\mathbb{R}^2)$	The parametrized curve has continuous velocity, but the curve can have sharp corners.	The trajectory of a car reversed out of a parking space. 
$\gamma \in C^2(I;\mathbb{R}^2)$	The parametrized curve has continuous acceleration, but the curve can have sharp corners.	Cycloid—the trajectory of a circumference point of a rolling wheel. 
$\gamma \in C^1(I;\mathbb{R}^2)$, $\lvert\gamma'\rvert \neq 0$	The curve has continuous tangents.	A pieced together track. 
$\gamma \in C^2(I;\mathbb{R}^2)$, $\lvert\gamma'\rvert \neq 0$	The curve has continuous tangents and curvature.	Elastic splines. 

Table 1.1 Curves parametrized by functions in various smoothness class.

Notation 1.2 We write $\phi \in \mathrm{Diff}^k(I_1; I_2)$ and say ϕ is a k-th order diffeomorphism if $\phi \in C^k(I_1; I_2)$, $\phi^{-1}\colon I_2 \to I_1$ exists, and $\phi^{-1} \in C^k(I_2; I_1)$. If $k \geq 1$, we require $|\phi'| \neq 0$ and $|\phi^{-1\prime}| \neq 0$.

Suppose $\gamma\colon I \to \mathbb{R}^2$ is a C^k regular parametrized curve. Any other C^k regular parametrization $\widetilde{\gamma}\colon \widetilde{I} \to \mathbb{R}^2$ for the same unparametrized curve arises in the form $\widetilde{\gamma} = \gamma \circ \phi$ for some diffeomorphism $\phi \in \mathrm{Diff}^k(\widetilde{I}; I)$. The new speed is given by $|\widetilde{\gamma}'(t)| = |\gamma'(\phi(t))||\phi'(t)|$.

In the rest of the chapter, we assume the curve is regular with a sufficient order of smoothness.

1.2 Geometric Quantities

What are geometric measurements? Well, it depends on what the geometric object is.

Let us say we are in a context where the geometric object of interest is a *parametrized* curve $\gamma\colon I \to \mathbb{R}^2$, such as the description of a car traversing in the plane where the kinematics are equally interesting as the shape of the trajectory. Then any expression in terms of γ is a valid geometric quantity. For example, the velocity vector $v(t) = \gamma'(t)$ is a geometric quantity of a parametrized curve.

If in a context where the geometric object is an *oriented* curve, then a geometric quantity as an expression involving some parametrization γ must be invariant under orientation-preserving reparametrizations.

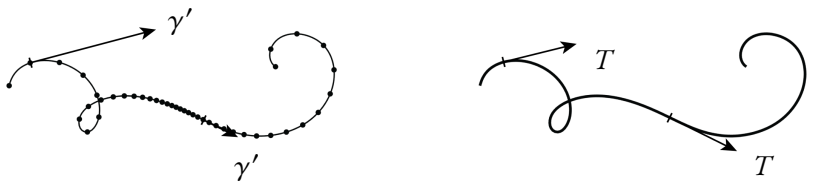


Figure 1.3 The velocity γ' is a measurement of a parametrized curve. The tangent vector $T = \gamma'/|\gamma'|$ is independent of the parametrization.

Proposition 1.1 *The tangent vector $T(t) = \gamma'(t)/|\gamma'(t)|$ is a geometric quantity of an oriented curve.*

Exercise 1.2 Show Proposition 1.1. That is, for each parametrized curve $\gamma\colon I \to \mathbb{R}^2$ and an orientation-preserving reparametrization $\phi\colon \widetilde{I} \to I$, $\phi' > 0$, which generates another parametrized curve $\widetilde{\gamma}\colon \widetilde{I} \to \mathbb{R}^2$ by $\widetilde{\gamma}(t) := \gamma(\phi(t))$, show that the new tangent vector

$$\widetilde{T}(t) := \frac{\widetilde{\gamma}'(t)}{|\widetilde{\gamma}'(t)|} \quad \text{satisfies} \quad \widetilde{T}(t) = T(\phi(t)).$$

∎

Suppose we are in a situation where the geometric object is an unparametrized curve with neither a distinguished orientation nor a distinguished parametrization. Then a geometric quantity is an expression involving γ that is invariant under any reparametrization.

Theorem 1.2 The **total arclength** of a parametrized curve $\gamma\colon I \to \mathbb{R}^2$

$$L[\gamma] := \int_I |\gamma'(t)|\, dt$$

is invariant under reparametrizations. Therefore the total arclength is a geometric quantity of the underlying unparametrized curve.

Proof. Let $\phi\colon \widetilde{I} \to I$ be a reparametrization, giving rise to another parametrized curve $\widetilde{\gamma}\colon \widetilde{I} \to \mathbb{R}^2$ given by $\widetilde{\gamma}(\tilde{t}) = \gamma(\phi(\tilde{t}))$. The total arclength of $\widetilde{\gamma}$ is

$$L[\widetilde{\gamma}] = \int_{\widetilde{I}} |\widetilde{\gamma}'(\tilde{t})|\, d\tilde{t} = \int_{\widetilde{I}} |\gamma'(\phi(\tilde{t}))\phi'(\tilde{t})|\, d\tilde{t} = \int_{\widetilde{I}} |\gamma'(\phi(\tilde{t}))||\phi'(\tilde{t})|\, d\tilde{t}$$
$$\overset{t=\phi(\tilde{t})}{=} \int_I |\gamma'(t)|\, dt = L[\gamma].$$

Thus the total arclength is invariant under reparametrization. □

Exercise 1.3 Is the **Dirichlet energy** of a parametrized curve

$$E[\gamma] := \frac{1}{2}\int_I |\gamma'(t)|^2\, dt$$

a geometric quantity of the underlying unparametrized curve? ▪

1.2.1 Arclength Parametrization

Here is a special class of parametrization for curves.

Definition 1.3 A parametrized curve γ is said to be **parametrized by arclength** if $|\gamma'| = 1$, *i.e.* γ is moving at unit speed.

Given an arclength-parametrized curve $\gamma\colon [0, L] \to \mathbb{R}^2$, the arclength of an interval $[t_1, t_2] \subset [0, L]$ is

$$\int_{t_1}^{t_2} |\gamma'(t)|\, dt = \int_{t_1}^{t_2} 1\, dt = t_2 - t_1, \tag{1.1}$$

which is just the length of the interval $[t_1, t_2]$ in the parameter space $[0, L]$. Thus an arclength parametrization is also called an **isometric parametrization**.

Theorem 1.3 Every regular parametrized curve $\gamma\colon [a, b] \to \mathbb{R}^2$ can be reparametrized by its arclength.

Proof. Let $L = \int_a^b |\gamma'(t)|\, dt$ be the total arclength. Consider $\phi\colon [0, L] \to [a, b]$ given by

$$\phi(s) = a + \int_0^s \frac{1}{|\gamma'(t)|}\, dt, \quad s \in [0, L].$$

Then $\widetilde{\gamma}\colon [0, L] \to \mathbb{R}^2$, $\widetilde{\gamma}(s) := \gamma(\phi(s))$, is parametrized by arclength. Indeed, $|\widetilde{\gamma}'| = |\gamma'||\phi'| = |\gamma'| \cdot 1/|\gamma'| = 1$. □

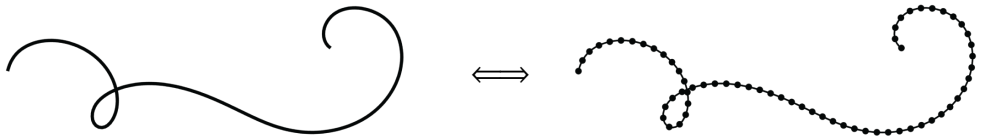


Figure 1.4 Each (oriented) unparametrized curve has a canonical representing parametrization, namely the arclength parametrization.

The readers can extend Theorem 1.3 to parametrized curves defined on unbounded intervals such as $[a, \infty)$ or $(-\infty, \infty)$.

Corollary 1.4 Every unparametrized regular curve has a unique (up to translation and reflection in the parameter) arclength parametrization.

Proof. The existence follows from Theorem 1.3. The uniqueness follows from the the rigidity of the length-preservation condition (1.1). □

In the earlier discussion we saw that finding a geometric quantity for an unparametrized curve is rather nontrivial. One would have to carefully check the expression for the quantity, written in terms of any parametrization, is invariant under reparametrization. Now, Corollary 1.4 gives a vast reduction of such trouble. What Corollary 1.4 suggests is that there *is* an essentially unique *canonical* parametrization, namely the arclength parametrization. Starting with any parametrized curve, one first "calibrates" the parametrization to the arclength one, before plugging the parametrization into a possibly parametrization-dependent expression. In other words:

R In the context of unparametrized regular curves, any expression written in terms of the arclength parametrization will automatically be geometric.

1.2.2 Tangent and Normal

The tangent vector defined in Proposition 1.1 is alternatively introduced in the following way. Given an oriented curve, we let $\gamma\colon [0, L] \to \mathbb{R}^2$ be its arclength parametric curve. That is we have $|\gamma'(s)| = 1$ for all $s \in [0, L]$.

Definition 1.4 The **tangent vector** of an oriented regular curve γ is given by

$$T(s) = \frac{d\gamma}{ds}(s) \tag{1.2}$$

where s is the arclength parameter for γ.

Eq. (1.2) reads "the tangent vector is the rate of change of the position per unit length."

In case one wants an expression for T in another parameter t, one replaces ds by $|{}^{d\gamma}/_{dt}|\, dt$ and obtains $T = {}^{1}/_{|{}^{d\gamma}/_{dt}|} \cdot {}^{d\gamma}/_{dt}$.

Definition 1.5 The **normal vector** N of an oriented regular curve is the 90° counterclockwise rotation of T. That is, if $T = (T_x, T_y)$, then $N = (-T_y, T_x)$.

Figure 1.5 Tangent vector and normal vector.

1.3 Curvature

Curvature measures how much the curve deviates from a straight line. If the tangent T and the normal N stay constant along the curve, then the curve does not change its direction and is therefore a straight line. This motivates us to define the curvature as *the rate of change of* T or N. Once again, to ensure the curvature is a geometric quantity for unparametrized (oriented) curve, the rate of change is measured per unit arclength.

Note that the derivative of the condition $|T|^2 = \langle T, T\rangle = 1$ yields

$$2\left\langle \frac{dT}{ds}, T\right\rangle = 0.$$

What it means is that the rate of change dT/ds of T is always orthogonal to T, *i.e.* dT/ds is a scalar multiple of N. We call this scalar multiple the curvature.

Definition 1.6 The **curvature** κ at each point along an oriented curve γ is defined by

$$\frac{dT}{ds} = \kappa N; \quad \text{equivalently,} \quad \kappa = \left\langle \frac{dT}{ds}, N\right\rangle. \tag{1.3}$$

Exercise 1.4 Show that

$$\frac{dN}{ds} = -\kappa T.$$

Hint *Differentiate* $\langle N, T\rangle = 0$. ▪

▪ **Example 1.2 — Curvature of a circle.** Consider a circle of radius R centered at the origin. Its arclength parametrization is given by

$$\gamma(s) = \begin{bmatrix} R\cos(s/R) \\ R\sin(s/R)\end{bmatrix}.$$

The tangent and normal vectors are

$$\gamma'(s) = T(s) = \begin{bmatrix} -\sin(t/R) \\ \cos(t/R)\end{bmatrix}, \quad N(s) = \begin{bmatrix} -\cos(t/R) \\ -\sin(t/R)\end{bmatrix}.$$

One sees that

$$T'(s) = \frac{1}{R}N(s).$$

Hence the circle has constant curvature $\kappa = 1/R$. ▪

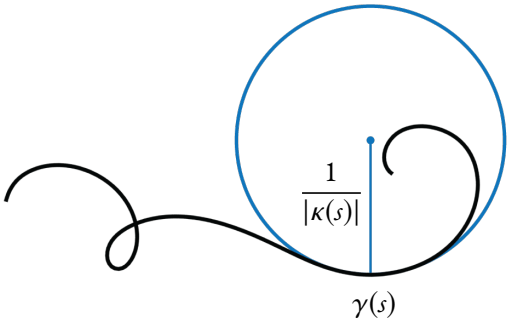


Figure 1.6 The osculating circle is the best approximating circle of a curve at a given point.

1.3.1 Osculating Circle

Due to Example 1.2, the reciprocal $1/|\kappa|$ of the curvature is known as the **radius of curvature**.

The **center of curvature**

$$\gamma(s) + \frac{1}{\kappa(s)} N(s)$$

is the center of the circle *best approximating* the curve at $\gamma(s)$. This best-approximating circle of the curve is known as the **osculating circle**.

R We can also use the notion of geometric continuity G^k to talk about the order of approximation. We say two plane curves are in G^0 contact if they intersect. We say they are in G^1 contact if they share the same tangent at the contact point. They are in G^2 contact if they share the same curvature. The osculating circle at $\gamma(s)$ is the unique circle that is in G^2 contact with γ at $\gamma(s)$.

R One may first introduce the notion of osculating circle, the unique circle that fits the curve to second order. Then define the curvature as the reciprocal of the circle radius. For oriented curve, the osculating circle should also be oriented clockwise or counterclockwise, so that the oriented circle is in oriented contact with the curve. The curvature is set with positive (resp. negative) sign if the osculating circle is counterclockwise (resp. clockwise). By this construction, the curvature is obviously geometric independent of the parametrization.

1.3.2 Turning Angle

The normal vector N is always the 90° counterclockwise rotation of the tangent vector T. It is convenient to identify the plane $\mathbb{R}^2$ as the complex plane $\mathbb{C}$. In that case, we view $\gamma\colon I \to \mathbb{C}$, $T\colon I \to \mathbb{C}$, $N\colon I \to \mathbb{C}$ as complex-valued functions. The relation between N and T becomes

$$N(s) = \mathbb{i}\, T(s). \tag{1.4}$$

Now let us look at $T\colon I \to \mathbb{C}$ as a parametrized curve in its own right. This parametrized curve traced out by γ' is called the **hodograph** of γ. Since T always has unit length, the hodograph lies in the unit circle, where the natural coordinate is the angle:

$$\theta\colon I \to \mathbb{R} \mod 2\pi, \quad T(s) = e^{\mathbb{i}\theta(s)}.$$

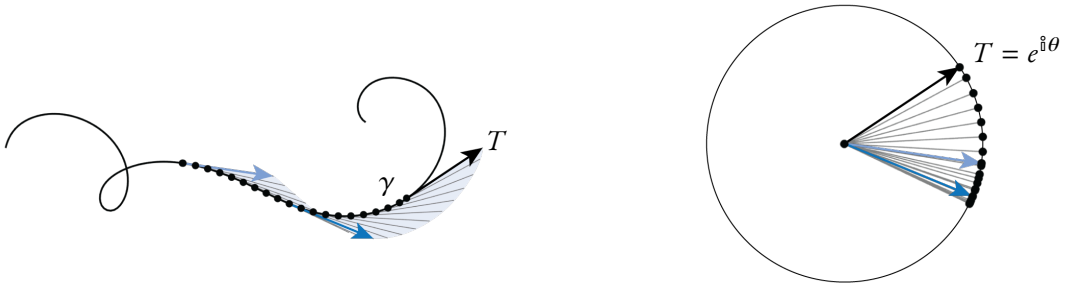


Figure 1.7 The tangent hodograph of a curve γ is the trajectory (right) traced out by the tangent vector T as a vector based at the origin. The tangent hodograph lies on the unit circle with a natural angular coordinate θ. The curvature κ of γ is the rate of change of θ per unit arclength of the original curve γ. In other words, the integral of $\kappa\, ds$ is the turning angle, *i.e.* the difference in θ.

The angle $\theta(s)$ is the angle $T(s)$ makes with the positive real axis. Now, Eq. (1.3) $dT/ds = \kappa N$ reads

$$\frac{d}{ds}e^{i\theta(s)} = i\frac{d\theta}{ds}e^{i\theta} \overset{(1.4)}{=} \frac{d\theta}{ds}N = \kappa N.$$

Therefore

$$\frac{d\theta}{ds} = \kappa. \tag{1.5}$$

A direct consequence is that the integral of κ is the total angle by which the direction T has turned

$$\Delta\theta = \int_{s_1}^{s_2} \kappa(s)\, ds. \tag{1.6}$$

In other words, we can write how the tangent vector has turned as

$$T(s_2) = e^{i\Delta\theta}T(s_1) = e^{i\int_{s_1}^{s_2}\kappa(s)\,ds}T(s_1). \tag{1.7}$$

R We can also write $\Delta\theta = \theta(s_2) - \theta(s_1) = \int_{s_1}^{s_2}\kappa(s)\,ds$ but only modulo 2π. Strictly speaking, there is a distinction between measuring the difference of the angle θ (defined only modulo 2π) and the turning angle (an actual real number). In expressions such as (1.7), the turning angle is used as a mod 2π quantity. One cannot distinguish whether a curve has turned $\Delta\theta$ or $\Delta\theta + 2\pi$ by measuring only the two end tangents $T(s_1), T(s_2)$. On the other hand, having the entire curve segment $\gamma(s)$, $s_1 \le s \le s_2$, available, one obtains the true turning angle $\int_{s_1}^{s_2}\kappa(s)\,ds$ which does distinguish $\Delta\theta$ from $\Delta\theta + 2\pi$, $\Delta\theta + 4\pi$, *etc.*

1.4 Integration along a Curve

The geometric quantities like the tangent vector, the normal vector and the curvature are local measurements; they are computed only using the geometry of the curve in a neighborhood of a

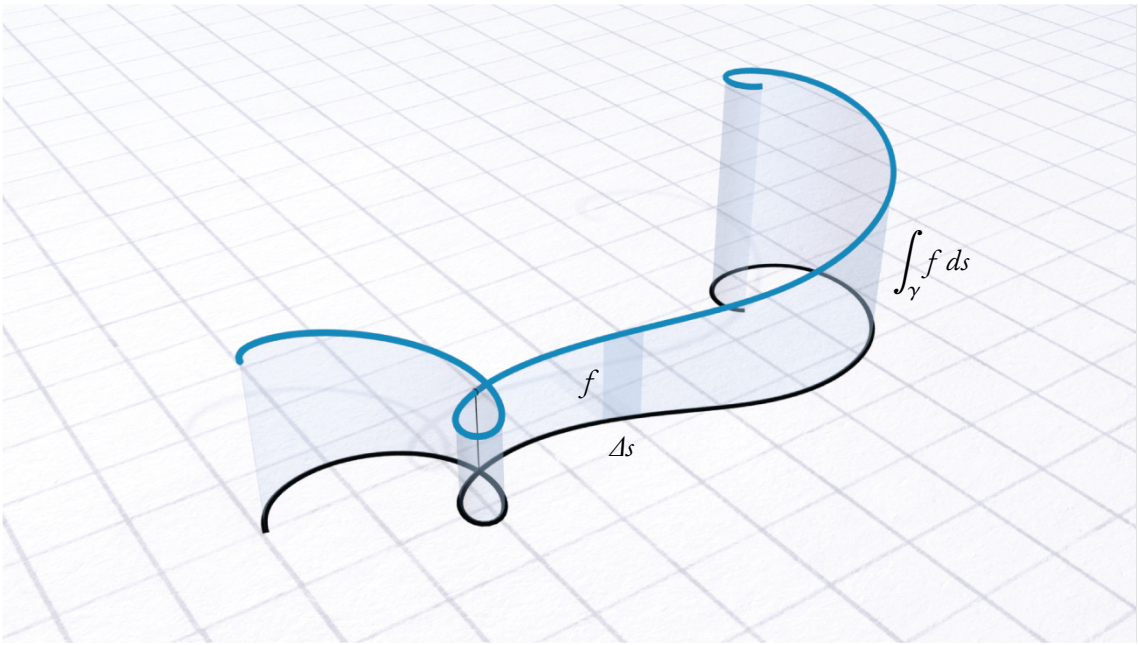


Figure 1.8 A function f defined over a curve γ. Its integral with respect to the arclength, denoted by $\int_\gamma f\,ds$, is the area beneath the function graph over the curve as shown in the figure. This is geometric independent of how the curve is parametrized.

point. Geometric quantities such as the total length $L[\gamma]$ are global measurements; it depends on the geometry of the entire curve. Similarly, the total turning angle (1.6) is a global geometric measurement. Global geometric measurements usually involve an integral that pieces together local geometric quantities. In this section we clarify the notion of integration along a curve.

In the previous examples, the geometric quantities $T(s)$, $N(s)$, $\kappa(s)$ all took the arclength s as the parametrization. That was because as we defined these quantities it was convenient to work with arclength. However, it is also fine to have a quantity f just described as a function on the curve, like shown in Figure 1.8; based on whichever parameter $t \in I$ the parametrization $\gamma(t)$ uses, f may be viewed as a function of $t \in I$.

Suppose $\gamma\colon I \to \mathbb{R}^2 = \mathbb{C}$ is a parametrized plane curve. The parameter $t \in I$ does not need to be arclength. Let $f\colon I \to \mathbb{R}$ be some quantity assigned at each point on the curve. Now we want to make sense of integrating the quantity f over the curve

$$\int_I f\,ds, \quad \text{also written as} \quad \int_\gamma f\,ds. \tag{1.8}$$

Note that the integrand $f(t)$ takes a different variable t than the arclength s that makes the infinitesimal ds.

Here are two ways to understand (1.8).

Stieltjes Integral

This is the standard formalism, first presented by Riemann in 1854 and then extended by Stiltjes in 1894. In the Riemann–Stieltjes integration theory, all integrals $\int_I (\cdot)\,d(\cdot)$ is understood first in the discrete setting as a summation $\sum(\cdot)\Delta(\cdot)$, and the integral is its limit under refinement of the discretization.

The arclength parameter s is viewed as an increasing function of the other parameter $s\colon I \to \mathbb{R}$, $s = s(t)$. The value $s(t)$, $t \in I = [a, b]$, is the total arclength from a to t. Now discretize the domain $[a, b]$ into some partition $a = t_0 < t_1 < \ldots < t_n = b$, and define $s_i = s(t_i)$ be the sampled arclength value. Then

$$\int_{t=a}^{b} f(t)\, ds(t) = \lim_{\substack{\text{discretization} \\ \text{refines}}} \sum_{i=1}^{n} f(t_i)\big(s_i - s_{i-1}\big).$$

In the case of arclength s, we know the explicit formula

$$s(t) = \int_0^t \left|\frac{d\gamma}{dt}(t)\right| dt, \quad \text{in other words,} \quad s'(t) = |\gamma'(t)|.$$

Then one shows that (1.8) can be written as an integration against dt

$$\int_{t=a}^{b} f(t)\, ds(t) = \int_{t=a}^{b} f(t)\ |\gamma'(t)|\ dt.$$

R Strictly speaking, in Riemann–Stieltjes' way to decode (1.8), an infinitesimal ds should always come with an integral sign $\int$. When $\int$ is paired with an infinitesimal, it begins to make sense as the limit of discrete summations. What comes next takes a different philosophy. The infinitesimal can be treated as a mathematical object in its own right even without an integral sign.

Differential Forms

Another viewpoint is that $\int_I f\, ds$ is the *integral* $\int$ of the *differential form* $f\, ds$ over the domain I. The elements ds and $f\, ds$ are **differential forms**, which are "to-be-integrated objects." These differential forms are mathematical objects and can be manipulated intuitively without an integral sign. The integral sign is just an *evaluation* operator that turns the differential forms into a definite number. One can go back to the Riemann–Stieltjes integral to formally define this evaluation process.

Such a viewpoint had been popular since the mid 18th century, until the early 20th century when such notations were discouraged in the mainstream calculus curricula. However, the theory of differential forms, and its full extension to multivariable setting called the exterior calculus (by Élie Cartan since 1899), is the necessary language one needs to speak in modern differential geometry.

For example, ds is called the **arclength differential form** defined over I, since for each interval $[t_1, t_2] \subset I$, its evaluation

$$\int_{t=t_1}^{t_2} ds(t) = s(t_2) - s(t_1)$$

measures the total arclength on the interval $[t_1, t_2]$.

A differential form ds multiplied by a function $f\colon I \to \mathbb{R}$ gives another differential form $f\, ds$. Its evaluation over I is (1.8), and it can be understood for example by the Riemann–Stieltjes machinery.

For a concrete example, the turning angle formula (1.6) motivates us to consider the differential form

$$\kappa\, ds.$$

Its evaluation over an interval $[t_1, t_2]$

$$\int_{t=t_1}^{t_2} \kappa(t)\, ds(t) \left(= \int_{t=t_1}^{t_2} \kappa(t)\ |\gamma'(t)|\ dt \right)$$

measures the turning angle $\theta(t_2) - \theta(t_1)$ of the tangent vector $T = e^{\mathbb{i}\theta}$. We call $\kappa(s)\, ds$ the **curvature form**, or the **turning form**.

We can discuss whether a differential form is a geometric object. Suppose $\gamma : I \to \mathbb{C}$ is some not-necessarily-arclength parametrized curve. If dt is the differential form that measures the displacement of the parameter, then the arclength measure is given by $ds = |d\gamma/dt|\, dt$. Changing dt by a reparametrization will leave ds invariant. If $f : I \to \mathbb{R}$ is a geometric quantity independent of the parametrization, then so is $f\, ds$. For instance, $\kappa\, ds$ is geometric.

R Given a regular curve $\gamma : I \to \mathbb{R}$, the domain I is endowed with an arclength differential form ds. Independent of what the parameter t that labels I, the differential form ds is a geometric quantity of the curve.

1.5 Topological Quantities

The geometric quantities like tangent vector, normal vector and curvature are local measurement; it is computed only using the geometry of the curve in a neighborhood of a point. Geometric quantities such as the total length $L[\gamma]$ are global measurements; it depends on the geometry of the entire curve.

A global geometric quantity becomes a **topological quantity** if it is invariant under smooth deformation of the geometry.

For example, the total length $L[\gamma]$ is *not* a topological quantity. We can perturb γ so that $L[\gamma]$ becomes bigger or smaller.

The main point we will focus on is that the **total curvature** of a *closed* curve

$$\text{TotalCurvature}[\gamma] = \int_{\mathbb{T}^1} \kappa\, ds$$

is a topological quantity.

1.5.1 Turning Number of a Closed Curve

Let $\gamma : \mathbb{T}^1 \to \mathbb{C}$ be a regular closed curve. Here $\mathbb{T}^1 = \mathbb{R} \mod 2\pi$, which means that the parametrized curve γ must be periodic $\gamma(t + 2\pi) = \gamma(t)$ for all $t \in \mathbb{R}$. This makes the curve a closed curve.

When we write $\int_{\mathbb{T}^1} f\, ds$ for a periodic function $f : \mathbb{T}^1 \to \mathbb{R}$, we mean the integral over one period $[t_0, t_0 + 2\pi]$ for any $t_0 \in \mathbb{R}$:

$$\int_{\mathbb{T}^1} f\, ds = \int_0^{2\pi} f(t) \left|\frac{d\gamma}{dt}\right| dt = \int_{t_0}^{t_0+2\pi} f(t) \left|\frac{d\gamma}{dt}\right| dt.$$

Any geometric measurement of a closed curve is automatically periodic. In particular, the tangent vector is periodic

$$T(t_0 + 2\pi) = T(t_0), \quad \text{for all } t_0 \in \mathbb{R}.$$

In terms of the turning form, we see

$$T(t_0 + 2\pi) = e^{\mathbb{i} \int_{\mathbb{T}^1} \kappa \, ds} T(t_0) = T(t_0).$$

Therefore,

$$e^{\mathbb{i} \int_{\mathbb{T}^1} \kappa \, ds} = 1.$$

In other words, we have:

Theorem 1.5 The total curvature of an oriented closed plane curve $\gamma : \mathbb{T}^1 \to \mathbb{C}$ is an integer multiple of 2π:

$$\int_{\mathbb{T}^1} \kappa \, ds = \int_0^{2\pi} \kappa(t) \left| \frac{d\gamma}{dt} \right| dt = 2\pi m \quad \text{for some } m \in \mathbb{Z}.$$

Definition 1.7 We call the integer m in Theorem 1.5 the **turning number** of the given oriented closed plane curve.

The turning number is also the **winding number** of the hodograph. It is the number of times $\gamma'(t)$ or $T(t)$ winds around the origin over one period $t \in [t_0, t_0 + 2\pi]$.

1.5.2 Regular Homotopy Type

The total curvature $\int_{\mathbb{T}^1} \kappa \, ds$ varies continuously under a continuous variation of the function κ. On the other hand, $\int_{\mathbb{T}^1} \kappa \, ds$ takes only quantized values (2π multiple of integers). Therefore, m must stay constant when γ is *regularly* deformed. By regular deformation we mean the type of deformation of γ so that $\kappa \, ds$ varies continuously. Such deformation is called a **regular homotopy**.

If we are allowed to deform the curve so that the curve does not stay regular (*e.g.* a cusp is developed), then $\int_{\mathbb{T}^1} \kappa \, ds$ does not need to stay constant. This shows the importance of being regular as we discuss the constancy of the total curvature.

The turning number m is invariant under regular deformation of the geometry. Therefore the turning number is a **topological quantity** in the context of regular oriented closed curves.

Definition 1.8 We say two regular oriented closed curves are of the same **regular homotopy type** if one can be deformed to the other and stay regular throughout the deformation.

Our earlier discussion can be concluded as:

Theorem 1.6 Any two regular oriented closed curves with the same regular homotopy type must have the same turning number.

In fact, the converse is also true:[1]

[1] A proof of the Whitney–Graustein Theorem is explained in detail in a 1974 computer animation "Regular Homotopies in the Plane, Part II" (`https://youtu.be/mY-V0TSMVCY`)

Theorem 1.7 — Whitney–Graustein 1937. Any two regular oriented closed curves having the same turning number must also be of the same regular homotopy type. That is, they admit a regular deformation that deforms one to the other.

R In other words, the turning number *classifies* the regular homotopy types of regular oriented closed curves.

1.6 Variational Approach to Curvature

We have seen that the curvature is defined either by

- the rate of change of tangent vector per unit arclength.
- the reciprocal of the osculating circle radius.

Here we introduce yet another approach to curvature based on **variational principle**. The curvature is

- the tension force that straightens the curve when the potential energy of a curve is the total length.

Such a modeling approach makes sense as soon as we recognize "the curvature measures how much the curve deviates from being straight" and "a straight line is the shortest path connecting two fixed ends."

1.6.1 Gradient of a Functional

Let us recall the **directional derivatives** in multivariable calculus. Suppose $f = f(x_1, \dots, x_n)$ is a function of several variables. Take a point $(x_1, \dots, x_n)$ and extend it to a one-parameter (ε) family of points

$$\big(\widetilde{x}_1(\varepsilon), \dots, \widetilde{x}_n(\varepsilon)\big), \quad \text{with } (x_1, \dots, x_n) = \big(\widetilde{x}_1(0), \dots, \widetilde{x}_n(0)\big).$$

The infinitesimal perturbation made by varying ε about $\varepsilon = 0$ gives rise to a vector

$$(\mathring{x}_1, \dots, \mathring{x}_n) := \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \big(\widetilde{x}_1(\varepsilon), \dots \widetilde{x}_n(\varepsilon)\big)\,.$$

Plug the ε-dependent variables into the function f of interest, and study the rate of change of the resulting value $f(\widetilde{x}_1(\varepsilon), \dots, \widetilde{x}_n(\varepsilon))$ when we perturb ε. This rate of change

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} f\left(\widetilde{x}_1(\varepsilon), \dots, \widetilde{x}_n(\varepsilon)\right)$$

is the *directional derivative of f in the direction* $(\mathring{x}_1, \dots, \mathring{x}_n)$. One then discovers that, provided that f is differentiable, the directional derivative always depends linearly in $\mathring{x}_i$'s:

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} f\left(\widetilde{x}_1(\varepsilon), \dots, \widetilde{x}_n(\varepsilon)\right) = \sum_{i=1}^{n} G_i \mathring{x}_i.$$

The coefficients G_i's form a vector $G = (G_1, \dots, G_n)$ called the **gradient** of f.

R The gradient vector G is also characterized as the direction in which f changes most rapidly per unit length. The magnitude of G is the slope of f in that direction.

1.6.2 Variation of the Length

The story is the same for a function, such as the total arclength $L[\gamma]$, that takes an entire curve as the argument, which has infinite many degrees of freedom. The variables x_i, indexed by $i = 1, \ldots, n$, in the multivariable calculus setting is now replaced by a parametrized curve $\gamma(t)$ indexed by $t \in I = [a, b]$.

Take a curve $\gamma\colon I \to \mathbb{R}^2$, say arclength parametrized. Extend it to a one-parameter (ε) family of parametrized curves

$$\widetilde{\gamma}(\cdot;\varepsilon)\colon I \to \mathbb{R}^2, \quad \text{with } \gamma(t) = \widetilde{\gamma}(t;0) \text{ for all } t \in I.$$

Note that we only assume that $\gamma(\cdot) = \widetilde{\gamma}(\cdot;0)$ is parametrized by arclength, but $\widetilde{\gamma}(\cdot;\varepsilon)$ are not necessarily parametrized by arclength. Now, the infinitesimal perturbation made by moving ε about $\varepsilon = 0$ gives rise to a vector field

$$\mathring{\gamma}\colon I \to \mathbb{R}^2, \quad \mathring{\gamma}(t) \coloneqq \left.\frac{\partial}{\partial\varepsilon}\right|_{\varepsilon=0} \widetilde{\gamma}(t;\varepsilon).$$

Plug the ε-dependent varying curve into the length functional

$$L[\widetilde{\gamma}(\cdot;\varepsilon)] = \int_a^b \left|\frac{\partial}{\partial t}\widetilde{\gamma}(t;\varepsilon)\right| dt,$$

and study its rate of change when we perturb ε. This rate of change

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} L[\widetilde{\gamma}(\cdot;\varepsilon)]$$

is the *directional derivative of L in the direction $\mathring{\gamma}$*. We expect the directional derivative to be written in an expression that is linear in $\mathring{\gamma}$, and the coefficient in the linear form will be the gradient of the functional. The calculation goes as

$$\begin{aligned}
\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} L[\widetilde{\gamma}(\cdot;\varepsilon)] &= \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_a^b \left|\frac{\partial}{\partial t}\widetilde{\gamma}(t;\varepsilon)\right| dt \\
&= \int_a^b \left.\frac{\partial}{\partial\varepsilon}\right|_{\varepsilon=0} \sqrt{\left\langle \frac{\partial}{\partial t}\widetilde{\gamma}(t;\varepsilon), \frac{\partial}{\partial t}\widetilde{\gamma}(t;\varepsilon)\right\rangle}\, dt \\
&= \int_a^b \left.\frac{\langle \frac{\partial}{\partial t}\widetilde{\gamma}, \frac{\partial^2}{\partial t\partial\varepsilon}\widetilde{\gamma}\rangle}{\sqrt{\langle \frac{\partial}{\partial t}\widetilde{\gamma}, \frac{\partial}{\partial t}\widetilde{\gamma}\rangle}}\right|_{\varepsilon=0} dt \\
&= \int_a^b \left\langle \frac{d\gamma}{dt}, \frac{d\mathring{\gamma}}{dt}\right\rangle dt = \int_a^b \langle T, \mathring{\gamma}'\rangle \, dt
\end{aligned}$$

where the last line evaluates $\widetilde{\gamma}$ at ε, and the denominator is 1 since γ is arclength parametrized. Taking integration by parts of the last expression yields

$$\begin{aligned}
\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} L[\widetilde{\gamma}(\cdot;\varepsilon)] &= \int_a^b \langle T, \mathring{\gamma}'\rangle \, dt \\
&= -\langle T(a), \mathring{\gamma}(a)\rangle + \langle T(b), \mathring{\gamma}(b)\rangle - \int_a^b \langle T', \mathring{\gamma}\rangle \, dt. \qquad (1.9)
\end{aligned}$$

Finally, this is an expression that is written in a simple linear dependency on $\mathring{\gamma}$. At this moment after setting $\varepsilon = 0$, we write t as s to emphasize it is an arclength parametrization. We conclude from (1.9) that the differential form

$$-T'\,ds, \quad s \in (a,b),$$

is the **functional gradient** of the functional L. At the exceptional points, *i.e.* the boundary points, the functional gradients are

$$-T(a) \text{ at } s = a, \text{ and } T(b) \text{ at } s = b.$$

In Mechanics

In the context of mechanics, if the functional L is the potential energy of the curve, then its negative gradient is the induced force. As the negative gradient, it is also the direction the geometry tends to move that most rapidly minimizes the potential energy.

Specifically, the force associated with total length L is called the **tension**. On the boundary points we do see a constant traction force

$$T(a) \text{ at } s = a, \text{ and } -T(b) \text{ at } s = b.$$

If the curve is not straight, then the tension also produces a net force

$$T'\,ds, \quad s \in (a,b)$$

over the curve. Indeed, this tension force measures the curvature of the curve

$$T'\,ds = \kappa N\,ds.$$

The curvature of the curve arises as the magnitude of the tendency that the curve minimizes its length.

2. Discrete Plane Curves

Discrete curves are representations of curves that only have finitely many degrees of freedom. Examples include polygonal curves, various versions of spline curves and subdivision curves, each of which only have finitely many control points. Here, we will only study polygonal curves.

For discrete curves, we aim to define geometric quantities that are analogous to those for smooth curves. We also desire to see similar theorems in the differential geometry of curves that hold for the discrete quantities for discrete curves.

As we have seen in Chapter 1 most of the interesting geometric quantities involve derivatives. One way to approach these derivatives is to apply discretization schemes, such as the finite difference method. However, the best one can hope for that comes out of these general discretization schemes are approximation. Based on approximation theory, an exact discrete analog of a geometric theorem is rare.

In contrast, a discrete differential geometric approach takes the discrete geometry as a geometric object in its own right. By taking the smooth theory as a reference, we find the corresponding geometric notions that compose a geometric theory that mimics the smooth counterpart.

In short, discretize the theory, not the equations.

2.1 Polygonal Curves

Instead of having the domain I of a parametrized curve be an interval, let the domain be an ordered index set

$$I = (0, 1, \ldots, n-1). \tag{2.1}$$

Here $n \in \mathbb{N}$ is the number of points.

Definition 2.1 A **discrete plane curve**, or more precisely a **polygonal curve** is a map $\gamma\colon I \to \mathbb{R}^2 = \mathbb{C}$ where I is the ordered index set (2.1).

The image

$$\gamma(I) = (\gamma_0, \gamma_1, \dots, \gamma_{n-1})$$

is an ordered sequence of point positions in the plane.

Let E be a collection of ordered paris of points representing the edges

$$E = \begin{cases} \{(0,1),(1,2),\cdots,(n-2,n-1)\} & \text{non-closed curve;} \\ \{(0,1),(1,2),\cdots,(n-2,n-1),(n-1,0)\} & \text{closed curve.} \end{cases} \tag{2.2}$$

For each pair of neighboring points $(i, i+1) \in E$, we have the **edge vector**

$$v\colon E \to \mathbb{R}^2, \quad v_{i,i+1} \coloneqq \gamma_{i+1} - \gamma_i \in \mathbb{R}^2 = \mathbb{C}. \tag{2.3}$$

If $v_{i,i+1} \neq 0$, then we define the **edge tangent vector**

$$T_{i,i+1} \coloneqq \frac{v_{i,i+1}}{|v_{i,i+1}|} = \frac{\gamma_{i+1} - \gamma_i}{|\gamma_{i+1} - \gamma_i|} \in \mathbb{R}^2 = \mathbb{C}. \tag{2.4}$$

The **edge normal** is the 90° counterclockwise rotation of the edge tangent

$$N_{i,i+1} \coloneqq \mathbb{i}\, T_{i,i+1} \tag{2.5}$$

Let us call the **edge length**

$$\ell\colon E \to \mathbb{R}_{\geq 0}, \quad \ell_{i,i+1} = |v_{i,i+1}| = |\gamma_{i+1} - \gamma_i|\,. \tag{2.6}$$

The total length of the curve is given by

$$L[\gamma] = \sum_{(i,i+1)\in E} \ell_{i,i+1}. \tag{2.7}$$

Definition 2.2 A polygonal curve is called **regular** if $\ell_{i,i+1} \neq 0$ for all edges, and

$$T_{i-1,i} \neq -T_{i,i+1} \tag{2.8}$$

for all $i = 1, \dots, n-2$ for non-closed curve, and for all $i = 0, \dots, n-1$ (in the mod n arithmetics) for closed curve.[a]

[a] There are other discrete theories where regular curves are defined by the condition that any three successive points are pairwise disjoint; *i.e.* $v_{i-1,i} \neq -v_{i,i+1}$.

That is, a regular curve never has two successive edges folding back on each other.

Definition 2.3 A polygonal curve is called **arclength parametrized** if

$$\ell_{0,1} = \ell_{1,2} = \cdots = \ell_{n-2,n-1} (= \ell_{n-1,0}) \neq 0.$$

R Unlike the smooth setting where all regular curves can be reparametrized by arclength, there is no obvious way to reparametrize discrete curves. That is, discrete curves are parametrized curves; parametrization is part of the geometry.

2.2 Discrete Curvatures

What is the notion of curvature for a polygonal curve? In fact there are several different discrete curvatures depending on which characterization of curvature we replicate from the smooth theory.

2.2.1 Exterior Angles as the Discrete Curvature Form

An approach we have in the smooth setting is that the curvature describes the rate of change of the tangent (or normal). That is, $dT/ds = \kappa N$ and $dN/ds = -\kappa T$. This idea can be translated to the discrete setting.

Each edge tangent vector $T_{i,i+1} = e^{\mathbb{i}\theta_{i,i+1}}$ defines an angle $\theta_{i,i+1} \in \mathbb{T}^1 = \mathbb{R}/2\pi\mathbb{Z}$ (the angle T makes with the positive real axis). The difference of successive angles

$$\alpha_i := \theta_{i,i+1} - \theta_{i-1,i} \mod 2\pi, \quad \text{taking the branch } \alpha_i \in (-\pi, \pi], \tag{2.9}$$

is our discrete curvature form:

$$(\kappa\, ds)_i = \alpha_i. \tag{2.10}$$

Geometrically, this is the exterior angle at each vertex of the polygon. In terms of $T_{i,i+1} = e^{\mathbb{i}\theta_{i,i+1}}$,

$$T_{i,i+1} = e^{\mathbb{i}\alpha_i} T_{i-1,i}. \tag{2.11}$$

Note that for regular curves $\alpha_i \neq \pi$.

It is more proper to call the exterior angle the discrete analog of the *curvature form* rather than the *curvature*. The curvature form measures the total angle T turns, whereas the curvature is the curvature form divided by arclength form. One also interprets a vertex of a polygon as the limit of a smooth curve where the curvature concentrates into an impulse, but the total contribution in the turning angle $\int \kappa\, ds$ across the vertex converges to the exterior angle of the vertex.

Turning Number

In the smooth theory, the total curvature $\int_{\mathbb{T}^1} \kappa\, ds$ of a closed curve is an integer multiple of 2π. Here, the total sum of the discrete curvature forms is also an integer multiple of 2π.

Theorem 2.1 The sum of the exterior angles of a closed polygon

$$\sum_{i=0}^{n-1} \alpha_i = 2\pi m, \quad \text{for some } m \in \mathbb{Z}. \tag{2.12}$$

The integer m is the **turning number** of the closed polygon.

Proof. By (2.11) we have

$$\begin{aligned} T_{n-1,0} &= e^{\mathbb{i}\alpha_{n-1}} T_{n-2,n-1} = e^{\mathbb{i}\alpha_{n-1}} e^{\mathbb{i}\alpha_{n-2}} T_{n-3,n-2} = \cdots = e^{\mathbb{i}\alpha_{n-1}} e^{\mathbb{i}\alpha_{n-2}} \cdots e^{\mathbb{i}\alpha_1} T_{0,1} \\ &= e^{\mathbb{i}\alpha_{n-1}} e^{\mathbb{i}\alpha_{n-2}} \cdots e^{\mathbb{i}\alpha_1} e^{\mathbb{i}\alpha_0} T_{n-1,0}. \end{aligned}$$

Therefore $e^{\mathbb{i}\alpha_{n-1}} \cdots e^{\mathbb{i}\alpha_0} = e^{\mathbb{i}\sum_{i=0}^{n-1}\alpha_i} = 1$, which is equivalent to $\sum_{i=0}^{n-1}\alpha_i = 0 \mod 2\pi$; *i.e.* $\sum_{i=0}^{n-1}\alpha_i$ is an integer multiple of 2π. □

Theorem 2.2 The turning number of a discrete closed curve is invariant under a regular homotopy, *i.e.* a continuous deformation during which the curve stays regular (*cf.* Definition 2.2).

Proof. Under a regular homotopy, the exterior angles change continuously and stay within $(-\pi, \pi)$. Therefore, $\sum_{i=0}^{n-1} \alpha_i$ changes continuously. By Theorem 2.1, the value $\sum_{i=0}^{n-1} \alpha_i$ is quantized as integer multiples of 2π. Hence $\sum_{i=0}^{n-1} \alpha_i$ must stay constant. □

Note that the continuity argument in the proof will fail if the curve does not stay regular under the deformation. When an edge folds back onto a neighboring edge, the exterior angle jumps by 2π due to the choice of the branch $\alpha \in (-\pi, \pi]$.

Real Angles v.s. Periodic Angles

In Section 1.3.2, we made a remark that the turning angle $\int \kappa \, ds$ contains more information than just comparing the angles between the tangents at the ends of the segments. The latter is a periodic quantity modulo 2π, contrasting the former which is a real number that distinguishes different periods.

We may bring this idea to the discrete setting. Given a discrete curve, we can only infer the exterior angle up to 2π integers. Alternatively, we describe the discrete curve not only by list of vertex positions $(\gamma_0, \ldots, \gamma_{n-1})$, but also a list of *real numbers* $(\alpha_0, \ldots, \alpha_{n-1})$, annotating the "extended" exterior angles that are not limited to $(-\pi, \pi)$, subject to one condition that $e^{\mathbb{i}\alpha_i} T_{i-1,i} = T_{i,i+1}$ for each i.

In this way, the discrete curvature form represents the true turning angle from one edge to the next. If $|\alpha_i| > \pi$, it means the curve actually makes one or many loops within that vertex that are not visible in the current discrete resolution (which can be a useful information for subdivision). With such a representation of discrete curve with real curvature form, we do not need the deformation to be regular in the sense of Definition 2.2 to obtain Theorem 2.2; we just need to keep track of α_i so that it changes continuously as a real number.

2.2.2 Osculating Circles for Discrete Curves

Another characterization of the curvature is the reciprocal of the radius of the osculating circle.

Vertex Osculating Circle

In the smooth theory, every point on a regular curve admits a unique circle that is in G^2 contact with the curve at the given point. A discrete analog of such an existence and uniqueness property is that every 3 disjoint points $\gamma_{i-1}, \gamma_i, \gamma_{i+1}$ admit a unique circle C_i^{cc} that passes through all three points. This circle is the circumcircle (hence the label "cc") of the triangle $\gamma_{i-1}, \gamma_i, \gamma_{i+1}$.

Recall from elementary geometry:

Proposition 2.3 *The center of C_i^{cc} is the intersection of the orthogonal bisector lines of $\gamma_{i-1}\gamma_i$ and $\gamma_i\gamma_{i+1}$.*

Definition 2.4 Let the circumcircle-based discrete curvature be $\kappa_i^{\mathrm{cc}} = 1/R_i^{\mathrm{cc}}$ where R_i^{cc} is the radius of the circumcircle of the triangle $\gamma_{i-1}, \gamma_i, \gamma_{i+1}$.

Once again, let us have α_i denote the exterior angle.

Theorem 2.4 The circumcircle-based discrete curvature is given by

$$\kappa_i^{\mathrm{cc}} = \frac{2\sin\alpha_i}{|\gamma_{i+1} - \gamma_{i-1}|} = \frac{2\sin\alpha_i}{\sqrt{\ell_{i-1,i}^2 + \ell_{i,i+1}^2 + 2\ell_{i-1,i}\ell_{i,i+1}\cos\alpha_i}}. \tag{2.13}$$

If the discrete curve is arclength parametrized with constant edge length ℓ, then (2.13) simplifies to

$$\kappa_i^{\mathrm{cc}} = \frac{2}{\ell}\sin\left(\frac{\alpha_i}{2}\right). \tag{2.14}$$

Proof. Use the law of sine

$$\frac{\kappa_i^{\mathrm{cc}}}{2} = \frac{1}{2R_i^{\mathrm{cc}}} = \frac{\sin\left(\gamma_i \angle^{\gamma_{i-1}}_{\gamma_{i+1}}\right)}{\overline{\gamma_{i-1}\gamma_{i+1}}}$$

and the law of cosine

$$\overline{\gamma_{i-1}\gamma_{i+1}}^2 = \ell_{i-1,i}^2 + \ell_{i,i+1}^2 - 2\ell_{i-1,i}\ell_{i,i+1}\cos\left(\gamma_i \angle^{\gamma_{i-1}}_{\gamma_{i+1}}\right).$$

□

Note that κ_i^{cc} is rather unstable when α_i is large. When α_i approaches π, we have according to (2.13) $\kappa_i^{\mathrm{cc}} \to 0$ as long as $\ell_{i-1,i} \neq \ell_{i,i+1}$, whereas according to (2.14) $\kappa_i^{\mathrm{cc}} \to 2/\ell$ when $\ell_{i-1,i} = \ell_{i,i+1} = \ell$.

What one may truly want for a discrete theory is to expect $\kappa_i \to \infty$ if α_i approaches π since the curve is approaching an irregular curve. This version of a discrete osculating circle addresses these concerns.

Circle Osculating Three Edges

There is a dual version of the vertex osculating circle. For each edge $(i, i+1)$ there are circles that are tangent to three lines extended from the successive edges $\overleftrightarrow{\gamma_{i-1}\gamma_i}$, $\overleftrightarrow{\gamma_i\gamma_{i+1}}$ and $\overleftrightarrow{\gamma_{i+1}\gamma_{i+2}}$. These circles are the inscribed circle by the three edges. In fact there are 4 such circles in general.

Among these 4 circles, exactly one of them becomes an oriented circle that is in oriented contact with the oriented lines $\overrightarrow{\gamma_{i-1}\gamma_i}$, $\overrightarrow{\gamma_i\gamma_{i+1}}$ and $\overrightarrow{\gamma_{i+1}\gamma_{i+2}}$. This circle $C_{i,i+1}^{\mathrm{ic}}$ is our three-edge osculating circle.

Proposition 2.5 *The center of $C_{i,i+1}^{\mathrm{ic}}$ is the intersection of the angle bisectors of $\gamma_i \angle^{\gamma_{i-1}}_{\gamma_{i+1}}$ and $\gamma_{i+1} \angle^{\gamma_i}_{\gamma_{i+2}}$.*

Definition 2.5 The edge-based discrete curvature $\kappa_{i,i+1}^{\mathrm{ic}}$ is defined per edge by the reciprocal of the radius of $C_{i,i+1}^{\mathrm{ic}}$.

Theorem 2.6 The edge-based discrete curvature is given by

$$\kappa_{i,i+1}^{\mathrm{ic}} = \frac{\tan(\alpha_i/2) + \tan(\alpha_{i+1}/2)}{\ell_{i,i+1}}. \tag{2.15}$$

Exercise 2.1 Show Theorem 2.6. ▪

Note that any of α_i and α_{i+1} approaches π will lead to $\kappa_{i,i+1}^{\mathrm{ic}} \to \infty$.

Chain of Circles

The discrete model of chain of circles is a variant of the previous discrete osculating circles tangent to edges.

Suppose for now the discrete curve is arclength parametrized, *i.e.* it has constant edge lengths. Then at each i there is a unique circle C_i that is tangent to the midpoint of $\gamma_{i-1}\gamma_i$ and the midpoint of $\gamma_i\gamma_{i+1}$. We call it the edge osculating circle. By drawing only these circular arcs, the polygonal curve is the control polygon that represents a piecewise circular G^1 smooth curve, which has a well-defined piecewise constant curvature.

Definition 2.6 Let the discrete curvature based on edge osculating circle κ_i^{ec} be the reciprocal of the radius of the edge osculating circle C_i.

Theorem 2.7 For an arclength parametrized discrete curve, the edge osculating circle C_i has curvature

$$\kappa_i^{\text{ec}} = \frac{2}{\ell} \tan\left(\frac{\alpha_i}{2}\right). \tag{2.16}$$

Note that as $\alpha_i \to \pi$, $\kappa_i^{\text{ec}} \to \infty$.

Chain of Circles for Non-arclength Parametrized Curves

Suppose the discrete curve is not arclength parametrized. Then there are still piecewise circular curves that are tangent to some point on each edge. In fact there are one-parameter family many of these piecewise circular curves.

Suppose a circular arc is tangent to the edge $\gamma_{i-1}\gamma_i$ at $\mu_{i-1,i}$ and to the edge $\gamma_i\gamma_{i+1}$ at $\mu_{i,i+1}$. Then, due to the property of an inscribed circle, we must have $\overline{\mu_{i-1,i}\gamma_i} = \overline{\gamma_i\mu_{i,i+1}}$. We call this common segment length *half of the arclength at the vertex*

$$\ell_i := 2\overline{\mu_{i-1,i}\gamma_i} = 2\overline{\gamma_i\mu_{i,i+1}}.$$

It is the total length of the polygonal curve from $\mu_{i-1,i}$ to $\mu_{i,i+1}$ (traversing along the polygon, not the circular arc). In terms of these vertex arclengths, we immediately have the curvature edge osculating circle by the same calculation as Theorem 2.7:

$$\kappa_i^{\text{ec}} = \frac{2}{\ell_i} \tan\left(\frac{\alpha_i}{2}\right). \tag{2.17}$$

The arclengths ℓ_i at the vertices must satisfy the following condition per edge

$$\frac{\ell_i + \ell_{i+1}}{2} = \ell_{i,i+1}, \quad i = 0, \ldots, n-2 \tag{2.18}$$

There are one-parameter family of solutions to the difference equation (2.18). Choosing ℓ_0 arbitrarily, then $\ell_1 = \ell_{0,1} - \ell_0$ is determined, and similarly $\ell_2 = \ell_{1,2} - \ell_1 = \ell_{1,2} - \ell_{0,1} + \ell_0$ is determined, *etc.*

$$\ell_i = (-1)^i \ell_0 - \sum_{j=0}^{i-1} (-1)^{i-j} \ell_{j,j+1}. \tag{2.19}$$

If γ is a closed curve, then there is an extra condition on the final edge, making (2.18) a linear system. For closed curve, one solves

$$\frac{1}{2}\begin{bmatrix} 1 & 1 & & & & \\ & 1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & & 1 & 1 \\ 1 & & & & & 1 \end{bmatrix}\begin{bmatrix} \ell_0 \\ \ell_1 \\ \vdots \\ \ell_{n-2} \\ \ell_{n-1} \end{bmatrix} = \begin{bmatrix} \ell_{0,1} \\ \ell_{1,2} \\ \vdots \\ \ell_{n-2,n-1} \\ \ell_{n-1,0} \end{bmatrix}. \tag{2.20}$$

The linear system has a unique solution when n is odd, and has one parameter family of solutions when n is even.

Another natural way to look at such a discrete curve is the following. Visualize the vertex length ℓ_i as a circle of radius ℓ_i centered at γ_i. Then the discrete curve is decorated with touching circles centered at each vertex.

In fact one may start with a chain of touching circles of various radii ℓ_i as the primary object. That saves us from solving the inverse problem (2.18). The represented discrete curve is either the polygon connecting the centers of the circles, or the piecewise circular arcs that connect the touching points $\mu_{i-1,i}$ and $\mu_{i,i+1}$ of the chain of circles, and are orthogonal to the touching circles.

In this picture of chains of circles, the curvature κ_i^{ec} is also the magnitude of the image of $\mu_{i,i+1}$ via the stereographic projection from $\mu_{i-1,i}$.

2.2.3 Discrete Curvature from Variational Approach

Another approach to model curvature is by taking the variation of the total length $L[\gamma] = \sum_{i=0}^{n-2} \ell_{i,i+1}$.

Note that the gradient is the direction in which a perturbation of the curve makes the most rapid increment of the functional. In $L[\gamma] = \sum_{i=0}^{n-2} \ell_{i,i+1}$, moving γ_i only affects two terms $\ell_{i-1,i}$ and $\ell_{i,i+1}$. The gradient of $\ell_{\ell_{i-1,i}}$ with respect to perturbing γ_i is $T_{i-1,i}$, as it is the direction that lengthens $\ell_{i-1,i}$ the most, and the rate of change is $1 = |T_{i-1,i}|$. Similarly, the gradient of $\ell_{i,i+1}$ with respect to ℓ_i is $-T_{i,i+1}$. Combining them we obtain:

Theorem 2.8 The gradient of L with respect to γ_i is

$$\frac{\partial L}{\partial \gamma_i} = T_{i-1,i} - T_{i,i+1}.$$

The magnitude is given by

$$\left|T_{i-1,i} - T_{i,i+1}\right| = \left|2\sin\left(\frac{\alpha_i}{2}\right)\right| \tag{2.21}$$

which gives us a variation-based discrete curvature form

$$(\kappa\, ds)_i^{\text{var}} := 2\sin\left(\frac{\alpha_i}{2}\right). \tag{2.22}$$

R For arclength parametrized discrete curve, the variation-based discrete curvature agrees with the circumcircle-based discrete curvature.

2.2.4 Summary of Discrete Curvatures

We introduced several different discrete curvatures. As discrete curvature forms on an arclength parametrized discrete curve,

$$(\kappa\, ds)_i = \alpha_i, \quad \kappa_i^{\mathrm{cc}}\ell = 2\sin\left(\frac{\alpha_i}{2}\right), \quad \kappa_{i,i+1}^{\mathrm{ic}}\ell = \tan\left(\frac{\alpha_i}{2}\right) + \tan\left(\frac{\alpha_{i+1}}{2}\right),$$
$$\kappa_i^{\mathrm{ec}}\ell = 2\tan\left(\frac{\alpha_i}{2}\right), \quad (\kappa\, ds)_i^{\mathrm{var}} = 2\sin\left(\frac{\alpha_i}{2}\right).$$

Each pair of these different notions of curvature forms approximate each other in the sense that, for example,

$$\frac{\kappa_i^{\mathrm{cc}}\ell}{\kappa_i^{\mathrm{ic}}\ell} \to 1 \quad \text{as } \alpha_i \to 0 \text{ and } \frac{\alpha_{i+1} - \alpha_i}{\alpha_i} \to 0.$$

From the point of view of numerical error analysis, all of these discrete curvatures are some (low order) approximations of the equation $^{dT}/_{ds} = \kappa N$. The point here is that by understanding the different geometric characterizations of the curvature, one reveals and distinguishes different notions of discrete curvatures, each of which fits some discrete analog of the curve theory. Discovering how a theory holds in the discrete setting provides insights to the nature of these theorems in the original smooth curve theory. In practice, a discrete curvature can have a greater advantage over another depending on the application, even though they have a similar order of accuracy.

3. Space Curves

In this chapter we explore curves in 3D. Similar to the plane curves, given a parameter space $I \subset \mathbb{R}$ a **parametrized space curve** is a continuous function $\gamma\colon I \to \mathbb{R}^3$. A G^k **regular space curve** has γ C^k-differentiable and $\gamma' \neq 0$. An unparametrized space curve is the equivalence class of parametrized space curves modulo reparametrizations.

▪ **Example 3.1** The intersection curve of a cylinder $\{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ and a surface $\{(x,y,z) \in \mathbb{R}^3 \mid z = x^2\}$ can be parameterized by a smooth (C^∞) function

$$\gamma(t) = \begin{bmatrix} \cos(t) \\ \sin(t) \\ \frac{1+\cos(2t)}{2} \end{bmatrix}, \quad t \in [0, 2\pi),$$

with nowhere vanishing

$$|\gamma'(t)| = \sqrt{\sin^2(t) + \cos^2(t) + \sin^2(2t)}.$$

Therefore this intersection curve is a G^∞ regular curve. ▪

Each space curve can be parametrized by arclength by the same argument as for the plane curves. In terms of differential form, I is endowed with an arclength measure ds.

▪ **Example 3.2** The following family of space curves are parameterized by arclength.

$$\gamma(t) = \frac{1}{r + r^{2k-1}} \begin{bmatrix} r\cos(t) + \frac{1}{2k-1} r^{2k-1} \cos((2k-1)t) \\ r\sin(t) - \frac{1}{2k-1} r^{2k-1} \sin((2k-1)t) \\ \frac{2}{k} r^k \sin(kt) \end{bmatrix}$$

where $r > 0$ and $k \in \mathbb{N}$. ▪

Definition 3.1 The tangent vector of an oriented regular space curve is given by

$$T = \frac{\gamma'}{|\gamma'|} = \frac{d\gamma}{ds} : I \to \mathbb{S}^2 \subset \mathbb{R}^3.$$

Here $\mathbb{S}^2$ denote the unit 2-sphere $\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \,|\, x^2 + y^2 + z^2 = 1\}$. That is, $|T| = 1$.

In the case plane curve, the normal vector N is the 90° counterclockwise rotation of T. What is now different is that there is no distinguished normal vector—at each point of a space curve, there is a circle worth of unit vectors that are orthogonal to T.

Definition 3.2 On a space curve, a function $N_1 : I \to \mathbb{S}^2$ satisfying $\langle N_1, T\rangle = 0$ is a **normal vector field**, or a **frame**. Note that a normal vector field N_1 gives rise to a **binormal vector field** N_2 by

$$N_2 = T \times N_1$$

so that T, N_1, N_2 form an orthonormal basis (and hence the name frame).

R A frame is an additional geometric structure on top of the space curve. We call a space curve γ with a choice of N_1 a **framed curve** or a **ribbon**.

3.1 There is More than One Way to Frame a Curve

A framed curve is a curve together with a choice of a normal vector field. The vector N_1 and its complementary basis vector N_2 can describe the guide line for a solid material thickened around the curve, the metal rings of a chain, or a model for DNA molecules. For example, a ribbon—a space curve extended into narrow surface along a normal vector field—is a framed curve. Similar to ribbons, any physical thin wire, tube or rod is a framed curve: how a piece of the wire is attached to the next piece of the wire defines how the normal vector is propagated along the curve.

The notion of framed curves are essential in 3D graphics designs. In order to build a tube of cross section radius $r : I \to \mathbb{R}_{>0}$ around a given space curve $\gamma : I \to \mathbb{R}^3$:

$$\begin{bmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{bmatrix} = \gamma(u) + r(u)\cos(v)N_1(u) + r\sin(v)N_2(u), \quad u \in I, \quad -\pi < v \le \pi, \tag{3.1}$$

we need a normal vector field N_1. Note that the result of the tube heavily depends on the choice of N_1.

As a classical subject and a practically important piece of mathematics, the theory of framed curves comes surprisingly late in the history. It is usually referred as the **ribbon theory** developed in the 1950's. Even today it is not widely known or taught in differential geometry. In almost every elementary differential geometry textbook on curves and surfaces, the only frame one would find is the **Frenet–Serret frame**. Frenet (1847) and Serret (1851) independently established the differential relationships between (T, N_1, N_2) for a *specific* frame, called the Frenet–Serret frame. The Frenet–Serret frame is adapted to the osculating circle, which is geometrically

interesting. However, the Frenet–Serret frame is usually not the frame that is best adapted to the physics or graphics application. It is neither the frame that produces a good tube around a space curve for graphics, nor the frame that arises when a physical wire is put at rest. It is necessary to look beyond the Frenet–Serret frame to work with space curves in practice.

In 1975, Richard Bishop wrote an article entitled "There is more than one way to frame a curve" that addressed the concerns about Frenet–Serret frames. Bishop suggested the notion of **parallel frame** (a.k.a. **Bishop frame**) which behaves much more natural than the Frenet–Serret frame. Contrasting the Frenet–Serret frame, parallel frames are the more relevant basis for modeling physical wires and building tubes around curves.

3.1.1 Complex Normal Plane

At each point on a framed space curve, the vectors N_1, N_2 span the 2-dimensional **normal space**. This two dimensional space can be regarded as a complex plane. For each normal vector

$$Z \in \operatorname{span}\{N_1, N_2\}, \quad Z = aN_1 + bN_2, \quad a, b \in \mathbb{R}$$

we consider its complex representation assembled from its coefficients:

$$Z_{\mathbb{C}} = z \coloneqq a + b\mathbb{i} \in \mathbb{C}.$$

A multiplication by $\mathbb{i}$ amounts to a 90° rotation about T:

$$\mathbb{i}Z_{\mathbb{C}} = (T \times Z)_{\mathbb{C}}.$$

If we take $\mathbb{i}$ as such a 90° rotation *operator*, then we can simply write

$$N_2 = \mathbb{i}N_1, \quad Z = (a + b\mathbb{i})N_1 = Z_{\mathbb{C}}N_1. \tag{3.2}$$

That is, the complex representation $Z_{\mathbb{C}}$ of $Z \in \operatorname{span}\{N_1, N_2\}$ is the complex scalar, as an operator, that transforms N_1 into Z.

3.1.2 Complex Curvature

Recall that in the case of plane curve, by taking the derivative of $\langle T, T\rangle = 1$ we find $\langle T', T\rangle = 0$. We then define the curvature κ by the scalar so that $T' = \kappa N$ (with the arclength parametrization). In the case of a space curve, we also have $\langle T', T\rangle = 0$ for the same reason. Based on the frame N_1 we define:

Definition 3.3 — Complex curvature. For a framed curve $\gamma\colon I \to \mathbb{R}^3, N_1\colon I \to \mathbb{S}^2$, $\langle N_1, T\rangle = 0$, the **complex curvature** is a complex-valued function over the curve

$$\psi\colon I \to \mathbb{C}$$

so that

$$\frac{dT}{ds} = \psi N_1. \tag{3.3}$$

The following describes how the complex curvature depends on the choice of frame N_1.

Theorem 3.1 Let N_1 and $\widetilde{N}_1$ be two different frames on the same curve, related by the rotation

$$\widetilde{N}_1 = e^{\mathbb{i}\phi} N_1, \quad \phi \colon I \to \mathbb{T}^1.$$

Let ψ and $\widetilde{\psi}$ be the complex curvature defined via N_1 and $\widetilde{N}_1$ respectively. Then

$$\widetilde{\psi} = e^{-\mathbb{i}\phi}\psi. \tag{3.4}$$

Proof. By definition (3.3) we have $\psi N_1 = \widetilde{\psi}\widetilde{N}_1 = \widetilde{\psi} e^{\mathbb{i}\phi} N_1$. Therefore, $\psi = \widetilde{\psi} e^{\mathbb{i}\phi}$ which implies (3.4). □

3.1.3 Frenet–Serret Frame

Similar to plane curves, the vector dT/ds is a normal vector scaled by the curvature. For example it arises as the variation of total length. It is also a related to the osculating circle.

Definition 3.4 — Curvature normal and Frenet–Serret frame. The vector $dT/ds = \psi N_1$ of Eq. (3.3) is called the **curvature normal**. The norm of the complex curvature is the **Frenet–Serret curvature**, **real curvature**, or simply **curvature**:

$$\kappa := |\psi| \colon I \to \mathbb{R}_{\geq 0}. \tag{3.5}$$

If at a point on the curve we have $\kappa \neq 0$, then we define the **Frenet–Serret normal**:

$$N_{\mathrm{FS}} := \frac{dT/ds}{|dT/ds|} = \frac{1}{\kappa}\frac{dT}{ds} = \frac{\psi}{\kappa} N_1, \quad \text{that is,} \tag{3.6}$$

$$\frac{dT}{ds} = \kappa N_{\mathrm{FS}}. \tag{3.7}$$

The **Frenet–Serret frame** is the frame where $N_1 = N_{\mathrm{FS}}$, which is properly defined only when κ is non-vanishing.

Another way to interpret (3.7) is that *the Frenet–Serret frame is the frame in which the complex curvature is real and positive: $\psi = \kappa$.*

R The curvature normal, the real curvature, and the Frenet–Serret frame are geometric object of the space curve itself independent of the frame N_1. This is because they can directly be written in terms of T and ds.

Notation 3.1 Under the Frenet–Serret frame $N_1 = N_{\mathrm{FS}}$, the vector $N_2 = \mathbb{i} N_1$ is called the **Frenet–Serret binormal** often denoted by the letter B:

$$B_{\mathrm{FS}} := T \times N_{\mathrm{FS}} = \mathbb{i} N_{\mathrm{FS}}. \tag{3.8}$$

Osculating Circle

The reciprocal $1/\kappa$ of the real curvature is the radius of curvature, and the center of curvature is given by

$$\gamma + \frac{1}{\kappa} N_{\mathrm{FS}}.$$

The circle centered at the center of curvature with radius $1/\kappa$ is the **osculating circle**, which is the unique circle that makes a G^2 contact with the curve at the given point.

Definition 3.5 The **osculating plane** is the plane the osculating circle lies in. That is, the osculating plane at a point $\gamma(s)$ on a space curve γ is the plane spanned by the vectors $T(s)$ and $N_{\mathrm{FS}}(s)$ rooted at $\gamma(s)$, provided that N_{FS} is defined. In other words, the osculating plane has $B_{\mathrm{FS}}(s)$ as its normal vector.

3.1.4 The Twist of a Frame

Recall that in the case of plane curve, once we have $T' = \kappa N$ then we have $N' = -\kappa T$. For a framed space curve, it is a bit different. The derivative of the frame N_1 contains not only the information similar to $-\kappa T$ of how much the frame is bent along the curve, but also another information that describes how the frame *twists* along the curve.

First of all, since $\langle N_1, N_1\rangle = 1$, we have $\langle dN_1/ds, N_1\rangle = 0$. Thus dN_1/ds must only be a linear combination of T and N_2:

$$\frac{dN_1}{ds} = \left\langle T, \frac{dN_1}{ds}\right\rangle T + \left\langle N_2, \frac{dN_1}{ds}\right\rangle N_2. \tag{3.9}$$

We already know a relation $dT/ds = \psi N_1$. By taking the derivative of the condition $\langle T, N_1\rangle = 0$, we obtain

$$\begin{aligned} 0 &= \left\langle \frac{dT}{ds}, N_1\right\rangle + \left\langle T, \frac{dN_1}{ds}\right\rangle = \langle \psi N_1, N_1\rangle + \left\langle T, \frac{dN_1}{ds}\right\rangle = \mathrm{Re}(\psi) + \left\langle T, \frac{dN_1}{ds}\right\rangle \\ \Longrightarrow \quad & \left\langle T, \frac{dN_1}{ds}\right\rangle = -\mathrm{Re}(\psi). \end{aligned}$$

Therefore, the first term of (3.9) can be expressed by the real part of the complex curvature.

$$\frac{dN_1}{ds} = -\mathrm{Re}(\psi)T + \left\langle N_2, \frac{dN_1}{ds}\right\rangle N_2. \tag{3.10}$$

We can do the same calculation for N_2, and arrive at

$$\frac{dN_2}{ds} = -\mathrm{Im}(\psi)T + \left\langle N_1, \frac{dN_2}{ds}\right\rangle N_1. \tag{3.11}$$

Proposition 3.2 $\langle N_2, dN_1/ds\rangle = -\langle N_1, dN_2/ds\rangle$.

Proof. Take the derivative of $\langle N_1, N_2\rangle = 0$. □

Indeed, the first terms of (3.10) and (3.11) mimic the case of the case of plane curves. In the planar case, if T tilts in the direction of N with rate κ, then the derivative of N tilts in the direction of $-T$ with rate κ. Now, in the case of space curve, N_1 tilts not only in the direction of T but also in the direction of N_2. The rate that N_1 tilts into N_2 is how fast N_1 rotates about T, *i.e.* how fast the frame twists along the curve.

Definition 3.6 — Twist. The **twist** of a frame is given by

$$\omega := \left\langle N_2, \frac{dN_1}{ds} \right\rangle \left(= -\left\langle N_1, \frac{dN_2}{ds} \right\rangle \right) = \left\langle \mathbb{i} N_1, \frac{dN_1}{ds} \right\rangle = \left\langle N_1, -\mathbb{i} \frac{dN_1}{ds} \right\rangle. \tag{3.12}$$

In terms of the twist, (3.10) and (3.11) become

$$\begin{aligned} \frac{dN_1}{ds} &= -\operatorname{Re}(\psi) T + \omega N_2 \\ \frac{dN_2}{ds} &= -\operatorname{Im}(\psi) T - \omega N_1. \end{aligned}$$

Just like the complex curvature, the twist is also a quantity that depends on the frame.

Theorem 3.3 With the same setup as Theorem 3.1, let ω and $\widetilde{\omega}$ be the twist of the frame N_1 and $\widetilde{N}_1$ respectively. Then

$$\widetilde{\omega} = \omega + \frac{d\phi}{ds}. \tag{3.13}$$

Proof.

$$\begin{aligned} \widetilde{\omega} &= \left\langle \widetilde{N}_1, -\mathbb{i} \frac{d\widetilde{N}_1}{ds} \right\rangle = \left\langle e^{\mathbb{i}\phi} N_1, -\mathbb{i} \frac{d(e^{\mathbb{i}\phi} N_1)}{ds} \right\rangle \\ &= \left\langle e^{\mathbb{i}\phi} N_1, -\mathbb{i} \frac{d(e^{\mathbb{i}\phi})}{ds} N_1 \right\rangle + \left\langle e^{\mathbb{i}\phi} N_1, -\mathbb{i} e^{\mathbb{i}\phi} \frac{dN_1}{ds} \right\rangle \\ &= \left\langle e^{\mathbb{i}\phi} N_1, \frac{d\phi}{ds} e^{\mathbb{i}\phi} N_1 \right\rangle + \left\langle N_1, -\mathbb{i} \frac{dN_1}{ds} \right\rangle = \frac{d\phi}{ds} + \omega. \end{aligned}$$

□

In sum, we have:

Corollary 3.4 A general frame (T, N_1, N_2) satisfies the **frame equation**

$$\begin{bmatrix} dT/ds \\ dN_1/ds \\ dN_2/ds \end{bmatrix} = \begin{bmatrix} 0 & \operatorname{Re}(\psi) & \operatorname{Im}(\psi) \\ -\operatorname{Re}(\psi) & 0 & \omega \\ -\operatorname{Im}(\psi) & -\omega & 0 \end{bmatrix} \begin{bmatrix} T \\ N_1 \\ N_2 \end{bmatrix}. \tag{3.14}$$

Under a transformation $\widetilde{N}_1 = e^{\mathbb{i}\phi} N_1$, the complex curvature and the twist are transformed by

$$\widetilde{\psi} = e^{-\mathbb{i}\phi} \psi, \quad \widetilde{\omega} = \omega + \frac{d\phi}{ds}. \tag{3.15}$$

▪ **Example 3.3 — Frame equation for the Frenet–Serret frame.** The Frenet–Serret frame is the frame so that the complex curvature ψ takes a positive real value $\psi = \kappa$. Then the frame equation reads

$$\begin{bmatrix} dT/ds \\ dN_{\mathrm{FS}}/ds \\ dB_{\mathrm{FS}}/ds \end{bmatrix} = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \omega_{\mathrm{FS}} \\ 0 & -\omega_{\mathrm{FS}} & 0 \end{bmatrix} \begin{bmatrix} T \\ N_{\mathrm{FS}} \\ B_{\mathrm{FS}} \end{bmatrix}. \tag{3.16}$$

The twist ω_{FS} of the Frenet–Serret frame is called the **Frenet torsion**, often denoted by $\tau = \omega_{\text{FS}}$. The last row of Eq. (3.16)

$$\frac{dB_{\text{FS}}}{ds} = -\tau N_{\text{FS}} \tag{3.17}$$

shows that the Frenet torsion measures the rate of change of the normal vector of the osculating plane. ▪

3.1.5 Fundamental Theorem of Framed Space Curves

Given a framed curve (γ, N_1), we derive from it the complex curvature ψ and the twist ω. Conversely, given an arbitrarily given complex curvature ψ and twist ω, we can uniquely reconstruct a curve up to a Euclidean motion.

Theorem 3.5 — Fundamental Theorem of framed space curves. Given continuous functions $\psi\colon [0,L] \to \mathbb{C}$ and $\omega\colon [0,L] \to \mathbb{C}$, there exists an arclength-parametrized curve $\gamma\colon [0,L] \to \mathbb{R}^3$ with a frame $N_1\colon [0,L] \to \mathbb{S}^2$ that has ψ and ω as its complex curvature and twist. The curve is unique up to a Euclidean motion. That is, any other solution $\widetilde{\gamma}$ differs from γ by a Euclidean motion $\widetilde{\gamma} = R\gamma + b$ for some 3D rotation R and some 3D translation $b \in \mathbb{R}^3$.

Proof. Given any (positively oriented) orthonormal basis $T(0), N_1(0), N_2(0)$ of $\mathbb{R}^3$ as the initial value, there is a unique solution $T(s), N_1(s), N_2(s)$ to the frame equation (3.14) as a result of the existence and uniqueness of initial value problems of ordinary differential equations. Any other choice of the initial frame $\widetilde{T}(0), \widetilde{N}_1(0), \widetilde{N}_2(0)$ must be a rotation R of $T(0), N_1(0), N_2(0)$. One then checks that $\widetilde{T}(s) = RT(s), \widetilde{N}_1(s) = RN_1(s), \widetilde{N}_2(s) = RN_2(s)$ is the consequent solution to the frame equation. The curve $\gamma(s)$ is obtained by integrating $d\gamma/ds = T(s)$; *i.e.* $\gamma(s) = \gamma(0) + \int_0^s T(s')\,ds'$. A different choice of the initial $\widetilde{T}$ leads to a rotation in γ; a different choice of $\gamma(0)$ yields a translation. Other than these Euclidean motions, γ is unique. □

3.1.6 Bishop (Parallel) Frame

A frame that has no twist is a natural consideration in practice. For example, in 3D geometric modeling, the frame used to scaffold a tube around a curve had better have no twist, as twists lead to uneven thickness of the tube.

Definition 3.7 A frame N_1 is called a **parallel frame** or a **Bishop frame** if it has no twist ($\omega = 0$).

Theorem 3.6 A non-closed curve $\gamma\colon [0,L] \to \mathbb{R}^3$ (as opposed to a closed curve which has a periodic domain) always has a parallel frame. The parallel frame is unique up to a global phase; *i.e.* any parallel frame $\widehat{N}_1$ is in a constant angle away from another parallel frame $e^{\mathbb{i}\phi_0} N_1(s) = \widehat{N}_1(s)$.

Proof. Suppose N_1 is an existing frame on γ which is not necessarily parallel, *i.e.* $\omega \neq 0$. We can "untwist" the frame, $\widetilde{N}_1(s) = e^{i\phi(s)} N_1(s)$, with $\phi(s)$ designed by the following. From the transformation rule (3.15), namely $\widetilde{\omega} = \omega + d\phi/ds$, we see that $\widetilde{N}_1$ will be parallel if $d\phi/ds = -\omega$; that is, $\phi(s) = \phi(0) - \int_0^s \omega(s')\,ds'$. Note that in general $\phi(L) \neq \phi(0)$, so we may not obtain a parallel frame on a closed curve without discontinuity in its frame.

The uniqueness up to global phase is clear again from the transformation rule $\widetilde{\omega} = \omega + {}^{d\phi}/_{ds}$. The phase shift ϕ is constant if and only if the twist modification ${}^{d\phi}/_{ds}$ remains zero. □

The frame equation (3.14) for a parallel frame becomes

$$\begin{bmatrix} {}^{dT}/_{ds} \\ {}^{dN_1}/_{ds} \\ {}^{dN_2}/_{ds} \end{bmatrix} = \begin{bmatrix} 0 & \operatorname{Re}(\psi_P) & \operatorname{Im}(\psi_P) \\ -\operatorname{Re}(\psi_P) & 0 & 0 \\ -\operatorname{Im}(\psi_P) & 0 & 0 \end{bmatrix} \begin{bmatrix} T \\ N_1 \\ N_2 \end{bmatrix} \tag{3.18}$$

where ψ_P denotes the complex curvature for a parallel frame.

R A parallel frame is a frame that has no twist. The Frenet–Serret frame is the frame whose complex curvature is real and positive.

Build a Parallel Frame from Scratch

In Theorem 3.6 the parallel frame is constructed by first giving an arbitrary frame, measuring its twist, and then untwisting it via a change of frame. An alternative way is by solving the initial value problem for the frame $N_1(s)$

$$\frac{dN_1}{ds}(s) = \left(T(s) \times \frac{dT}{ds}(s) \right) \times N_1(s), \quad T(s) \text{ and } N_1(0) \perp T(0) \text{ are given.} \tag{3.19}$$

By the cross product identity $(A \times B) \times C = \langle A, C \rangle B - \langle B, C \rangle A$, an equivalent way of writing (3.19) is

$$\frac{dN_1}{ds}(s) = -\left\langle \frac{dT}{ds}(s), N_1(s) \right\rangle T(s). \tag{3.20}$$

Or, by the Frenet–Serret curvature normal ${}^{dT}/_{ds} = \kappa N_{\mathrm{FS}}$, Eq. (3.19) can also be written as

$$\frac{dN_1}{ds}(s) = \kappa(s) B_{\mathrm{FS}}(s) \times N_1(s). \tag{3.21}$$

It is easy to verify from (3.20) that these (equivalent) equations for N_1 are precisely the equation for N_1 in (3.18).

Frenet Torsion in terms of a Parallel Frame

Suppose ψ_P is the complex curvature of a parallel frame N_1. In polar form

$$\psi_P = \kappa e^{\mathbb{i}\theta}, \quad \text{for some } \theta \colon I \to \mathbb{T}^1.$$

If we apply a change of frame $\widetilde{N}_1 = e^{\mathbb{i}\theta} N_1$, which implies $\widetilde{\psi} = e^{-\mathbb{i}\theta} \psi_P = \kappa$, then we arrive at the Frenet–Serret frame $\widetilde{N}_1 = N_{\mathrm{FS}}$. Since the parallel frame N_1 has no twist, the Frenet–Serret twist (*i.e.* the Frenet torsion) is given by $\tau = 0 + {}^{d\theta}/_{ds}$. In conclusion:

Theorem 3.7 Let $\psi_P = \kappa e^{\mathbb{i}\theta}$ be the complex curvature with respect to a parallel frame. Then the Frenet torsion is the derivative of its complex phase

$$\tau = \frac{d\theta}{ds} = \frac{1}{|\psi|^2} \left\langle \psi, -\mathbb{i} \frac{d}{ds} \psi \right\rangle. \tag{3.22}$$

In other words, given the real curvature κ and the Frenet torsion τ, the parallel frame complex

curvature is given by

$$\psi_P(s) = \kappa(s) e^{\mathbb{i}\theta(0)} e^{\mathbb{i}\int_0^s \tau(s')\,ds'}. \tag{3.23}$$

The freedom to select the constant phase $\theta(0)$ reflects the uniqueness up to a global phase of parallel frames.

3.1.7 Frames with Uniform Twist

As noted in the proof of Theorem 3.6, a parallel frame does not always exist for a closed curve. It can also be seen more directly from (3.19) that the solution $N_1(s)$ of initial value problem has $N_1(0) \neq N_1(L)$ in general.

R Another scenario where parallel frame fails to exist is to have prescribed boundary frames. Suppose $\gamma\colon [0, L] \to \mathbb{R}^3$ is a non-closed curve, with prescribed $N_1^{\text{given}}(0)$ and $N_1^{\text{given}}(L)$. Imagine a physical wire with clamped boundaries where not only the curve but also the frame is fixed. Then the unique solution $N_1(s)$ of (3.19) using the initial value $N_1^{\text{given}}(0)$ generally does not yield $N_1(L) = N_1^{\text{given}}(L)$.

In case that the parallel frame does not exist, the next "best" frame is a frame with uniform twist.

Definition 3.8 A framed curve (γ, N_1) has **uniform twist** if the twist $\omega(s) = \omega$ is constant along the curve.

Any frame can be modified into a uniformly twisted frame without changing the **total twist**. Suppose N_1 is an arbitrary frame with twist function $\omega(s)$. Call the total twist

$$\text{Twist} = \int_0^L \omega(s)\,ds. \tag{3.24}$$

The uniformly twisted frame with the same total twist must have the twist given by

$$\widetilde{\omega} = \frac{\text{Twist}}{L}, \quad \text{so that } \widetilde{\text{Twist}} = \int_0^L \widetilde{\omega}\,ds = \frac{\text{Twist}}{L}\int_0^L ds = \text{Twist}\,. \tag{3.25}$$

This constant twist $\widetilde{\omega} = {}^{\text{Twist}}/_{L}$ can be achieved by the change of frame:

$$\phi(s) = \phi(0) - \int_0^s \omega(s')\,ds' + \widetilde{\omega}s, \quad \widetilde{N}_1(s) = e^{\mathbb{i}\phi(s)} N_1(s). \tag{3.26}$$

R The notion of uniform twist is natural in elastic rods and wires. A basic model is that the part of the elastic potential energy due to twisting is given by

$$E_{\text{twist}} = \int_0^L \omega(s)^2\,ds.$$

By the *Cauchy–Schwarz inequality* we have

$$\left(\int_0^L \omega(s)\,ds\right)^2 \leq \left(\int_0^L ds\right)\left(\int_0^L \omega(s)^2\,ds\right)$$

with equality holds if and only if $\omega(s)$ is a constant function. In other words,

$$E_{\text{twist}} \geq \frac{(\text{Twist})^2}{L}, \quad \text{minimum is reached when } \omega(s) \text{ is constant.}$$

Therefore, for an elastic rod or wire where the twisting is irresistibly relaxed the lowest potential state under a fixed total twist, we always observe a uniformly twisted frame.

3.2 Geometry of Closed Space Curves

Here we are concerned with closed curves.

3.2.1 Quantized Total Twist

Let $\gamma\colon \mathbb{T}^1 \to \mathbb{R}^3$ be a closed curve. We demand a frame N_1 over the curve to also be a periodic function $N_1\colon \mathbb{T}^1 \to \mathbb{R}^3$, $N_1(t+2\pi) = N_1(t)$.

Recall that the total twist of N_1 is given by

$$\text{Twist}[\gamma, N_1] = \int_{\mathbb{T}^1} \omega_{N_1}\, ds \in \mathbb{R}. \tag{3.27}$$

Here $\omega_{N_1}\colon \mathbb{T}^1 \to \mathbb{R}$ denotes the twist of N_1. Now, observe that an arbitrary change of frame $\widetilde{N}_1 = e^{i\phi} N_1$ yields

$$\text{Twist}[\gamma, \widetilde{N}_1] = \int_{\mathbb{T}^1} \omega_{\widetilde{N}_1}\, ds = \int_{\mathbb{T}^1} \left(\omega_{N_1} + \frac{d\phi}{ds}\right) ds = \text{Twist}[\gamma, N_1] + \int_{\mathbb{T}^1} \frac{d\phi}{ds}\, ds. \tag{3.28}$$

Since $\widetilde{N}_1 = e^{i\phi} N_1$ should stay periodic, we have that after a period ϕ differs by an integer multiple of 2π:

$$\phi(t+2\pi) = \phi(t) + 2\pi m \quad \text{for some } m \in \mathbb{Z}. \tag{3.29}$$

In other words,

$$\int_{\mathbb{T}^1} \frac{d\phi}{ds}\, ds = 2\pi m. \tag{3.30}$$

Therefore, (3.28) reads that *the total twists of any two arbitrary frames always differ by a* 2π *integer*

$$\text{Twist}[\gamma, \widetilde{N}_1] = \text{Twist}[\gamma, N_1] + 2\pi m, \quad m \in \mathbb{Z}. \tag{3.31}$$

Theorem 3.8 The total twist modulo 2π is a geometric quantity of a closed curve, independent of the frame.

▪ **Example 3.4** The total Frenet torsion $\int_{\mathbb{T}^1} \tau\, ds$ is the same, mod 2π, as the total twist of any other frame. ▪

3.2.2 Total Torsion and the Hodographic Area

Recall that the tangent hodograph of a curve is the trajectory of the tangent $T = {}^{d\gamma}/_{ds}$. In the case of a closed space curve, the hodograph is a closed curve on the 2-sphere $T\colon \mathbb{T}^1 \to \mathbb{S}^2$.

Definition 3.9 The **total torsion** of a closed space curve is a number in $\mathbb{R}/4\pi\mathbb{Z}$

$$\text{Torsion}[\gamma] = 2\pi - \text{Area}_{\mathbb{S}^2}[T] \mod 4\pi. \tag{3.32}$$

Here $\text{Area}_{\mathbb{S}^2}[T]$ is the spherical area enclosed by the tangent hodograph T, which is a mod-4π quantity.

▪ **Example 3.5** The total torsion of a closed plane curve (as a space curve that lies in a plane) of an even turning number is zero. If the turning number of the plane curve is odd, then the total torsion is 2π. This is because for plane curves the hodograph lies in the equator, enclosing a 2π or 0 hemisphere area (mod 4π) depending on the turning number. ▪

Note that a curve has a Frenet–Serret frame only when κ is non-vanishing, *i.e.* T' is never zero, *i.e.* the tangent hodograph T is a regular curve.

Theorem 3.9 If γ has a Frenet–Serret frame, that is the tangent hodograph T is a regular parametrized curve on $\mathbb{S}^2$, then the total torsion agrees with the total Frenet torsion

$$\text{Torsion}[\gamma] = \int_{\mathbb{T}^1} \tau \, ds \mod 4\pi. \tag{3.33}$$

Proof. The Frenet torsion is given by $\tau = \langle dN_{\text{FS}}/ds, B_{\text{FS}} \rangle$. Note that N_{FS} is the normalized T' and thus it is the tangent vector of the tangent hodograph. Also, T is the normal vector of the sphere at $T \in \mathbb{S}^2$, and hence $\tau = \langle dN_{\text{FS}}/ds, B_{\text{FS}} \rangle = \langle dN_{\text{FS}}/ds, T \times N_{\text{FS}} \rangle$ is the turning speed of the tangent vector N_{FS} of the hodograph per unit s (which is not the arclength of the hodograph). Hence the total Frenet torsion $\int_{\mathbb{T}^1} \tau \, ds$ is the total turning angle of the hodograph. By the **Gauss–Bonnet Theorem** on the unit sphere, which we will study in the surface theory, the total turning angle of any curve on the sphere equals to $2\pi -$ the area enclosed by the curve.

The discrete version of the spherical Gauss–Bonnet Theorem is in fact elementary, which the readers will explore in Exercise 3.2. □

R Definition 3.9 is more general than integrating a Frenet torsion, since the Frenet–Serret frame does not always exist. On the other hand, the enclosed area of a spherical curve like the hodograph, not necessarily regular or smooth, is always well-defined (up to 4π).

3.2.3 Linking Number between Two Closed Curves

Let $\gamma : \mathbb{T}^1 \to \mathbb{R}^3$ be an oriented closed space curve. Picture it as an electric wire that carries a unit current. The current generates a magnetic field $X_\gamma : \mathbb{R}^3 \to \mathbb{R}^3$. In 1820, Jean-Baptiste Biot and Félix Savart discovered through experiment that

$$X_\gamma(\mathbf{x}) = \frac{1}{4\pi} \oint_\gamma T \times \frac{\mathbf{x} - \gamma}{|\mathbf{x} - \gamma|^3} \, ds. \tag{3.34}$$

At the time the implication of the Biot–Savart law was not completely clear. In 1825, André-Marie Ampère also concluded from experiments an alternative statement to the magnetic fields generated by electric currents: The line integral of the magnetic field around an arbitrary chosen path C is (proportional to) the net electric current enclosed by the path

$$\oint_C \langle X_\gamma, dC \rangle = \text{number of times } C \text{ winds around } \gamma. \tag{3.35}$$

In 1833 Carl Friedrich Gauss put (3.34) and (8.17) together:

Definition 3.10 The **Gauss Linking number** between two non-intersecting closed curves C and γ is given by

$$\begin{aligned}\mathrm{Link}[C,\gamma] &:= \text{number of times } C \text{ winds around } \gamma \\ &= \frac{1}{4\pi}\oint_C\oint_\gamma \left\langle dC, d\gamma \times \frac{C-\gamma}{|C-\gamma|^3}\right\rangle. \end{aligned} \tag{3.36}$$

Exercise 3.1 Show that $\mathrm{Link}[C,\gamma] = -\mathrm{Link}[\gamma, C]$. ∎

In particular, the integral (3.36) is always an integer. Under smooth deformations of the two curves, as long as the two curve do not intersect, the integral stays invariant. It is a topological invariant.

3.2.4 Total Writhe

We saw in Section 3.2.1 that a change of frame leads to a 2π integer jump in the total twist. At first glance there does not seem to be a canonical choice among all these possible twists (2π apart from each other) that can be assigned to a curve. Which 2π branch the twist belongs to depends on the chosen frame. But actually there is a way to assign a canonical branch.

Suppose the curve γ does not self intersect. Then with a sufficiently small constant $\varepsilon > 0$, the offset curve $\gamma_\varepsilon(t) := \gamma(t) + \varepsilon N_1(t)$ also does not self intersect and makes no intersection with γ. One can view γ and γ_ε as the two boundary components of a ribbon; the ribbon is guided by the frame N_1.

Any two non-intersecting closed space curves, such as γ and γ_ε, have a topological invariant—the linking number.

Definition 3.11 Let us overload the notation of linking number. The **linking number** of a framed curve (γ, N_1) is defined by

$$\begin{aligned}\mathrm{Link}[\gamma, N_1] :=& \text{The linking number } \mathrm{Link}[\gamma_\varepsilon, \gamma] \\ & \text{of } \gamma_\varepsilon = \gamma + \varepsilon N_1 \text{ winding around } \gamma. \end{aligned} \tag{3.37}$$

Every change of frame which leads to a 2π jump in the twisting angle amounts to adding one to the linking number $\mathrm{Link}[\gamma, \mathrm{N}_1]$. A special branch of the twisting angle is the one where the frame makes no linking number $\mathrm{Link}[\gamma, \mathrm{N}_1] = 0$, *i.e.* the two boundaries of the ribbon is unlinked.

Definition 3.12 We call the total twist $\mathrm{Twist}[\gamma, N_1]$ of the frame N_1 that satisfies $\mathrm{Link}[\gamma, N_1] = 0$ the **total writhe**, or the **total writhing angle**, of the curve

$$\mathrm{Writhe}[\gamma] = \mathrm{Twist}[\gamma, N_1] \quad \text{where } N_1 \text{ is any frame with } \mathrm{Link}[\gamma, N_1] = 0. \tag{3.38}$$

R Note that the total writhing angle is a geometric quantity of the space curve, and the total twisting angle is a geometric quantity of the framed space curve.

Theorem 3.10 — Călugăreanu Theorem.

$$\mathrm{Link}[\gamma, N_1] = -\frac{1}{2\pi}\,\mathrm{Writhe}[\gamma] + \frac{1}{2\pi}\,\mathrm{Twist}[\gamma, N_1]. \tag{3.39}$$

R The sign convention in Theorem 3.10 is probably different from many expositions found in ribbon theory. The theorem is rather obvious using our definition of the writhe. The significance of the theorem is that the writhe is usually defined alternatively by

$$\frac{1}{2\pi}\,\mathrm{Writhe}[\gamma] = \frac{1}{4\pi}\oint_\gamma \oint_\gamma \left\langle \frac{\gamma(s)-\gamma(s')}{|\gamma(s)-\gamma(s')|^3}, d\gamma(s)\times d\gamma(s') \right\rangle \tag{3.40}$$

which one can see that it is the line integral of the Biot–Savart vector field along the generating curve γ itself. This is also why the writhe is also called the **self-linking number**. Starting from (3.40) it is less obvious that Theorem 3.10 holds.

R Another beautiful geometric fact is that (3.40) measures the *averaged crossing number*. Given the space curve γ, take an arbitrary position of the camera at infinity (an $\mathbb{S}^2$ degree of freedom of choice) to produce an orthographic view of the space curve. In the camera view, the curve becomes a plane curve with self-crossing. Depending on the orientation of the curve each crossing can be assigned $+1$ or -1. Each orthographic camera view is then assigned with the statistic of total crossing number. It turns out that $\frac{1}{4\pi}\,\mathrm{Writhe}[\gamma]$ is the crossing number averaged over all orthographic view (uniformly over $\mathbb{S}^2$).

3.3 Discrete Space Curves

Here we expect the readers to be familiar with Appendix A.

A discrete (polygonal) space curve is an ordered sequence of points in $\mathbb{R}^3$

$$\gamma : I = (0, 1, \ldots, n-1) \to \mathbb{R}^3. \tag{3.41}$$

The index set for the edges is given by

$$E = \begin{cases} \{(0,1),(1,2),\ldots,(n-2,n-1)\} & \text{non-closed curve;} \\ \{(0,1),(1,2),\ldots,(n-2,n-1),(n-1,0)\} & \text{closed curve.} \end{cases} \tag{3.42}$$

Each edge $(i, i+1) \in E$ ($i+1 \bmod n$ in the case of closed curve) is assigned with a **edge vector** and a **tangent vector**

$$\begin{aligned} v &: E \to \mathbb{R}^3, \quad v_{i,i+1} \coloneqq \gamma_{i+1} - \gamma_i, \\ T &: E \to \mathbb{S}^2, \quad T_{i,i+1} \coloneqq \frac{v_{i,i+1}}{|v_{i,i+1}|}. \end{aligned}$$

We assume the the curve is **regular**, namely v is never zero, and $T_{i-1,i} \neq -T_{i,i+1}$ for all i.

Each edge is therefore endowed with a normal plane

$$T^\perp_{i,i+1} = \left\{ u \in \mathbb{R}^3 \,\middle|\, \langle u, T_{i,i+1}\rangle = 0 \right\}. \tag{3.43}$$

A **frame** for a curve is an assignment of a unit vector in each normal plane

$$N\colon E \to \mathbb{S}^2, \quad N_{i,i+1} \in T^{\perp}_{i,i+1} \cap \mathbb{S}^2. \tag{3.44}$$

Each normal plane is identified with the complex plane,

$$T^{\perp}_{i,i+1} \leftrightarrow \mathbb{C}, \quad aN_{i,i+1} + bT_{i,i+1} \times N_{i,i+1} \leftrightarrow a + b\mathbb{i}. \tag{3.45}$$

3.3.1 Discrete Curvature and Twist

Similar to discrete plane curves, one can define the discrete curvature as the angle

$$\alpha_i = \cos^{-1}(\langle T_{i-1,i}, T_{i,i+1}\rangle) \tag{3.46}$$

between the edge tangents, the reciprocal of the radius of a notion of discrete osculating circle, or the gradient of the total length. These are more about the discrete version of real curvature. What is more about space curve is the complex curvature. It contains the information of not only the magnitude the curve is bending, but also the 3D direction in which the tangents are bent.

Definition 3.13 At those vertices i that has $T_{i-1,i} \neq T_{i,i+1}$ (and $T_{i-1,i} \neq -T_{i,i+1}$ as we have assumed throughout), we define the **discrete Frenet–Serret binormal vector**

$$B_{\mathrm{FS}i} := \frac{T_{i-1,i} \times T_{i,i+1}}{\left|T_{i-1,i} \times T_{i,i+1}\right|} \in \mathbb{S}^2. \tag{3.47}$$

The **dihedral rotation** (Appendix A.4.1) between two unit vectors $T_{i-1,i}, T_{i,i+1} \in \mathbb{S}^2$ is the rotation about their common normal

$$R_{\mathrm{Dihedral}(T_{i-1,i},T_{i,i+1})} = \begin{cases} \text{rotation by angle } \alpha_i \text{ about } B_{\mathrm{FS}i} = \frac{T_{i-1,i}\times T_{i,i+1}}{|T_{i-1,i}\times T_{i,i+1}|} & T_{i-1,i} \neq T_{i,i+1} \\ \text{identity} & T_{i-1,i} = T_{i,i+1}. \end{cases} \tag{3.48}$$

When we represent this rotation by quaternion, we take the branch where the quaternion has positive real part

$$\mathrm{Dihedral}(T_{i-1,i}, T_{i,i+1}) := \begin{cases} \cos\left(\frac{\alpha_i}{2}\right) + \sin\left(\frac{\alpha_i}{2}\right) B_{\mathrm{FS}i}, & T_{i-1,i} \neq T_{i,i+1}, \\ \mathbb{1} & T_{i-1,i} = T_{i,i+1} \end{cases} \tag{3.49}$$

$$= \begin{cases} \sqrt{\frac{1+\langle T_-, T_+\rangle}{2}} + \sqrt{\frac{1-\langle T_-, T_+\rangle}{2}} \frac{T_-\times T_+}{|T_-\times T_+|} & T_- \neq T_+ \text{ and } T_- \neq -T_+ \\ \mathbb{1} & T_- = T_+. \end{cases} \tag{3.50}$$

Here T_-, T_+ are shorthands for $T_{i-1,i}, T_{i,i+1}$ respectively. The dihedral rotation is among all rotation that maps $T_{i-1,i}$ to $T_{i,i+1}$ the most straightforward rotation; *i.e.* the rotation angle is exactly the angle between the two vectors, the vector $T_{i-1,i}$ travels along a great circle on the 2-sphere to $T_{i,i+1}$, and the unit quaternion on $\mathbb{S}^3$ representing the rotation is closest to $\mathbb{1}$.

The dihedral rotation is the **parallel transport** of a normal plane to the next normal plane

$$R_{\mathrm{Dihedral}(T_{i-1,i},T_{i,i+1})}\colon T^{\perp}_{i-1,i} \to T^{\perp}_{i,i+1}. \tag{3.51}$$

Definition 3.14 A discrete frame N over a discrete regular curve is called a **parallel frame** (**Bishop frame**) if

$$R_{\text{Dihedral}(T_{i-1,i},T_{i,i+1})}N_{i-1,i} = N_{i,i+1} \quad \text{for all interior vertex } i. \tag{3.52}$$

The twist of a frame measures how much a frame deviates from being parallel.

Definition 3.15 At each interior vertex i the **discrete twist** of a discrete frame N over a discrete regular curve is the angle $\omega_i \in (-\pi, \pi]$ by which $R_{\text{Dihedral}(T_{i-1,i},T_{i,i+1})}N_{i-1,i}$ needs to rotate about $T_{i,i+1}$ to arrive at $N_{i,i+1}$.

Theorem 3.11 Under a change of frame $\widetilde{N}_{i-1,i} = e^{\mathbb{i}\phi_{i-1,i}}N_{i-1,i}$ with some arbitrary function $\phi : E \to \mathbb{T}^1$, the discrete twist is transformed by

$$\widetilde{\omega}_i = \left[\omega_i + \phi_{i,i+1} - \phi_{i-1,i}\right]_{(-\pi,\pi]}. \tag{3.53}$$

Here $[\cdot]_{(-\pi,\pi]}$ means taking the $(-\pi, \pi]$ branch for the angle. In a more analytic form,

$$e^{\mathbb{i}\widetilde{\omega}_i} = e^{\mathbb{i}\omega_i}\frac{e^{\mathbb{i}\phi_{i,i+1}}}{e^{\mathbb{i}\phi_{i-1,i}}}. \tag{3.54}$$

The rotation $\text{Dihedral}(T_{i-1,i}, T_{i,i+1})$ plays the role of discrete curvature. It is similar to a finite difference for dT/ds that measures the deviation between two successive tangents. It is just that the difference is measured in the rotation instead of directly taking difference.

R Here the discrete curvatures and discrete twists are all discrete differential forms. That is, they measure angle, instead of measuring angle divided by some notion of edge length.

To make a parallel theory to the smooth one, it may be desirable to introduce a complex curvature $\psi_i \in \mathbb{C}$ as the representation of a full-blown quaternion $\text{Dihedral}(T_{i-1,i}, T_{i,i+1})$ or rotation matrix $R_{\text{Dihedral}(T_{i-1,i},T_{i,i+1})}$. The following is one example of a notion of complex curvature.

Chain of Spheres

At each vertex i take a small sphere centered at γ_i. One may set the radii of the spheres as some ℓ_i that satisfies $\ell_{i,i+1} = (\ell_i + \ell_{i+1})/2$ similar to the setting of chains of circles for discrete curves; so that we have a chain of touching spheres. Here we do not bother the detail of choosing the radii. If we only discuss the discrete curvature *differential form*, we only need a scale-invariant setup.

At the ith sphere the incoming tangent $T_{i-1,i}$ and outgoing $T_{i,i+1}$ gives two points on the sphere called p_- and p_+. We normalize the sphere as unit sphere, so $p_+ = T_{i,i+1} \in \mathbb{S}^2$ and $p_- = -T_{i-1,i} \in \mathbb{S}^2$. Using the frame at the previous edge $N_{i-1,i}$ as the reference coordinate system, we place a complex coordinate system over $\mathbb{S}^2 \setminus \{p_-\}$ through the **stereographic projection**

$$\begin{aligned}
&\text{StereoProj}_{p_-,N_{i-1,i}} : \mathbb{S}^2 \setminus \{p_-\} \to \mathbb{C},\\
&\text{StereoProj}_{p_-,N_{i-1,i}}\left(xN_{i-1,i} + y(p_- \times N_{i-1,i}) + zp_-\right) \to \frac{2x}{1-z} + \frac{2y}{1-z}\mathbb{i}.
\end{aligned} \tag{3.55}$$

with inverse formula

$$\begin{aligned}\mathrm{StereoProj}^{-1}_{p_-,N_{i-1,i}}\left(a+b\mathring{\mathbb{i}}\right) = \Bigg(&\frac{4a}{4+a^2+b^2}N_{i-1,i}+\frac{4b}{4+a^2+b^2}(p_-\times N_{i-1,i})\\&+\frac{-4+a^2+b^2}{4+a^2+b^2}p_-\Bigg).\end{aligned}\tag{3.56}$$

Through the stereographic projection we define the complex curvature

$$\psi_i := \mathrm{StereoProj}_{p_-,N_{i-1,i}}\left(p_+\right)\in\mathbb{C}.\tag{3.57}$$

One can check that

$$|\psi_i| = 2\tan\left(\frac{\alpha_i}{2}\right)\tag{3.58}$$

$$\arg(\psi_i) = \text{the angle from } N_{i-1,i} \text{ to } -\frac{T_{i-1,i}\times B_{\mathrm{FS}i}}{\left|T_{i-1,i}\times B_{\mathrm{FS}i}\right|} \text{ (rotating about } p_-).\tag{3.59}$$

Under a change of frame $\widetilde{N}_{i-1,i} = e^{\mathring{\mathbb{i}}\phi_{i-1,i}}N_{i-1,i}$ the complex curvature is transformed by $\widetilde{\psi}_i = e^{\mathring{\mathbb{i}}\phi_{i-1,i}}\psi_i$.

3.3.2 Hodograph of a Discrete Curve

The Frenet torsion is the rate of twisting of the Frenet–Serret frame. In the discrete setting B_{FS} sits on vertices. Therefore a natural place to define the Frenet torsion is on the edges.

Consider the tangent hodograph $T\colon E\to\mathbb{S}^2$, as a discrete curve on $\mathbb{S}^2$. The edges of γ become the vertices of the hodograph, and the vertices of γ correspond to the edges of the hodograph. In general, the hodograph may not be regular; it can happen that $T_{i-1,i} = T_{i,i+1}$. For the sake of the exposition, we will assume that $T_{i-1,i}\neq T_{i,i+1}$ for all i; namely T is a discrete regular curve on $\mathbb{S}^2$.

We call T a **spherical polygon**. All the connecting edges between the vertices are great circles on the sphere. The great circle connecting the vertex $T_{i-1,i}$ and the vertex $T_{i,i+1}$ is the rotation about the discrete Frenet–Serret binormal $B_{\mathrm{FS}i}$.

The torsion $\tau_{i,i+1}$ assigned to $(i,i+1)$ (a vertex for the spherical polygon hodograph) is the amount of change of the binormal from $B_{\mathrm{FS}i}$ to $B_{\mathrm{FS}i+1}$. This change is measured as the angle between them, which coincides with the angle between the two great circles, *i.e.* the exterior angle of the spherical polygon.

Definition 3.16 The discrete Frenet torsion $\tau_{i,i+1}$ is the exterior angle of the spherical polygon $(T_{n-1,0}, T_{0,1},\ldots,T_{n-2,n-1})$.

Exercise 3.2 In this exercise you will show that the sum $\sum_i\theta_i$ of the exterior angles θ_i of a spherical polygon is $2\pi - A$, where A is the area enclosed by the spherical polygon.

(a) Show that the area of a spherical triangle on the unit sphere with interior angles α_1, α_2,α_3 is

$$A = \alpha_1+\alpha_2+\alpha_3-\pi.\tag{3.60}$$

Hint *Consider the areas A_1, A_2, A_3 of the three **spherical lunes**.*

(b) Show that the area of a spherical n-gon with interior angles $\beta_1, \ldots, \beta_n$ is

$$A = (2 - n)\pi + \sum_{i=1}^{n} \beta_i. \tag{3.61}$$

Hint *Partition the polygon into triangles.*

(c) Conclude that the area of a spherical n-gon with exterior angles $\tau_1, \ldots, \tau_n$ is

$$A = 2\pi - \sum_{i=1}^{n} \tau_i. \tag{3.62}$$

■

4. Geometry of Discrete Surfaces

In this chapter and the next chapter, we study the elementary differential geometry of smooth and discrete surfaces. We first study the discrete surfaces, mainly triangle meshes. Then in the next chapter, we study the continuous analog of the discrete theory.

4.1 Mathematical Surfaces

A surface is a 2-dimensional abstract space called a 2-dimensional manifold. It is usually realized in the 3-dimensional space, which is more precisely called a realized surface. If the surface realization is "nice" enough, we call it an immersed surface or an embedded surface in 3D.

In general, an n-dimensional manifold is an abstract space that is locally parametrizable by patches in a Cartesian space $\mathbb{R}^n$.

For example, the curve we have seen is a 1-dimensional manifold immersed in a 2 or 3 dimensional space. It is because a connected 1-dimensional manifold is always diffeomorphic to an interval, an unbounded interval, or a circle, that we say the curve as a whole is parametrized by the interval or the circle. However, in the case of a surface, for example a sphere or a torus, it is harder to technically say the surface is parametrized by a single piece of $\mathbb{R}^2$ region (like an interval in the 1D case).

Therefore, in mathematics, a parametrized manifold is usually introduced as a collection of multiple patches called **charts**. This collection of charts is an **atlas**.

4.1.1 Manifold

Exactly like in geography, an atlas is a collection of maps (charts), each of which describes how each region of the world corresponds to a region under a 2D coordinate system. A collection of charts bundled in a book is not yet a useful atlas. An atlas also has a description how each chart is neighbored with each other. Every overlapping charts should have consistent maps.

A parametrized manifold is a set M (with technical assumptions such as being a paracompact Hausdorff topological space) together with a collection of open sets U_i that covers M, *i.e.*

$\mathcal{M} = \bigcup_i U_i$, and invertible maps $\mathbf{r}_i \colon U_i \to \mathbf{r}_i(U_i) \subset \mathbb{R}^2$ so that on each overlapping region $U_i \cap U_j$ the function $\mathbf{r}_j \mathbf{r}_i^{-1}$ is a diffeomorphism (a reparametrization) from $\mathbf{r}_i(U_i \cap U_j) \subset \mathbb{R}^2$ to $\mathbf{r}_j(U_i \cap U_j) \subset \mathbb{R}^2$. Here $\mathbf{r}_i = (u, v) \colon U_i \to \mathbb{R}^2$ is called a **coordinate**. The inverse $\mathbf{r}_i^{-1} \colon \mathbf{r}_i(U_i) \to U_i$ is called a **parametrization**.

Two atlases $(U_i, \mathbf{r}_i)_i$ and $(\widetilde{U}_j, \widetilde{\mathbf{r}}_j)_j$ are equivalent if when the two atlases are bundled together as a new atlas, the new atlas is self consistent. That is, on every overlapping region $U_i \cap \widetilde{U}_j$ between charts from either atlases, the function $\widetilde{\mathbf{r}}_j \mathbf{r}_i^{-1}$ is a diffeomorphism.

The equivalence class of atlased manifold is called a **differentiable manifold**.

If there exists an atlas consisting of charts for which all the transition reparametrizations $\mathbf{r}_j \mathbf{r}_i^{-1}$ preserve the orientation of the coordinate (the Jacobian $d(\mathbf{r}_j \mathbf{r}_i^{-1})$ has positive determinant), then we say the manifold is **orientable**.

4.1.2 Classification of Compact Surfaces

Two abstract surfaces $\mathcal{M}_1, \mathcal{M}_2$ are said to have the same topology if there exists a bijective continuous map $\varphi \colon \mathcal{M}_1 \to \mathcal{M}_2$ with continuous inverse $\varphi^{-1} \colon \mathcal{M}_2 \to \mathcal{M}_1$. We call $\mathcal{M}_1, \mathcal{M}_2$ are **homeomorphic**.[1]

For example the Bunny is a topological sphere, since it is homeomorphic to a sphere. A topological sphere is not homeomorphic to a torus. A torus with one disk removed is topologically different from a torus with two disks removed. Homeomorphisms map boundaries to boundaries.

A **closed surface** is a compact surface without any boundary.

The topological types of connected compact surfaces are completely classified. A *closed* connected compact surface is homeomorphic to exactly one of the following topological surfaces

1. a sphere $\mathbb{S}^2$;
2. a g-torus, which is the connected sum $\underbrace{\mathbb{T}^2 \# \mathbb{T}^2 \# \cdots \# \mathbb{T}^2}_{g}$ of g many copies of tori $\mathbb{T}^2$; we call the number g the **genus**;
3. the connected sum $\underbrace{\mathbb{RP}^2 \# \mathbb{RP}^2 \# \cdots \# \mathbb{RP}^2}_{k}$ of k number of copies of projective planes $\mathbb{RP}^2$.

A not-necessarily-closed connected compact surface is homeomorphic to one of the above types of surfaces with b number of disks removed.

Therefore, when we say "$\mathcal{M}$ is a torus with one disk removed" it is a complete description to the topological information of the abstract surface. Every version of tori with a disk removed will be homeomorphic to each other!

Here the projective plane $\mathbb{RP}^2$ is a non-orientable surface defined by identifying the antipodal points of $\mathbb{S}^2$. It is also known as the **Boy surface** whose first regular immersion into $\mathbb{R}^3$ is discovered by Werner Boy in 1901.

Every surface involving $\mathbb{RP}^2$ is non-orientable. Every orientable surface is either a sphere (with a few disks removed) or a g-torus (with a few disks removed).

▪ **Example 4.1** A disk is a sphere with a disk removed. An annulus is a sphere with two disks

[1] Every homeomorphic 2D differentiable manifolds are diffeomorphic. This is perhaps intuitive as most of the invertible continuous functions are either smooth or easy to be smoothed. Strangely enough, for n-dimensional manifolds, $n \geq 4$, there can be two differentiable manifolds that are homeomorphic but not diffeomorphic. The same topological manifold can admit different differentiable structures. Examples are *exotic 4-spaces* and *exotic 7-spheres*. Fortunately, the situation for 2D manifolds is quite simple.

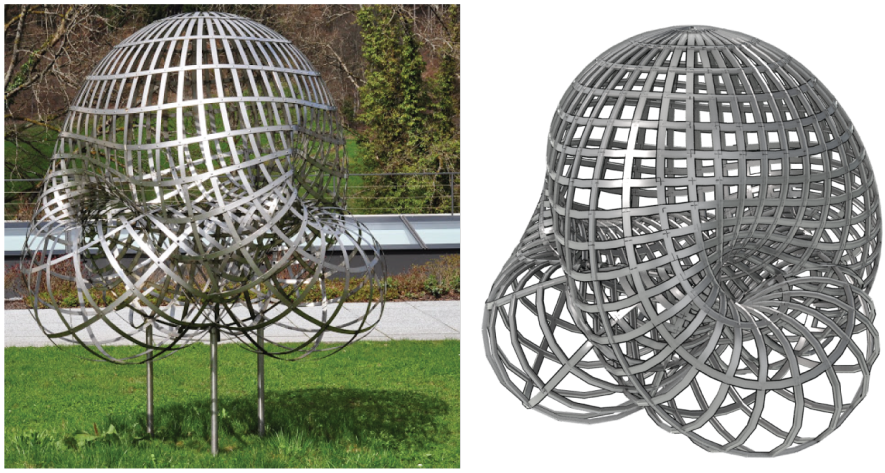


Figure 4.1 The Boy surface sculpture, constructed in 1991 by Mercedes-Benz, at the Oberwolfach Mathematical Research Institute.

removed. The Möbius strip is an $\mathbb{RP}^2$ with a disk removed. The Klein bottle is homeomorphic to $\mathbb{RP}^2 \# \mathbb{RP}^2$. ▪

▪ **Example 4.2 — The Boy surface at Oberwolfach.** The Boy surface sculpture at the Oberwolfach Mathematical Research Institute (Figure 4.1) is parametrized by the unit sphere via the following formula given by Kusner and Bryant. First, each $(X, Y, Z) \in \mathbb{S}^2$ is associated with the complex number $\zeta = \frac{X}{1-Z} + \mathbb{i}\frac{Y}{1-Z}$ (this is the stereographic projection). Let $p = 3$ (you may try different integers). Now parametrize the surface $(x(\zeta), y(\zeta), z(\zeta))$ as

$$x = \mathrm{Re}\left(\frac{\mathbb{i}(\zeta^{2p-1} - \zeta)}{\zeta^{2p} + 2\frac{\sqrt{2p-1}}{p-1}\zeta^p - 1}\right), \quad y = \mathrm{Re}\left(\frac{\zeta^{2p-1} + \zeta}{\zeta^{2p} + 2\frac{\sqrt{2p-1}}{p-1}\zeta^p - 1}\right),$$
$$z = \mathrm{Re}\left(\frac{\frac{p-1}{p}\mathbb{i}(\zeta^{2p} + 1)}{\zeta^{2p} + 2\frac{\sqrt{2p-1}}{p-1}\zeta^p - 1}\right). \tag{4.1}$$

The resulting surface will extend to infinity. Perform the following Möbius transformation: translate the surface $(x', y', z') = (x + a, y + b, z + c)$ for some constant vector $(a, b, c) \in \mathbb{R}^3$, and invert it in the unit sphere

$$\widetilde{x} = \frac{x'}{x'^2 + y'^2 + z'^2}, \quad \widetilde{y} = \frac{y'}{x'^2 + y'^2 + z'^2}, \quad \widetilde{z} = \frac{z'}{x'^2 + y'^2 + z'^2}. \tag{4.2}$$

The Boy surface sculpture at Oberwolfach takes $(a, b, c) = (0, 0, 0.33)$. ▪

4.1.3 Immersions and Embeddings

So far M is just an abstract space (together with an equivalence class of atlases). A surface immersed in 3D consists of a 2D abstract manifold M and a function $f : M \to \mathbb{R}^3$. The function is required to be **regular**, in the sense similar to that for curves, that its derivative df is full-rank. What this means is that

- Under a chart of any atlas, the function $f \circ \mathbf{r}^{-1} \colon \mathbf{r}(U) \subset \mathbb{R}^2 \to \mathbb{R}^3$ has the 3-by-2 derivative

$$d(f \circ \mathbf{r}^{-1}) = \begin{bmatrix} \frac{\partial (f_1 \circ \mathbf{r}^{-1})}{\partial u} & \frac{\partial (f_1 \circ \mathbf{r}^{-1})}{\partial v} \\ \frac{\partial (f_2 \circ \mathbf{r}^{-1})}{\partial u} & \frac{\partial (f_2 \circ \mathbf{r}^{-1})}{\partial v} \\ \frac{\partial (f_3 \circ \mathbf{r}^{-1})}{\partial u} & \frac{\partial (f_3 \circ \mathbf{r}^{-1})}{\partial v} \end{bmatrix}$$

  that has rank 2.
- In other words, for any regular curve $\gamma = (\gamma_u, \gamma_v) \colon I \to \mathbf{r}(U) \subset \mathbb{R}^2$, the composition $f \circ \mathbf{r}^{-1} \circ \gamma \colon I \to \mathbb{R}^3$ as a space curve is always a regular curve, since

$$\frac{d}{dt}\left(f \circ \mathbf{r}^{-1} \circ \gamma\right) = \begin{bmatrix} \frac{\partial (f_1 \circ \mathbf{r}^{-1})}{\partial u} & \frac{\partial (f_1 \circ \mathbf{r}^{-1})}{\partial v} \\ \frac{\partial (f_2 \circ \mathbf{r}^{-1})}{\partial u} & \frac{\partial (f_2 \circ \mathbf{r}^{-1})}{\partial v} \\ \frac{\partial (f_3 \circ \mathbf{r}^{-1})}{\partial u} & \frac{\partial (f_3 \circ \mathbf{r}^{-1})}{\partial v} \end{bmatrix} \begin{bmatrix} \frac{d\gamma_u}{dt} \\ \frac{d\gamma_v}{dt} \end{bmatrix} \neq 0$$

  due to the full-rankness.

We call such a function f an **immersion**.

The map f is further called an **embedding** if $f(M)$ has no self-intersection.

Theorem 4.1 An immersion $f \colon M \to \mathbb{R}^3$ is always a local embedding. That is, for each point $p \in M$ there exists a neighborhood $U \ni p$ so that $f(U)$ has no self-intersection.

Sketch. Observe that the image of an immersion is locally the level set of the (signed) distance function to the surface in $\mathbb{R}^3$ in a small neighborhood of the surface. The full-rankness of df ensures a well-defined normal, allowing the existence of a neighborhood where the distance function is smooth. □

Due to this property, immersions are also characterized as surface realizations that are local embeddings.

4.2 Discrete 2D Manifolds

A discrete surface is a **polygonal surface**. It is also described as a component of an abstract discrete 2D manifold M, and a function $f \colon M \to \mathbb{R}^3$ that realizes it in 3D.

A discrete 2D manifold is a polygon mesh with the property that it is locally a 2D Euclidean space.

4.2.1 Polygon Mesh

A **polygon mesh** consists of a collection of vertices V, a collection of edges E, and a collection of faces F.

Each (unoriented) edge $e = \{i, j\} \in E$ is described by two distinct vertices $i, j \in V$ as a set of two elements. That is, it is an unordered list $e = \{i, j\} = \{j, i\}$.

Each (unoriented) face f is a polygon, described by some number n_f of distinct edges $f = \{\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{n_f-1}, i_{n_f}\}, \{i_{n_f}, i_1\}\}$, so that every vertex involved in these edges occur exactly twice. A face is therefore also equivalent to a cyclically ordered list of vertices $f \equiv [i_1, \ldots, i_{n_f}]$ (for unoriented face we would also identify the reversed list $[i_{n_f}, \ldots, i_1]$ the same as $[i_1, \ldots, i_{n_f}]$).

R We have ruled out the possibility that a polygon may have two non-neighboring corners or sides to share the same vertex or edge. The notion of **cell complexes** allows these configurations.

We use the notation "$\prec$" to indicate the incidence relation. For example for $f = \{\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{n_f-1}, i_{n_f}\}, \{i_{n_f}, i_1\}\}$ we say the edge $e = \{i_2, i_3\}$ is incident to f, and write $e \prec f$. Similarly we have $i_2 \prec e$. If $f \in F$, then any e such that $e \prec f$ should be included in E. If $e = \{i, j\} \in E$, then both i, j should be included in V. If $i \prec e$ and $e \prec f$, then we also say $i \prec f$.

4.2.2 Manifold Conditions

Definition 4.1 A polygon mesh (V, E, F) is a manifold if both conditions hold:
(i) each edge $e \in E$ is incident to one or two faces;
(ii) the faces incident to each vertex $i \in V$ form a closed fan or an open fan.

As the discrete analog of the smooth theory, the ring of faces incident to a vertex is a chart of an atlas. Overlaps of charts are one or two faces.

An edge that is incident to only one face is a boundary edge. Its incident vertices are both boundary vertices, whose incident faces form just an open fan.

Notation 4.1 Let $V^\circ \subset V$ and $E^\circ \subset E$ denote the collection of non-boundary vertices and non-boundary edges respectively. They are the interior vertices and the interior edges.

4.2.3 Boundary Components

Observe that the collection of boundary edges always form cycles. (This is because starting with any unvisited boundary edge, one can uniquely find the next boundary edge using the fact that there are exactly two boundary edges of every open fan, attached to every boundary vertex. The singly linked list traversing through finitely many elements must form cycles.) Each of these cycles is called a **boundary component**.

R If we fill these boundary component with a polygon, defined exactly by these cycles, then the polygon mesh will be completed as a manifold without boundary.

Hence, an alternative definition for a manifold is that it is a polygon mesh V, E, F with each edge $e \in E$ incident to exactly two faces, and the incident faces of each vertex always form a closed fan. A subset of faces $B \subset F$ can be labeled as the set of boundary components if any two distinct boundary component faces $b_1, b_2 \in B$ never share the same incident edge or vertex.

4.2.4 Triangulated Surfaces

In most cases we will focus on polygonal surfaces where all faces are triangles. Those discrete 2D manifolds are called **triangle meshes** or **triangulated surfaces**. Another often known term for triangulated surfaces is **simplicial surfaces**.

Definition 4.2 The **valence** of an interior vertex of a polygonal surface is the number of edges incident to the vertex. For a triangulated surface, we call a vertex **regular** if the valence is 6. Otherwise we call it irregular.

Exercise 4.1 Illustrate a closed triangulated surface where all vertices are regular. ∎

4.3 Euler Characteristic

When we say the topology of the discrete manifold, we often mean the connectivity of the faces, edges and vertices. But sometimes we also refer it as the classification of surfaces like in the continuous setting. Note that a discrete surface is an atlased manifold, as a smooth manifold. Therefore it is classified into either a sphere (possibly with disks removed), a connected sum of a few tori (possibly with some number of disks removed), or a connected sum of a few Boy surfaces (possibly with disks removed).

Theorem 4.2 — Euler polyhedral formula. Every polygonal disk with $|V|$ vertices, $|E|$ edges and $|F|$ faces satisfies

$$|V| - |E| + |F| = 1. \tag{4.3}$$

Every polygonal sphere satisfies

$$|V| - |E| + |F| = 2. \tag{4.4}$$

Exercise 4.2 Show Theorem 4.2.
Hint *For a polygonal disk, what happens if you delete faces or vertices without destroying the topological disk? What happens if you only have one face left?* ∎

Definition 4.3 The **Euler characteristic** of $\mathcal{M} = (V, E, F)$ is defined by by

$$\chi(\mathcal{M}) := |V| - |E| + |F|. \tag{4.5}$$

Theorem 4.3 — Euler–Poincaré formula. The Euler characteristic $\chi(\mathcal{M})$ of a discrete 2D manifold $\mathcal{M} = (V, E, F)$ depends only on the topological type. Precisely,

- if $\mathcal{M}$ is a topological sphere with b disks removed, then $\chi(\mathcal{M}) = 2 - b$;
- if $\mathcal{M}$ is a genus g torus with b disks removed, then $\chi(\mathcal{M}) = 2 - 2g - b$;
- if $\mathcal{M}$ is the connect sum of k Boy surfaces with b disks removed, then $\chi(\mathcal{M}) = 2 - k - b$.

Exercise 4.3 Recall that a vertex in a triangle mesh is regular if it is valent 6. Show that the only closed triangulated surface where every vertex is regular is a torus or a Klein bottle. ∎

Exercise 4.4 Show that the minimum possible number m of irregular vertices in a closed, orientable, triangulated surface of genus g is

$$m(g) = \begin{cases} 4, & g = 0 \\ 0, & g = 1 \\ 1, & g \geq 2. \end{cases}$$

∎

Exercise 4.5 Show that the mean valence $\frac{1}{|V|}\sum_{i\in V}\text{valence}(i)$ of a closed simplicial surface of a fixed topology converges to 6 as $|V| \to \infty$. Conclude that the ratio of vertices to edges to faces approaches the following ratio

$$|V| : |E| : |F| \to 1 : 3 : 2. \tag{4.6}$$

■

4.4 Half-Edges

Recall that a polygon $\phi \in F$ is a cyclic ordered list of its incident vertices without distinguishing $\phi = [i_1, \ldots, i_{n_\phi}]$ or the reversed cyclic ordered list $\phi = [i_{n_\phi}, \ldots, i_1]$. An **oriented polygon** is a cyclic ordered list of the vertices; *i.e.* $[i_1, \ldots, i_{n_\phi}] \neq [i_{n_\phi}, \ldots, i_1]$.

An oriented polygon $\phi = [i_1, \ldots, i_{n_\phi}]$ assigns an oriented **half-edge** $h = (i_j, i_{j+1})$ for each incident edge $e = \{i_j, i_{j+1}\} \prec \phi$. Two neighboring oriented polygons are said to have consistent orientation if the half-edges they assigned to their common edge have opposite orientations.

An oriented discrete surface $M = (V, E, F)$ has each of its polygon $\phi \in F$ equipped with an orientation so that all neighboring polygons have consistent orientation. An oriented discrete surface is therefore endowed with a collection H of half edges, 2 oppositely oriented half edges per interior edge, and one oriented half edge per boundary edge.

4.4.1 Half-Edge Data Structure

In a typical half-edge data structure, an oriented discrete surface is described by a collection of vertices V, a collection of half-edges H, and a collection of faces F. In addition, there are the following pointers or maps that describe their incidence:

- $\textsc{src}: H \to V$, the source vertex i of a half-edge $h = (i, j)$;
- $\textsc{face}: H \to F$, the face the half-edge belongs to;
- $\textsc{next}: H \to H$, the next half-edge along the orientation of the face;
- $\textsc{flip}: H \to H \cup \{\textsc{null}\}$, the other half-edge sharing the same edge, null if the edge a boundary edge.

By composition of these functions, we obtain for example

- $\textsc{dst}: H \to V$, $\textsc{dst} = \textsc{src} \circ \textsc{next}$, the destination vertex j of a half-edge $h = (i, j)$;
- $\textsc{next} \circ \textsc{flip}: H \to H$, the next half-edge incident to the same (interior) vertex.

Using these functions, one can easily visit the neighboring edges, faces, or vertices around a vertex, or visit the incident edges of a face.

4.4.2 Minimalistic Half-Edge Data Structure

The minimalistic setup for a half-edge data structure is the following.

Definition 4.4 A closed oriented discrete 2D manifold is described by

- a finite set $H = \{h_1, \ldots, h_{|H|}\}$,
- a permutation $\textsc{next}: H \to H$, and
- a fixed-point-free involution $\textsc{flip}: H \to H$.

Here permutation is the synonym for an invertible map on a finite set to itself. A permutation $\rho: H \to H$ is called an involution if $\rho^2 = \text{id}$.

Lemma 4.4 *The orbit $(h, \sigma(h), \sigma^2(h), \ldots)$ generated by a permutation $\sigma : H \to H$ on an element h in a finite set H must form a cycle;* i.e. *$\sigma^n(h) = h$ for some n.*

Proof. Since there are only finitely many elements in H, the infinite sequence $(h, \sigma(h), \sigma^2(h), \ldots)$ must have at least two members in common, say $\sigma^k(h) = \sigma^\ell(h)$, $\ell > k$. Permutations are invertible. Therefore $h = \sigma^{-k}\sigma^\ell(h) = \sigma^{\ell-k}(h) = \sigma^n(h)$, with $n = \ell - k$. □

Definition 4.5 The orbits of NEXT are called faces. The orbits of (NEXT ∘ FLIP) are called vertices. The orbits of flip are called edges.

For manifolds with boundaries, just label a few faces as boundary component face. (See the remark in Section 4.2.3.)

▪ **Example 4.3** $H = \{h_1, h_2\}$, $\text{NEXT}(h_1) := h_1$, $\text{NEXT}(h_2) := h_2$, $\text{FLIP}(h_1) := h_2$, is a sphere, discretized into two 1-gons glued together. ▪

▪ **Example 4.4** $H = \{h_1, h_2\}$, $\text{NEXT}(h_1) := h_2$, $\text{NEXT}(h_2) := h_1$, $\text{FLIP}(h_1) := h_2$, is a sphere, discretized into one 2-gon. ▪

4.5 Piecewise Linear Surfaces

A triangulated surface $M = (V, E, F)$ is realized as a surface in 3D by assigning a vertex position per vertex

$$f : V \to \mathbb{R}^3.$$

These vertex positions are linearly (more precisely, affine) interpolated to form straight edges and flat triangular faces. We call the resulting surface a **piecewise linear (PL) surface**.

Definition 4.6 A PL surface $f : V \to \mathbb{R}^3$ is an **immersion** if
- each triangle has a nonzero area;
- the triangles incident to each vertex do not intersect each other.

The PL surface is an **embedding** if it has no self intersection.

In particular, the bending angle between two neighboring triangles in an immersed PL surface is not π or $-\pi$.

Definition 4.7 Given an oriented triangulated surface $M = (V, E, F)$ with a surface realization $f : V \to \mathbb{R}^3$, we have the following geometric measurements.
- For each oriented edge $ij \in H$, let

$$v_{ij} := f_j - f_i \in \mathbb{R}^3 \tag{4.7}$$

denote the **edge vector**. Note that $v_{ij} = -v_{ji}$. Let

$$\ell_{ij} := |v_{ij}| \tag{4.8}$$

be the edge length. Define the normalized edge vector

$$\hat{v}_{ij} := \frac{v_{ij}}{\ell_{ij}} \in \mathbb{R}^3. \tag{4.9}$$

- For each triangle $[ijk] \in F$ let

$$N_{ijk} := \frac{v_{ij} \times v_{ik}}{|v_{ij} \times v_{ik}|} \in \mathbb{S}^2 \tag{4.10}$$

be the **triangle normal**. Let

$$A_{ijk} := \frac{1}{2} |v_{ij} \times v_{ik}| \tag{4.11}$$

be the triangle area.

- For each corner i in a triangle $[ijk]$, define the **interior angle**

$$\theta^i_{jk} = {}_i\angle^k_j := \cos^{-1}\left(\langle \hat{v}_{ij}, \hat{v}_{ik} \rangle\right). \tag{4.12}$$

- For each interior edge $\{j,k\} \in E^\circ$ incident to $[ijk] \in F$ and $[\ell kj] \in F$, let the **bending angle** be

$$\alpha_{jk} := \sin^{-1}\left(\langle \hat{v}_{jk}, N_{ijk} \times N_{\ell kj} \rangle\right). \tag{4.13}$$

Note that $\alpha_{jk} = \alpha_{kj}$ and therefore it is a property of the edge, not only for the half-edge. Do not confuse it with the **dihedral angle**

$$\cos^{-1}\left(\langle N_{ijk}, N_{\ell kj} \rangle\right) \tag{4.14}$$

between the two triangles.

4.5.1 Intrinsic Quantities

A geometric measurement of a triangle mesh is an **intrinsic quantity** if it only depends on the edge lengths $(\ell_e)_{e \in E}$ of the mesh. Otherwise we call it an **extrinsic quantity**. In the following, we will give a few examples of intrinsic quantities. Suppose we are computing something on the mesh that only involves the intrinsic geometric measurements, then it does not depend on how the surface M is realized in 3D. All we need is an assignment of edge lengths to each edge. We call such a structure a **discrete metric** on a triangle mesh.

Definition 4.8 A **discrete metric** (a.k.a. **polyhedral metric**) on an abstract triangulated surface $M = (V, E, F)$ is a family of positive edge lengths $\ell : E \to \mathbb{R}_{>0}$ that satisfy the **triangle inequality**

$$\ell_{ij} + \ell_{jk} > \ell_{ki} \quad \text{for all } [ijk] \in F. \tag{4.15}$$

That is, the total length of any two edges of a triangle must be greater than the length of the third edge.

The edge vectors (4.7) and (4.9) are extrinsic. Since all quantities (4.10)–(4.14) are defined using the edge vectors, it seems that they are extrinsic. However, if we analyze them more carefully, the areas (4.11) and the interior angles (4.12) are intrinsic.

Indeed, edge lengths of a triangle uniquely determines the Euclidean shape of the triangle in the plane. Therefore, Euclidean geometric measurements such as the interior angles and the

area of each triangle is intrinsic.

Theorem 4.5 — Heron's Formula. The area of a triangle can be expressed in its edge lengths by

$$A_{ijk} = \frac{1}{4}\sqrt{\left(\ell_{ij}+\ell_{jk}+\ell_{ki}\right)\left(-\ell_{ij}+\ell_{jk}+\ell_{ki}\right)\left(\ell_{ij}-\ell_{jk}+\ell_{ki}\right)\left(\ell_{ij}+\ell_{jk}-\ell_{ki}\right)}. \tag{4.16}$$

Theorem 4.6 — Law of Cosine. The interior angle θ^i_{jk} of a triangle can be expressed in the edge lengths by

$$\theta^i_{jk} = \cos^{-1}\left(\frac{\ell_{ij}^2+\ell_{ki}^2-\ell_{jk}^2}{2\ell_{ij}\ell_{ki}}\right). \tag{4.17}$$

4.6 Angle Defects

Let there be a discrete metric $(\ell_e)_{e\in E}$ over $\mathcal{M}$.

As mentioned before, the discrete metric uniquely determines the Euclidean shape of each triangle. In particular, given any curve in a triangle, we have a well-defined total length, curvature, turning angle, *etc.* for this plane curve.

Suppose we have a curve on the triangle mesh. We denote it by $\gamma\colon I \to \mathcal{M}$. Assume that the curve crawls about faces and occasionally across edges, but never intersects any vertex. Up to perturbation we do not lose the generality with this assumption.

For segments of the curve within one triangles, we can measure the local geometric quantities. What if the curve crawls across an edge?

When the curve passes across an edge $jk \in E$, we take the neighboring triangles $[ijk], [\ell kj] \in F$ and realize them in the Euclidean space as two Euclidean triangles pieced together. This is possible because the two Euclidean triangles agrees on the length of the shared edge. Therefore the portion of the curve within the union of the triangles $[ijk]$ and $[\ell kj]$ can also be treated as a plane curve. In particular, there is no difficulty defining geometric measurements of the curve right at its intersection with the edge jk.

In other words, the geometry of a triangulated surface with a discrete metric is the same as a flat, unfolded polyhedron, with vertices being the exceptional points. Therefore the metric is also called a polyhedral metric.

Definition 4.9 For each interior vertex $i \in V$ define the **angle defect** or the **discrete Gaussian curvature form** by

$$(K\,dA)_i := 2\pi - \sum_{[ijk]\succ i} \theta^i_{jk}. \tag{4.18}$$

Exercise 4.6 — Descartes Theorem. Show that the total angle defect of a closed trian-

gulated surface M with polyhedral metric is

$$\sum_{i \in V} (K\, dA)_i = 2\pi \chi(M). \tag{4.19}$$

Here $\chi(M)$ is the Euler characteristic of M. ∎

4.6.1 Geodesic Curvature

By treating curves as plane curves in the unfolded polyhedron whose geometry is determined by the discrete metric, every consequential geometric quantities of the curves are automatically intrinsic measurements.

Definition 4.10 Let $\gamma : I \to M$ be a regular curve on a triangle mesh with a discrete metric. Then by laying the triangles flat locally, we have a plane curve. The curvature of the plane curve is called the **geodesic curvature**, denoted by $\kappa_g : I \to \mathbb{R}$.

Theorem 4.7 Let γ be a closed curve on M. If γ does not wind around any vertex, *i.e.* it is contractable to a point without touching any vertex, then the total turning number $\int_\gamma \kappa_g\, ds$ is a 2π integer.

Theorem 4.8 — Discrete Gauss–Bonnet Theorem (Turning Angle). Let γ be a regular simple closed curve that winds around a vertex $i \in V$ counterclockwise once. Then the total turning number of the curve is given by

$$\int_\gamma \kappa_g\, ds = \sum_{[ijk] > i} \theta^i_{jk}, \tag{4.20}$$

the sum of the interior angles at vertex i. In other words,

$$(K\, dA)_i + \int_\gamma \kappa_g\, ds = 2\pi. \tag{4.21}$$

Exercise 4.7 Show Theorem 4.8. ∎

4.6.2 Parallel Transport and Holonomy

Let p_1, p_2 be two points in the Euclidean plane. Let X be an arrow based at p_1. We call X a tangent vector at p_1. The parallel transport of X from p_1 to p_2 is the same vector X but now based at p_2.

On a surface, we cannot simply assign the parallel transport of a vector to be the same vector, since each point has a different tangent plane. Copying the vector to a different base point does not keep it tangential to the surface. Hence, the next best thing one could do is to transport the tangent vector little by little along a path, at each time keep it tangential to the surface and make sure the vector does not rotate about the normal. Such a parallel transport of a tangent vector depends on the path. Let us explain this process concretely for a discrete surface.

Let M be a triangle mesh with a polyhedral metric. Let γ be a curve on M without intersecting any vertex. The curve starts at p_1 and end at p_2, two points in some triangles. Let X be a

tangent vector at p_1.

Definition 4.11 The parallel transport of X along γ is by the following:

- In each segment of γ that is contained in one triangle, parallel transport X along the curve as in the Euclidean plane.
- When γ crawls over an edge, unfold the neighboring triangle so that they are both in the same Euclidean plane; transport the vector along the curve as parallel transport in a Euclidean plane.

The resulting vector when we arrive at the end point p_2 of γ is the parallel transport of X along γ.

Theorem 4.9 — Discrete Gauss–Bonnet Theorem (Holonomy). Let γ be a closed curve that winds around a vertex $i \in V$ counterclockwise once. Then the parallel transport of a vector along the closed curve will come back as another vector rotated by an angle. This angle is called the **holonomy angle**. The holonomy angle of such a loop equals to the angle defect $(K\,dA)_i$ of the vertex.

Exercise 4.8 Show Theorem 4.9. ∎

R Theorem 4.9 and Theorem 4.8 are very similar. The advantage of Theorem 4.9 is that the parallel transport can be done even if the curve is not regular, and is insensitive to having multiple additional loops along the curve. This contrasts the turning angle, which is sensitive to having additional loops which increase the total turning angle by a 2π integer (additional turning number). Note that in Theorem 4.8 we have assumed that the curve is simple, *i.e.* it has no self-intersection.

4.6.3 Area of the Gauss Map

Let $M = (V, H, F)$ be an oriented triangle mesh. Let there be a surface realization $f : V \to \mathbb{R}^3$, and let $N : F \to \mathbb{S}^2$ its normal vectors. Let $(\ell_e)_{e\in E}$ be the induced polyhedral metric. The metric defines intrinsic measurements, such as the angle defect $(K\,dA)_i$ for each $i \in V$. On the other hand, $N : F \to \mathbb{S}^2$ is an extrinsic measurement; (that is, knowing $(\ell_e)_{e\in E}$ alone is not enough to reconstruct N).

Definition 4.12 The normal vector as a map $N : F \to \mathbb{S}^2$ over the faces is called the **Gauss map**.

Here is a geometric law discovered by Gauss that relates the extrinsic Gauss map and the intrinsic angle defect.

Theorem 4.10 — Discrete Gauss–Bonnet Theorem (Area of the Gauss Map). Let γ be a simple closed curve that winds around a vertex $i \in V$ counterclockwise once; *i.e.* it is the boundary curve of a small disk D covering i but not any other vertex. Observe that $N \circ \gamma$ is a closed spherical polygon on $\mathbb{S}^2$, for which the vertices correspond to the normal of each face. (As the curve crawls across an edge, the normal N sweeps out a great circle by rotating about the edge. Therefore $N \circ \gamma$ is indeed a spherical polygon with great circular edges.) Then

the area of the region $N(D)$ enclosed by the spherical polygon $N \circ \gamma$ is the angle defect

$$\text{Area}(N(D)) = (K\,dA)_i = 2\pi - \sum_{[ijk]>i} \theta^i_{jk}. \tag{4.22}$$

Exercise 4.9 Show Theorem 4.10.
Hint *What are the exterior angles of this spherical polygon? Use the result of Exercise 3.2 to compute the enclosed area of a spherical polygon.* ▪

This implies a nontrivial rigidity of the normal vectors to a surface. Gauss called it "Theorema Egregium," which is "remarkable theorem" in Latin.

Corollary 4.11 — Discrete Gauss's Theorema Egregium. The area of the extrinsic Gauss map $\text{Area}(N(D))$ is intrinsic. That is, it only depends on the polyhedral metric $(\ell_e)_{e\in E}$.

4.7 Steiner's Formula

The angle defect measures the intrinsic curvature of the polyhedral surface. Here we talk about the extrinsic quantity α_{ij}, the bending angle (4.13), which also measures some aspect of the curviness of the surface.

Given an oriented surface realized in $\mathbb{R}^3$. Consider an ε normal offset of the surface. What we will obtain is that each triangle is translated in the normal direction with distance ε. Each edge is fanned out into a cylinder with radius ε centered about the edge as the axis, connecting the neighboring offset faces. Each vertex is fanned out into a spherical polygon on a sphere with radius ε centered at the vertex, capping the final missing pieces for the offset surface.

Between the original surface and the offset surface is the extruded region.

If α_{ij} is negative or the angle defect is negative, the extruded region forms more complicated caustics. We formally take the offset face as connected by cylinders and spheres, just with a negative weight. For visual explanation, we will mostly assume α_{ij} and $(K\,dA)_i$ are positive.

Definition 4.13 Define the **total mean curvature** around an edge $ij \in E$ as

$$(H\,dA)_{ij} := \frac{1}{2}\alpha_{ij}\ell_{ij} \in \mathbb{R}. \tag{4.23}$$

The total mean curvature around an edge describes the curvature of the corresponding cylinder of the offset surface. The total turning angle is α_{ij} in the rotation direction of the cylinder; on the other hand the cylinder is completely straight in other direction (the extent of the cylinder). Roughly speaking, the factor $\alpha_i/2$ in (4.23) arises by averaging α_{ij} and 0.

Theorem 4.12 — Steiner's Formula for Polyhedral Surfaces. The volume Vol_ε of the extruded region between a polyhedral surface and its ε normal offset

$$\text{Vol}_\varepsilon = \varepsilon \sum_{\phi\in F} A_\phi + \frac{\varepsilon^2}{2} \sum_{e\in E} 2(H\,dA)_e + \frac{\varepsilon^3}{3} \sum_{i\in V} (K\,dA)_i. \tag{4.24}$$

Here $\sum_{\phi\in F} A_\phi$ is the **total area** of the surface, $\sum_{e\in E}(H\,dA)_e$ is the **total mean curvature** of

the surface, and $\sum_{i \in V} (K\, dA)_i$ is the **total Gaussian curvature** of the surface. Note that by Descartes Theorem (Exercise 4.6) the total Gaussian curvature is a constant $2\pi\chi(M)$.

5. Geometry of Smooth Surfaces

In the last chapter, we studied the geometry of discrete surfaces. An abstract discrete surfaces are usually considered as triangle meshes. An abstract discrete surface equipped with edge lengths is a surface with a metric, which gives rise to intrinsic geometric notions, such as the discrete Gaussian curvature form (the angle defect). If we have the surface realized in 3D, then not only we have an induced metric but also extrinsic geometric notions such as the normal vectors (the Gauss map) and the bending angle (discrete mean curvature).

In this chapter, we look at the smooth counterpart of these notions.

Recall that an abstract surface, a.k.a. a 2D manifold, is a set (topological space) M of points structured with a bit of additional information—an equivalence class of atlases. An atlas is a collection of charts, each of which is an open subset of M that has a coordinate which maps the chart to a Cartesian space ($\mathbb{R}^2$), and each pair of overlapping charts enjoys a smooth coordinate transformation. Two different atlases are said to depict the same manifold if together they form an atlas (by combining the collections of charts) that still has consistent coordinate transformations between charts. In other words, the manifold is the set together with an equivalence class of atlases.

We will rarely talk about this formal definition of a manifold again. The message about the atlas formality is the following. Whenever we want, we have a local coordinate on the surface. But we need to make sure that any geometric quantity or geometric notion (if written in coordinates) should be consistent under coordinate transformations.

5.1 Tangent Spaces

The **tangent space** of a smooth 2D manifold M at $p \in M$ is the collection of all possible velocities of parametrized curves passing through the point.

First, one declare that the tangent space of M at p is a two-dimensional vector space, denoted by T_pM. Each element of T_pM is a tuple (p, X), called a **tangent vector** based at p. In particular, T_pM and T_qM are two different vector spaces (that share no common element) provided that

$p \neq q$. When the context is clear, we abbreviate (p,X) as X, and write $X \in T_pM$. As in any two-dimensional vector space, we can perform linear combination: $c_1X + c_2Y \in T_pM$ for each $X, Y \in T_pM$ and $c_1, c_2 \in \mathbb{R}$.

Now, the rule that associates tangent vectors to the manifold is the following. Given any parametrized curve $\gamma\colon I \to M$, $t_0 \in I$, the position $\gamma(t_0)$ and the velocity $\gamma'(t_0)$ at time t_0 provides a tangent vector

$$\big(\gamma(t_0), \gamma'(t_0)\big) \in T_{\gamma(t_0)}M, \quad \text{or simply} \quad \gamma'(t_0) \in T_{\gamma(t_0)}M. \tag{5.1}$$

Conversely, for every tangent vector $(p,X) \in T_pM$, there exists curves $\gamma\colon (-\varepsilon,\varepsilon) \to M$ such that

$$\gamma(0) = p \quad \text{and} \quad \gamma'(0) = X. \tag{5.2}$$

5.1.1 Under Coordinates

For example, suppose we take a coordinate $\mathbf{r} = (u,v)\colon U \to \mathbb{R}^2$, where $U \ni p$ is a neighborhood of p. Let $(u_0, v_0) := \mathbf{r}(p) \in \mathbb{R}^2$. Consider the curves traversing along the horizontal and the vertical axes at unit speed

$$C_1\colon (-\varepsilon,\varepsilon) \to \mathbb{R}^2, \quad C_1(t) := (u_0 + t, v_0), \tag{5.3}$$

$$C_2\colon (-\varepsilon,\varepsilon) \to \mathbb{R}^2, \quad C_2(t) := (u_0, v_0 + t). \tag{5.4}$$

By composing with $\mathbf{r}^{-1}$ we obtain a pair of curves $(\mathbf{r}^{-1} \circ C_1)(t) = \mathbf{r}^{-1}(C_1(t))$, $(\mathbf{r}^{-1} \circ C_2)(t) = \mathbf{r}^{-1}(C_2(t))$ that pass through p at $t = 0$ at velocities

$$e_1 := \left.\frac{dC_1}{dt}\right|_{t=0}, \quad e_2 := \left.\frac{dC_2}{dt}\right|_{t=0}. \tag{5.5}$$

Then these velocities e_1, e_2 will form a basis (called a **coordinate basis**) for the vector space T_pM. That is,

$$T_pM = \operatorname{span}\{e_1, e_2\}. \tag{5.6}$$

For any other curve $\gamma\colon (-\varepsilon,\varepsilon) \to M$ with $\gamma(0) = p$, the velocity $\gamma'(0) \in T_pM$ can be written as a linear combination of the basis vectors e_1, e_2

$$\gamma'(0) = c_1e_1 + c_2e_2, \quad c_1, c_2 \in \mathbb{R}$$

where the coefficients c_1, c_2 are given by the components of the velocity vector $\frac{d}{dt}|_{t=0}(\mathbf{r} \circ \gamma)$ of the plane curve $(\mathbf{r} \circ \gamma)\colon (-\varepsilon,\varepsilon) \to \mathbb{R}^2$:

$$(c_1, c_2) = \left.\frac{d}{dt}(\mathbf{r} \circ \gamma(t))\right|_{t=0}. \tag{5.7}$$

R Note that both the coordinate basis vectors e_1, e_2 and the coefficients c_1, c_2 depend on the coordinate $\mathbf{r}$. However, their combination $\gamma'(0) = c_1e_1 + c_2e_2$ is independent of the coordinate.

In the introductory multivariable calculus, most formulas are written under a coordinate basis and geometric objects (such as vectors) are written in terms of their corresponding coefficients under the basis. By working directly at the level of tangent spaces T_pM and tangent vectors γ', we can deal with multivariable calculus on manifolds without drowning in coordinate-dependent notations.

5.1.2 Tangent Maps

Let M, W be two manifolds, and let $g\colon M \to W$ be a smooth function. Consider $x \in M$ and $y = g(x) \in Y$. Then the derivative $dg|_x$ of g at x, called the **tangent map** or the **pushforward** by g, is a linear map

$$dg|_x\colon T_xM \to T_yW. \tag{5.8}$$

This linear operator maps "amounts of change" from the domain to "amounts of change" in the range, *i.e.*, it maps the tangent space of the domain to the tangent space of the range. For each tangent vector $v \in T_xM$, which is the velocity of some path $\gamma\colon (-\varepsilon,\varepsilon) \to M$, $\gamma(0) = x$, $\gamma'(0) = v$, the resulting tangent vector $dg|_x(v) \in T_yW$ is defined by

$$dg|_x(v) = dg|_{\gamma(0)}(\gamma'(0)) \coloneqq (g\circ\gamma)'(0), \tag{5.9}$$

which is the velocity of the image $g\circ\gamma$ of the path on W.

▪ **Example 5.1** Let M be a 2-dimensional manifold, and let $f\colon M \to \mathbb{R}^3$ be a surface realization. Then for each $x \in M$, we have a tangent map $df|_x\colon T_xM \to T_{f(x)}\mathbb{R}^3 = \mathbb{R}^3$. The range of $df|_x$ in $\mathbb{R}^3$ is the tangent plane visualized in $\mathbb{R}^3$. If $df|_x$ is everywhere rank-2, then we say f is an immersion (*cf.* Section 4.1.3). ▪

Theorem 5.1 — Chain rule. Suppose $g\colon M_1 \to M_2$ and $h\colon M_2 \to M_3$, $x \in M_1$, $y = g(x) \in M_2$, and $z = h(y) \in M_3$. Then

$$d(h\circ g)|_x = dh|_y \circ dg|_x. \tag{5.10}$$

Roughly speaking, $dg|_x$ is the linear map that best approximates g around x, in the sense that for each $v \in T_xM$ and small ε

$$\text{“}g(x+\varepsilon v) = g(x) + \varepsilon\, dg|_x(v) + O(\varepsilon^2)\text{.”} \tag{5.11}$$

Tangent Maps under Coordinates

Note that it is not clear what the offset operation “$+\,\varepsilon v$” or “$g(x)+\varepsilon\cdots$” means on a manifold (it only makes sense in an affine space like $\mathbb{R}^2, \mathbb{R}^3$). To understand (5.11), one may use local coordinates

$$\begin{aligned} \mathbf{r}\colon U \to \mathbb{R}^m, &\quad U \subset M, &\quad x \in U,\\ \mathbf{s}\colon V \to \mathbb{R}^n, &\quad V \subset W, &\quad y \in V. \end{aligned}$$

By composition, the function g is represented by a map between Cartesian spaces

$$\mathbf{s}\circ g\circ \mathbf{r}^{-1}\colon \mathbf{r}(U) \subset \mathbb{R}^m \to \mathbf{s}(V) \subset \mathbb{R}^n,$$

whose derivative (at $\mathbf{r}(x)$) is a linear map $d(\mathbf{s}\circ g\circ\mathbf{r}^{-1})\colon \mathbb{R}^m \to \mathbb{R}^n$ explicitly represented by the Jacobian matrix

$$d(\mathbf{s}\circ g\circ\mathbf{r}^{-1}) = \begin{bmatrix} \frac{\partial(s_1\circ g\circ\mathbf{r}^{-1})}{\partial r_1} & \frac{\partial(s_1\circ g\circ\mathbf{r}^{-1})}{\partial r_2} & \cdots & \frac{\partial(s_1\circ g\circ\mathbf{r}^{-1})}{\partial r_m} \\ \frac{\partial(s_2\circ g\circ\mathbf{r}^{-1})}{\partial r_1} & \frac{\partial(s_2\circ g\circ\mathbf{r}^{-1})}{\partial r_2} & \cdots & \frac{\partial(s_2\circ g\circ\mathbf{r}^{-1})}{\partial r_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial(s_n\circ g\circ\mathbf{r}^{-1})}{\partial r_1} & \frac{\partial(s_n\circ g\circ\mathbf{r}^{-1})}{\partial r_2} & \cdots & \frac{\partial(s_n\circ g\circ\mathbf{r}^{-1})}{\partial r_m} \end{bmatrix} \tag{5.12}$$

5.2 Metric

Recall that a metric on a discrete triangle mesh is an assignment of lengths on each edge that satisfies the triangle inequality. That structure fixes the Euclidean shape of each triangle, and thus endows the surface a Euclidean metric everywhere except for the vertices.

Before introducing its smooth analog, let us describe what a Euclidean vector space is. A Euclidean vector space is a vector space V together with an inner product structure $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ so that it is

- bilinear: $\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle$ and $\langle w, au + bv \rangle = a\langle w, u \rangle + b\langle w, v \rangle$ for $u, v, w \in V$ and $a, b \in \mathbb{R}$,
- symmetric: $\langle u, v \rangle = \langle v, u \rangle$, and
- positive definite: $\langle u, u \rangle$ (also denoted as $|u|^2$) ≥ 0, and $|u|^2 = 0$ only when $u = 0$.

We call $\langle \cdot, \cdot \rangle$ an **inner product (or dot product) structure**, a **Euclidean structure**, or a **Hilbert structure**. A finite dimensional space equipped with an inner product is called a **Euclidean vector space** or a (finite dimensional) **Hilbert space**.

Definition 5.1 A **metric** on a smooth manifold M is an assignment an inner product operator $\langle \cdot, \cdot \rangle$ on every tangent space. That is, for each $x \in M$, the metric $\langle \cdot, \cdot \rangle : T_xM \times T_xM \to \mathbb{R}$ is a symmetric positive-definite bilinear operator on T_xM. The metric on a manifold is also called a **Riemannian structure**. A manifold equipped with a Riemannian structure is called a **Riemannian manifold**.

R A frequently seen notation is to write $\langle \cdot, \cdot \rangle = g(\cdot, \cdot)$ to emphasize that it is assigning a tensor field g over the manifold. In many classical expositions of differential geometry, the surface theory is written under coordinates, and thus it is important to emphasize that $\langle \cdot, \cdot \rangle$ is not the standard dot product of the Cartesian coordinate.

In our setting, there is no coordinate. There is no notion of universal "standard" dot product structure $\langle \cdot, \cdot \rangle$ to be confused with. Each tangent space T_xM is distinct from any other tangent space T_yM, and there is no way one compares the inner product structures on two different tangent spaces (unless there is a natural linear map between the two spaces).

▪ **Example 5.2 — Induced metric from an immersion.** Let M be a 2D manifold and $f : M \to \mathbb{R}^3$ be an immersion. Note that $\mathbb{R}^3$ is viewed as a Euclidean space, where each $T_{f(x)}\mathbb{R}^3$ is a Euclidean vector space equipped with the standard inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^3}$. This gives rise to an **induced Riemannian metric** $\langle \cdot, \cdot \rangle_f$ on M defined by

$$\langle u, v \rangle_f := \langle df(u), df(v) \rangle_{\mathbb{R}^3}, \quad u, v \in T_xM.$$

▪

▪ **Example 5.3** One may also start with a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ without speaking about a surface realization in 3D. If one finds an immersion $f : M \to \mathbb{R}^3$ so that the induced metric $\langle \cdot, \cdot \rangle_f$ is exactly the same as the prescribed Riemannian structure $\langle \cdot, \cdot \rangle$, then we call f an **isometric immersion**. In general, finding an isometric immersion (or embedding) for a given Riemannian manifold is a difficult problem. ▪

5.2.1 Intrinsic Measurements

A Riemannian structure $\langle \cdot, \cdot \rangle$ on a manifold M allows us to measure various geometric quantities. Obviously we can measure the inner product $\langle u, v \rangle$ of two vectors u, v in the same tangent space, and the norm of a vector $|v| = \sqrt{\langle v, v \rangle}$.

Angle

The inner product $\langle u, v\rangle$ and the angle θ between two vectors $u, v \in T_x M$ are related by $\langle u, v\rangle = |u||v| \cos\theta$. Therefore, the inner product structure $\langle \cdot, \cdot \rangle$ gives rise to the notion of angle between tangent vectors:

$$\theta = \text{angle between } u, v := \cos^{-1}\left(\frac{\langle u, v\rangle}{|u||v|}\right). \tag{5.13}$$

The angle of intersection between two regular curves on the manifold is defined by the angle of their velocity as tangent vectors at the intersection point.

Area

The area of the parallelgram made by two tangent vectors u, v is given by

$$|\det(u, v)| = |u||v| \sin\theta. \tag{5.14}$$

The classical **Gram determinant** formula relates the area and the inner products:

$$|\det(u, v)|^2 = \det\begin{pmatrix}\langle u, u\rangle & \langle u, v\rangle \\ \langle v, u\rangle & \langle v, v\rangle\end{pmatrix}. \tag{5.15}$$

We call $|\det(\cdot, \cdot)|$ the (**unsigned**) **area element**, or sometimes **density**.

If we have in addition an orientable manifold, then we may assign signs to $|\det(u, v)|$ and becomes the skew-symmetric bilinear form

$$\det(u, v) = \pm|\det(u, v)| \begin{cases} > 0 & \text{if } u, v \text{ are positively oriented} \\ < 0 & \text{if } u, v \text{ are negatively oriented} \\ = 0 & \text{if } u, v \text{ are parallel.} \end{cases} \tag{5.16}$$

We call $\det(\cdot, \cdot)$ the **signed area element** or **area form**.

The inner product structure $\langle \cdot, \cdot \rangle$ gives rise to the notion of unsigned area element, as given by (5.15). An inner product structure and an orientation give us the area form.

With the area element, we can compute the area of any region on the manifold. First partition the region into parametrizable patches. The area of the region is the sum of the area of each patch. Suppose a patch U has a coordinate system $\mathbf{r}\colon U \to \mathbf{r}(U) \subset \mathbb{R}^2$; that is, $\mathbf{r}^{-1}\colon \mathbf{r}(U) \to U$ is a parametrization from the $r_1 r_2$-plane to the patch $U \subset M$. Then the area $\text{Area}(U) = \int_U |\det|$ is given by the Jacobian formula

$$\text{Area}(U) = \int_U |\det| = \iint_{\mathbf{r}(U)} \left|\det\left(\frac{\partial \mathbf{r}^{-1}}{\partial r_1}, \frac{\partial \mathbf{r}^{-1}}{\partial r_2}\right)\right| dr_1\, dr_2. \tag{5.17}$$

This area is independent of the coordinate/parametrization.

Here are a few useful notion to keep in mind:

- We write the total area of M as

$$\text{Area}(M) = \int_M |\det| \overset{\text{if } M \text{ is orientable}}{=} \int_M \det,$$

  without writing the Jacobian formula (5.17) in coordinates all the time.

- Suppose $h\colon M \to \mathbb{R}$ is a scalar function. Then we can compute the integral of h (the totality of h over M) by

$$\int_M h\,|\det|, \quad \text{or in the case of an orientable } M \quad \int_M h\,\det. \tag{5.18}$$

- Suppose $L\colon U \to V$ is a linear map between 2D vector spaces U, V. Let U and V be equipped with area form $\det_U(\cdot,\cdot)$, $\det_V(\cdot,\cdot)$. Then we can measure the *determinant of the linear map* L as

$$\det(L) := \frac{\det_V(L(u_1), L(u_2))}{\det_U(u_1, u_2)} \tag{5.19}$$

  for any linearly independent $u_1, u_2 \in U$. The quantity $\det(L)$ measures the amount of area stretching L does, with respect to the given notion of area on U and V.
- Suppose M, W are both orientable 2D Riemannian manifolds, and $g\colon M \to W$ is some general map. We will let $g(M)$ as the oriented-multi-set swept out by the map. We can talk about the oriented area of $g(M)$, which will be denoted and given by

$$\int_{g(M)} \det_W = \int_M \det(dg)\,\det_M.$$

5.3 The Gauss Map

Let M be an oriented 2D manifold, and $f\colon M \to \mathbb{R}^3$ be an immersion into $\mathbb{R}^3$. That is, at each $x \in M$, we have a rank-2 linear map $df|_x\colon T_xM \to \mathbb{R}^3$ that spans a 2D linear subspace in $\mathbb{R}^3$. Concretely, if v_1, v_2 is a positively oriented basis for the vector space T_xM, then $df(T_xM)$ is spanned by $df|_x(v_1)$ and $df|_x(v_2)$. Then, there is a unique *unit* vector $N(x) \in \mathbb{R}^3$ such that $(df(v_1), df(v_2), N)$ makes a positively-oriented basis for $\mathbb{R}^3$, which is given by

$$N(x) = \frac{df|_x(v_1) \times df|_x(v_2)}{\left|df|_x(v_1) \times df|_x(v_2)\right|}.$$

Note that this is independent of the choice of v_1, v_2 as long as they form a positively oriented basis for T_xM.

Definition 5.2 This vector $N(x)$ is the **normal vector** of the immersed surface $f\colon M \to \mathbb{R}^3$ at $x \in M$. As a map $N\colon M \to \mathbb{S}^2$ (that takes in a point $x \in M$ and outputs a unit $\mathbb{R}^3$ vector $N(x)$) it is also called the **Gauss map** of the immersed surface.

A flat plane in 3D is a surface whose Gauss map is a constant map (that takes a constant value on $\mathbb{S}^2$, the normal of the plane). For a curved surface the Gauss map $N(x)$ takes different directions as we vary $x \in M$. The faster N changes the more curved the surface is.

By studying the rate of change of the value of the Gauss map $N\colon M \to \mathbb{S}^2$ we define curvature for the immersed surface.

5.4 The Shape Operator

Let $N\colon M \to \mathbb{S}^2$ be the Gauss map (normal vector field) of a surface immersed in 3D, $f\colon M \to \mathbb{R}^3$. The tangent map of N at each $x \in M$ is a linear map

$$dN|_x\colon T_xM \to T_{N(x)}\mathbb{S}^2. \tag{5.20}$$

Note that $T_{N_x}\mathbb{S}^2$ is a plane in $\mathbb{R}^3$ that is parallel to tangent plane $df(T_xM)$ of the surface (as they are both normal to N). Thus, by identifying $T_{N_x}\mathbb{S}^2$ with $df(T_xM)$ we may also write

$$dN|_x \colon T_xM \to df(T_xM). \tag{5.21}$$

Here is a way to picture how dN taking value in the tangent plane. If we walk from $x \in M$ in a direction $v \in T_xM$, we observe that the normal vector N is tilting at $dN|_x(v)$. Since N is always a normal vector, infinitesimally this rate of change of N must tilts tangentially.

Each vector in the tangent plane $df(T_xM)$, such as $dN|_x(v)$, must be $df(u)$ for some vector $u \in T_xM$ in the abstract tangent space T_xM of the 2D manifold M. This gives us the following:

Definition 5.3 The **shape operator** (also known as the **Weingarten operator**) at each point $x \in M$ for an immersion $f \colon M \to \mathbb{R}^3$ is a linear operator on the tangent space

$$S_x \colon T_xM \to T_xM \tag{5.22}$$

that represents the derivative of the Gauss map

$$dN|_x(v) = df(S_x(v)) \quad \text{for all } v \in T_xM. \tag{5.23}$$

The shape operator encapsulates all the curvature informations.

Theorem 5.2 — Symmetry of the shape operator. The shape operator $S_x \colon T_xM \to T_xM$ is self-adjoint with respect to the induced metric $\langle \cdot, \cdot \rangle_f$. That is,

$$\langle u, S_x(v) \rangle_f = \langle v, S_x(u) \rangle_f, \quad u, v \in T_xM; \tag{5.24}$$

equivalently,

$$\langle df(u), dN(v) \rangle_{\mathbb{R}^3} = \langle df(v), dN(u) \rangle_{\mathbb{R}^3}, \quad u, v \in T_xM. \tag{5.25}$$

Proof. We will postpone the proof until we learn the basic operations in exterior calculus. With exterior calculus, we have a one-liner proof: $\langle df, N \rangle = 0 \implies d\langle df, N \rangle = -\langle df \wedge dN \rangle = 0$ which means $\langle df(u), dN(v) \rangle - \langle df(v), dN(u) \rangle = 0$ for all $u, v \in T_xM$. □

5.4.1 Normal Curvature

Definition 5.4 Let $f \colon M \to \mathbb{R}^3$ be an immersed surface, N be the Gauss map, and S denote the shape operator. Let $\langle \cdot, \cdot \rangle_f$ be the induced metric on M. Let $x \in M$ and $v \in T_xM$, $|v|_f = 1$. The **normal curvature** $\kappa_n(x, v)$ at x in the direction of v is given by

$$\kappa_n(x, v) \coloneqq \langle v, S_x(v) \rangle_f = \langle df(v), dN(v) \rangle_{\mathbb{R}^3}. \tag{5.26}$$

Here is the geometric picture of the normal curvature. Taking the tangent vector $df|_x(v)$ and the normal vector $N(x)$, construct a normal plane P_x spanned by them

$$P_{x,v} \colon \text{ 2D plane in } \mathbb{R}^3 \text{ that contains } f(x), f(x) + df|_x(v), \text{ and } f(x) + N(x).$$

This normal plane intersects the surface $f(M)$ in $\mathbb{R}^3$ in a curve that is regular around $f(x)$. Curvature of this curve is the **normal curvature** $\kappa_n(x, v) \in \mathbb{R}$. Let us call this intersection curve γ,

which is a space curve. Note that this intersection curve is also a plane curve that is contained in $P_{x,v}$, and that is why its curvature can be described by a real number instead of a complex number or a 3D vector. We take the sign so that $\kappa_n > 0$ if $\gamma''(0)$ is aligned with $-N(x)$ and $\kappa_n < 0$ if $\gamma''(0)$ is aligned with $N(x)$.

5.4.2 Principal Curvatures

Observe from (5.26) that the normal curvature $\kappa_n(x, v)$ at a fixed x is a quadratic function of $v \in T_xM$. If we sweep over the circle $|v|_f^2 = 1$ in T_xM, we will see a maximum $\kappa_n(x, v)$ and a minimum $\kappa_n(x, v)$. By the theory of quadratic forms and Rayleigh quotients in linear algebra, these extrema happen at the eigenvectors of S_x and the extremal value are the eigenvalues of S_x:

$$\kappa_1(x) = \min_{v \in T_xM, |v|_f^2 = 1} \langle v, S_x(v) \rangle_f, \quad v_1(x) = \operatorname*{argmin}_{v \in T_xM, |v|_f^2 = 1} \langle v, S_x(v) \rangle_f \tag{5.27}$$

$$\kappa_2(x) = \max_{v \in T_xM, |v|_f^2 = 1} \langle v, S_x(v) \rangle_f, \quad v_2(x) = \operatorname*{argmax}_{v \in T_xM, |v|_f^2 = 1} \langle v, S_x(v) \rangle_f \tag{5.28}$$

$$\iff \quad S_x v_i(x) = \kappa_i(x) v_i(x), \quad i = 1, 2, \quad \kappa_2 \geq \kappa_1. \tag{5.29}$$

Definition 5.5 The eigenvalues $\kappa_1, \kappa_2 \in \mathbb{R}$ of the shape operator S are called the **principal curvatures**. The corresponding normalized eigenvectors $v_1, v_2 \in T_xM$ are called the **principal directions**.

Since the shape operator S is self-adjoint with respect to $\langle \cdot, \cdot \rangle_f$, the eigenvalues are real and the eigenvectors are orthonormal with respect to $\langle \cdot, \cdot \rangle_f$.

Theorem 5.3 Let $v \in T_xM$, $|v|_f^2 = 1$. Represent v in terms of the principal directions as $v = \cos(\theta) v_1 + \sin(\theta) v_2$ for some angle θ. Then

$$\kappa_n(x, v) = \cos^2(\theta) \kappa_1 + \sin^2(\theta) \kappa_2. \tag{5.30}$$

Next we look at the natural quantities (invariants) of the shape operators. They are all functions of the principal curvatures. Specifically, we will look at $(\kappa_1 + \kappa_2)/2$ which measures the normal curvature in average, $(\kappa_2 - \kappa_1)$ that measures the deviation from uniformly curved, and $\kappa_1 \kappa_2$ which measures the area stretched by the Gauss map.

5.4.3 Mean Curvature and Hopf Differential

Definition 5.6 The **mean curvature** $H \in \mathbb{R}$ at a point on an immersed surface is given by

$$H := \frac{1}{2} \operatorname{tr}(S) = \frac{\kappa_1 + \kappa_2}{2}. \tag{5.31}$$

The trace-free part (deviator part) of the shape operator is called the **Hopf differential**

$$Q := S - H \operatorname{id}. \tag{5.32}$$

Here $\operatorname{id}: T_xM \to T_xM$ is the identity map at each $x \in M$.

Theorem 5.4 The mean curvature is also the normal curvature averaged over all directions: Let $v_\theta = \cos(\theta)e_1 + \sin(\theta)e_2$ where e_1, e_2 are an orthonormal basis of T_xM. Then

$$\frac{1}{2\pi}\int_0^{2\pi} \kappa_n(x, v_\theta)\, d\theta = H. \tag{5.33}$$

Proof. Without loss of generality, take $e_1 = v_1$ and $e_2 = v_2$ where v_1, v_2 are the principal directions. (This change of orthogonal basis only amounts to a translation in the angle $\theta \mapsto \theta - \theta_0$ which does not change the integral of interest.) Applying (5.30), we have

$$\frac{1}{2\pi}\int_0^{2\pi} \kappa_n(x, v_\theta)\, d\theta = \frac{\kappa_1}{2\pi}\int_0^{2\pi} \cos^2(\theta)\, d\theta + \frac{\kappa_2}{2\pi}\int_0^{2\pi} \sin^2(\theta)\, d\theta = \frac{\kappa_1 + \kappa_2}{2}.$$

□

In general, the normal curvature depends on the direction. The Hopf differential measures the deviation of this curvature from being constant.

Definition 5.7 A point $x \in M$ of an immersed surface is called an **umbilical point** if $Q(x) = 0$. That is, $\kappa_1 = \kappa_2$, or $\kappa_n(x, v)$ is independent of v, at that point.

▪ **Example 5.4** A sphere of radius r has mean curvature $H = 1/r$ at every point. Every point on a sphere is an umbilical point. In fact the converse is also true: if every point in an open set (region) of a surface is umbilical, then that region must be a portion of a round sphere. ▪

At an non-umbilical point, the Hopf differential has eigenvalues $\pm\frac{\kappa_1-\kappa_2}{2}$ and the same set of eigenvectors v_1, v_2 as that of S. One reconstructs the entire shape operator by providing the mean curvature, the magnitude $\frac{\kappa_1-\kappa_2}{2}$ and the axis v_1, v_2 of the Hopf differential.

▪ **Example 5.5** In an eye-glasses prescription, one would find the *spherical*, *cylindrical* and *axis* data of each eye. The spherical data corresponds to the mean curvature, and the other two characterizes the Hopf differential operator. The mean curvature of the lens describes the degree of near/far-sightedness it corrects. The Hopf differential of the lens is to correct astigmatism. ▪

5.4.4 Gaussian Curvature

Definition 5.8 The **Gaussian curvature** $K \in \mathbb{R}$ at a point on an immersed surface $f : M \to \mathbb{R}^3$ is given by

$$K := \det(S) = \det(dN) = \kappa_1\kappa_2. \tag{5.34}$$

This definition is the smooth analog of (4.22). For each region $U \subset M$, we find a corresponding image $N(U)$ on $\mathbb{S}^2$ swept out by the normal vector (Gauss map). If the immersed surface is a flat plane, then N maps U to a single point. If the immersed surface bulges or dents like a portion of a sphere, then N maps U to a nontrivial area in an orientation preserving fashion. If the immersed surface is shaped like a saddle in U, then N maps U to a nontrivial area in an orientation reversing way. This signed area of the (signed-multi-)set $N(U)$ is given by

$$\int_{N(U)} \det_{\mathbb{S}^2} = \int_U \det(dN)\, \det_f = \int_U K\, \det_f. \tag{5.35}$$

Here $\det_{\mathbb{S}^2}$ is the area form on the unit sphere, and $\det_f$ is the induced area form (*i.e.* the area form associated with the metric $\langle \cdot, \cdot \rangle_f$ induced by the immersion f) on M.

Another way to phrase (5.34) is that the Gaussian curvature K at a point x is the limit of the ratio of the signed area of $N(U)$ to the area of U as the region U shrinks to the point x:

$$K = \lim_{U \searrow \{x\}} \frac{\text{SignedArea}(N(U))}{\text{Area}(U)}. \tag{5.36}$$

The formula $K = \kappa_1 \kappa_2$ gives us a good sense of the value of the Gaussian curvature. The maximal and minimal normal curvature κ_1, κ_2 have

- the same sign if only if $K > 0$ (*e.g.* convex or concave dome);
- the opposite signs if only if $K < 0$ (*e.g.* saddle);
- one or both of them zero if only if $K = 0$ (*e.g.* cylinder, cone, plane, or any transitional locus between positive and negative Gaussian curvature regions on a surface).

5.5 Gauss–Bonnet Theorem

We have already seen the remarkable facts about the Gaussian curvature in the discrete setting:

- (Definition 4.9 and Theorem 4.10) The area of the Gauss map is the angle defect. The former is defined using the immersion and the normal vector, and surprisingly the latter only depends on the intrinsic metric.
- (Theorem 4.8) The turning angle of a curve on a surface is 2π minus the total Gaussian curvature on the disk enclosed by the curve.
- (Definition 4.11 and Theorem 4.9) The parallel transport of a vector along a closed curve on a surface will acquire a holonomy angle, which is given by the total Gaussian curvature enclosed by the curve.

The smooth analog of these theorems, which we list below, are best understood by the discrete version (Section 4.6) with concrete geometric proofs. This contrasts the classical exposition of these theorems, where one typically performs calculation in a coordinate chart, with the introduction of coordinate dependent numbers such as the Christoffel symbols.

5.5.1 Parallel Transport and the Covariant Derivative of Tangent Vectors

First let us talk about the **parallel transport** of a tangent vector along a path.

Here is parallel transport in a nutshell. Each tangent space $T_x M$ is a different vector space from any other tangent space $T_y M$, $x \neq y$. Unlike the case of $M = \mathbb{R}^2$ where we can parallel translate a vector from $T_x M$ to $T_y M$, for a general surface M there is no canonical linear map $T_x M \to T_y M$. What we can do instead is to consider a path $\gamma : [0,1] \to M$ that travels from x to y ($\gamma(0) = x, \gamma(1) = y$), and infinitesimally transport the vector along the path such that the vector does not rotate in the tangent plane during the transportation. This is called the parallel transport of a tangent vector at $T_x M$ to $T_y M$ along a path. The measurement of this rotation, called the **covariant derivative** of a vector field along the path, is solely a function of the inner product structure $\langle \cdot, \cdot \rangle$ assigned on each tangent space. That is, the notion of the parallel transport is an intrinsic notion.

This idea is much more transparent in the discrete setting. Recall that in the discrete setting, we can parallel transport a vector from one triangle to a adjacent triangle. The parallel transport

from a triangle to a distant triangle depends on the path (triangle strip) along which the vector is parallel transported.

Notation 5.1 — Tangent vector field. A tangent vector field u on M is assigning a vector $u(x) \in T_x M$ for each $x \in M$ (with $u(x)$ being sufficiently smooth with respect to varying x in any coordinate chart). The space of all such tangent vector fields is denoted by $\Gamma(TM)$. We write $u \in \Gamma(TM)$ and say u is a (smooth) tangent vector field.

Notation 5.2 — Tangent vector field along a curve. Let $\gamma : I \to M$ be a curve on M. A family of tangent vectors of M defined along the curve is denoted by $u \in \Gamma(T_\gamma M)$, which means that $u(t) \in T_{\gamma(t)} M$. The object $T_\gamma M$ is a collection of tangent spaces $T_{\gamma(t)} M$ parametrized by $t \in I$, *i.e.* $(T_\gamma M)_t = T_{\gamma(t)} M$.

The notation $T_\gamma M$ in Notation 5.2 is not so commonly adopted. The standard notation in differential geometry literature is to write $\gamma^* TM$ instead of $T_\gamma M$, and to write $u \in \Gamma(\gamma^* TM)$ instead of $u \in \Gamma(T_\gamma M)$. One calls $\gamma^* TM$ a pullback vector bundle.

In the following, we concretely define covariant derivatives and parallel transportation on an immersed surface $f : M \to \mathbb{R}^3$.

Definition 5.9 — Covariant derivative. Let $\gamma : [0,1] \to M$ be a path. Let $u \in T_\gamma M$ be a tangent vector field defined along the curve. The **covariant derivative** of u with respect to the parameter t of the curve is denoted by

$$\frac{\nabla}{dt} u(t) \in T_{\gamma(t)} M, \tag{5.37}$$

which is given by the derivative of the $\mathbb{R}^3$ vector (via the tangent map of the immersion) followed by the orthogonal projection $P_{df(T_{\gamma(t)}M)} : \mathbb{R}^3 \to df(T_{\gamma(t)} M)$ to the tangent plane $df(T_{\gamma(t)} M) \subset \mathbb{R}^3$:

$$df\left(\frac{\nabla}{dt} u(t)\right) := P_{df(T_{\gamma(t)}M)} \frac{d}{dt}\Big(df\big(u(t)\big)\Big). \tag{5.38}$$

The covariant derivative is the tangent-space-component of the rate of change of the vector field when realized in 3D. In other words,

$$\frac{d}{dt}\Big(df\big(u(t)\big)\Big) = \underbrace{df\left(\frac{\nabla}{dt} u(t)\right)}_{\text{tangent component}} + \underbrace{\left\langle \frac{d}{dt}\Big(df\big(u(t)\big)\Big), N \right\rangle N}_{\text{normal component}}. \tag{5.39}$$

In fact, we already know what this normal component is! Observe that since $df(u)$ is orthogonal to the normal vector, we have $\langle df(u(t)), N \rangle = 0$ for all t; taking the derivative of this condition yields

$$\begin{aligned} 0 &= \frac{d}{dt} \langle df(u(t)), N(\gamma(t)) \rangle = \left\langle \frac{d}{dt}\Big(df\big(u(t)\big)\Big), N \right\rangle + \left\langle df\big(u(t)\big), \frac{d}{dt}\Big(N(\gamma(t))\Big) \right\rangle \\ &= \left\langle \frac{d}{dt}\Big(df\big(u(t)\big)\Big), N \right\rangle + \langle u(t), S_{\gamma(t)} \gamma'(t) \rangle \end{aligned} \tag{5.40}$$

where S is the shape operator. Note that since S is self-adjoint, $\langle u, S\gamma' \rangle = \langle \gamma', Su \rangle$. Therefore, we can rewrite the normal component of (5.39) in terms of the shape operator:

Theorem 5.5 Let $u \in T_\gamma M$ be a vector field defined along a curve $\gamma : I \to M$. Let $f : M \to \mathbb{R}^3$ be an immersion. Then the tangenti and normal component of the derivative of the 3D vector $df(u(t))$ is given by

$$\frac{d}{dt}\Big(df\big(u(t)\big)\Big) = df\left(\frac{\nabla}{dt}u(t)\right) - \langle \gamma'(t), Su(t) \rangle N. \tag{5.41}$$

Definition 5.10 Suppose $u \in \Gamma(TM)$ is a vector field defined over M. Let $v \in T_x M$ be a tangent vector at one point x. Then the **directional covariant derivative** of u in the direction v is a vector denoted by

$$\nabla_v u \in T_x M \tag{5.42}$$

defined by

$$\nabla_v u := \frac{\nabla}{dt}u(t)\Big|_{t=0} \tag{5.43}$$

using any curve $\gamma : (-\varepsilon, \varepsilon) \to M$ with $\gamma(0) = x$ and $\gamma'(0) = v$, and by setting $u(t) = u(\gamma(t))$.

R The operator $\nabla_{(\cdot)}(\cdot)$ takes in the first slot a vector at a point, and in the second slot a vector field, and returns a vector at a point. This covariant derivative operator is also called the **Levi-Civita connection**. It satisfies the following properties:

- Linearity in the direction: $\nabla_{c_1 v_1 + c_2 v_2} u = c_1 \nabla_{v_1} u + c_2 \nabla_{v_2} u_2$.
- Additivity for the vector field: $\nabla_v(u_1 + u_2) = \nabla_v u_1 + \nabla_v u_2$.
- Product rule for the vector field multiplied with a scalar function: $\nabla_v(hu) = (d_v h)u + h\nabla_v u$, where $h : M \to \mathbb{R}$. (Here $d_v h = dh(v)$ denotes the directional derivative of a scalar function h.)
- Compatibility with metric: $d_v\langle u_1, u_2 \rangle = \langle \nabla_v u_1, u_2 \rangle + \langle u_1, \nabla_v u_2 \rangle$.
- Torsion free: If $\mathbf{r} : U \subset M \to \mathbb{R}^2$ is a coordinate, and $u_1 = \frac{\partial \mathbf{r}^{-1}}{\partial r_1}, u_2 = \frac{\partial \mathbf{r}^{-1}}{\partial r_2}$ are coordinate vector fields, then $\nabla_{u_1} u_2 = \nabla_{u_2} u_1$.

Any operator that satisfies the first three properties is called a **connection**. It turns out that given a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ there exists a unique operator $\nabla_{(\cdot)}(\cdot)$ that satisfies all five properties; such an operator is the Levi-Civita connection of the Riemannian manifold. This is known as the **Fundamental Theorem of Riemannian Geometry**.

The moral of the story is that there is a canonical way of measuring the directional derivative of a vector field on a Riemannian manifold, and it is denoted by $\nabla_v u$.

Definition 5.11 — Parallel transport. Let $u \in T_x M$, and $\gamma : [0,1] \to M$ be a path from $\gamma(0) = x$ to $\gamma(1) = y$. Then the parallel transport of u along the path γ is to extend u to a

vector field $\widetilde{u} \in \Gamma(T_\gamma M)$ along the path such that

$$\frac{\nabla}{dt} u(t) = 0 \quad \text{for all } t \in [0,1]. \tag{5.44}$$

We write

$$\mathcal{P}_\gamma \colon T_x M \to T_y M, \quad \mathcal{P}_\gamma u := \widetilde{u}(t=1). \tag{5.45}$$

R The parallel transport operator $\mathcal{P}_\gamma$ depends on the path γ, but is independent of the parametrization of the curve γ. The operator $\mathcal{P}_\gamma \colon T_{\gamma(0)}M \to T_{\gamma(1)}M$ is a linear map. The parallel transport linear map also preserves metric (*i.e.* it is a linear isometry): $\langle u_1, u_2 \rangle_{T_{\gamma(0)}M} = \langle \mathcal{P}_\gamma u_1, \mathcal{P}_\gamma u_2 \rangle_{T_{\gamma(1)}M}$.

5.5.2 Geodesic Curvature

The covariant derivative of tangent vector field allow us to measure the rate of change of the vector field within the tangent space. Using the covariant derivative, we can measure the curvature of a curve on the manifold similar to that of plane curves.

Definition 5.12 — Geodesic curvature. Let $\gamma \colon I \to M$ be a regular curve on an oriented Riemannian manifold M. Let ds denote the arclength measure of γ. That is, $d\gamma/ds \in \Gamma(T_\gamma M)$ is the unit tangent vector of the curve. Let $\mathbb{i} \colon T_x M \to T_x M$ denote the 90° counterclockwise rotation in each tangent space. Define the **geodesic curvature** κ_g of γ as

$$\kappa_g = \left\langle \frac{\nabla}{ds}\frac{d\gamma}{ds}, \mathbb{i}\frac{d\gamma}{ds} \right\rangle, \quad \text{equivalently} \quad \frac{\nabla}{ds}\frac{d\gamma}{ds} = \kappa_g \mathbb{i}\frac{d\gamma}{ds}. \tag{5.46}$$

The integral

$$\int_I \kappa_g \, ds \tag{5.47}$$

is the total **turning angle** of the curve.

Recall that in the case of an immersed surface $f \colon M \to \mathbb{R}^3$, the covariant derivative is the tangential component of the derivative taken in $\mathbb{R}^3$. What this means is the following: via the immersion, the curve γ is placed in 3D as a space curve $f \circ \gamma$; the geodesic curvature is the tangent component of the Frenet curvature normal of this space curve

$$\frac{d}{ds}\frac{d(f\circ\gamma)}{ds} = \underbrace{\kappa_g N \times \frac{d(f\circ\gamma)}{ds}}_{\text{tangential part}} + \underbrace{\left\langle \frac{d}{ds}\frac{d(f\circ\gamma)}{ds}, N \right\rangle N}_{\text{normal part}} \tag{5.48}$$

Here, N is the Gauss map of the surface, not the Frenet normal of the space curve. We have used $df(\mathbb{i}\, d\gamma/ds) = N \times df(d\gamma/ds) = N \times d(f\circ\gamma)/ds$. By a similar calculation as (5.40), we find that

the normal component is given by the normal curvature in the $d\gamma/ds$ direction:

$$\begin{aligned}\left\langle \frac{d}{ds}\frac{d(f\circ\gamma)}{ds}, N\circ\gamma \right\rangle &= -\left\langle \frac{d(f\circ\gamma)}{ds}, \frac{d(N\circ\gamma)}{ds} \right\rangle \\ &= -\left\langle df\left(\frac{d\gamma}{ds}\right), dN\left(\frac{d\gamma}{ds}\right) \right\rangle = -\left\langle \frac{d\gamma}{ds}, S\frac{d\gamma}{ds} \right\rangle \\ &= -\kappa_{\mathrm{n}}\left(\gamma(s), \frac{d\gamma}{ds}\right). \end{aligned} \tag{5.49}$$

Hence:

Theorem 5.6 Let $\gamma : I \to M$ be a regular curve on an oriented surface M immersed in 3D $f : M \to \mathbb{R}^3$. Then the space curve $(f \circ \gamma)$ has Frenet curvature normal $d^2(f\circ\gamma)/ds^2 \in \mathbb{R}^3$ whose tangential and normal components are

$$\frac{d^2(f\circ\gamma)}{ds^2} = \kappa_{\mathrm{g}} N \times \frac{d(f\circ\gamma)}{ds} - \kappa_{\mathrm{n}}\left(\gamma, \frac{d\gamma}{ds}\right) N. \tag{5.50}$$

In particular, the Frenet curvature κ of the space curve is

$$\kappa = \sqrt{\kappa_{\mathrm{g}}^2 + \kappa_{\mathrm{n}}(\gamma, d\gamma/ds)^2} \tag{5.51}$$

5.5.3 Gauss–Bonnet Theorems

As the smooth analog of Theorem 4.8, we have:

Theorem 5.7 — Local Gauss–Bonnet Theorem. Let $U \subset M$ be a disk on an oriented Riemannian manifold M, and let γ be the boundary curve of U. Then the total Gaussian curvature on U is 2π minus the total turning angle of the curve:

$$\int_U K \ \det = 2\pi - \oint_\gamma \kappa_{\mathrm{g}}\, ds. \tag{5.52}$$

As the smooth version of the Descartes Theorem (Exercise 4.6) we have:

Theorem 5.8 — Global Gauss–Bonnet Theorem. Let M be a closed (without boundary) oriented Riemannian manifold. Then

$$\int_M K \ \det = 2\pi\chi(M) \tag{5.53}$$

where $\chi(M)$ is the Euler characteristic of M.

As the smooth counterpart of Theorem 4.9, we have:

Theorem 5.9 — Gauss–Bonnet Theorem and the Holonomy. Let $U \subset M$ be a disk on an oriented Riemannian manifold M, and let $\gamma : [0,1] \to \partial U$ be the boundary curve of U. Note that $\gamma(0) = \gamma(1)$. Let $\mathcal{P}_\gamma : T_{\gamma(0)}M \to T_{\gamma(1)}M = T_{\gamma(0)}M$ be the parallel transport map along the loop. Since $\mathcal{P}_\gamma$ is a linear isometry from $T_{\gamma(0)}M$ to itself, it is a rotation operator

$\mathcal{P}_\gamma = e^{\mathbb{i}\theta}$ for some angle $\theta \in \mathbb{T}^1$. Then

$$\theta = \int_U K \det \qquad \mod 2\pi. \tag{5.54}$$

▪ **Example 5.6 — Foucault pendulum.** The Foucault pendulums are large and heavy pendulums displayed in several science museums over the world. They swing like a usual pendulum in a vertical plane. After several hours (say a day), this vertical plane has also rotated a certain angle. This is an example of Theorem 5.9. The direction of the pendulum is a tangent vector of the earth. The rotation of the globe brings this device along the circle of constant latitude (λ° counted from the equator), and in the least dissipative manner (adiabatic process) the tangent vector (swing plane of the pendulum) is parallel transported along this circle. After a day, the path has closed up and enclosed an area with total Gaussian curvature

$$\int_{\text{latitude}\geq\lambda^\circ} K \det = \frac{1}{2}\left(90^\circ - \lambda^\circ - \frac{1}{2}\sin(2\lambda^\circ)\right) \tag{5.55}$$

which will be the holonomy one observes after a day. Since the whole museum is rotated with the earth, what we actually observe is the total turning angle, given by 360° minus the value of (5.55). The resulting angle divided by 24 is the total angle the pendulum appears to rotate each hour. ▪

Finally, the smooth analog of Corollary 4.11 is

Theorem 5.10 — Gauss's Theorema Egregium. The Gaussian curvature $K = \det(dN)$ depends only on the intrinsic metric $\langle\cdot,\cdot\rangle$. That is, any isometric deformation of the surface must leave the Gaussian curvature invariant.

▪ **Example 5.7** Paper is made of an isometric material. That is, any deformation of a piece of paper (such as bending and folding the piece of paper) is isometric (doesn't change the induced metric). By Theorem 5.10 the Gaussian curvature must remain the same before and after the deformation. A standard flat piece of paper has $K = 0$. Therefore, the deformed piece of paper, though seemingly curved, still has $K = 0$. That is, if the maximal normal curvature is $\kappa_2 > 0$, then in order to have $K = \kappa_1\kappa_2 = 0$ we must have $\kappa_1 = 0$ in its orthogonal direction! Examples include a piece of cylinder or a piece of cone. In general, these surfaces with $K = 0$ are called **developable surfaces**. ▪

II Exterior Calculus

6. Exterior Calculus

Throughout the notes, we assume that the reader is familiar with basic multivariable calculus as well as linear algebra. So what about this **exterior calculus**? Why do we need this special flavor of calculus? The main motivation for us is that we want a concise notational apparatus to talk about differential and integral calculus on manifolds. Under this framework, one carefully distinguishes coordinate-dependent multivariable quantities, and directly works with their geometrically meaningful, coordinate-free representations. This exposes the underlying structures far better than piles of indices pointing to various coordinate components. One may write down equations that relate these geometric measurements, say in a continuum, at a higher level that is closer to our physical and geometric intuition. In many cases, even without having the domain to be set on general manifolds, one finds much more transparent and direct geometric reasoning to phenomena in a physical system governed by these differential equations. We will take a glimpse into a few of these differential equations that have played fundamental roles in geometry processing as well as in physical simulations.

6.1 From Multivariable Calculus to Exterior Calculus

When it comes to multivariable integration in the introductory calculus, there are two important topics: change of variables (change of coordinates) formula, and integral theorems namely the Green, Gauss, Stokes theorems. The main motivation of exterior calculus is to work in a framework that the change of variables are automatic, and all the Stokes-like theorems are unified.

Here is a quickly recap of these topics in multivariable calculus. We follow one of the very first written textbooks on these topics, Maxwell's 1873 "a Treatise on Electricity and Magnetism, Vol. 1, Preliminary." At the dawn of vector calculus, Maxwell has made several remarks indicating that the natural structure to unify multivariable calculus is to use the skew-symmetric structure of differential forms. What exterior calculus is about is the linear algebra of these skew-symmetric forms.

6.1.1 Line and Surface Integral

An important operation is the integration of the component of a vector field projected along a line or a curve. This is called the **circulation** of a vector field along a curve (not necessarily closed).

Let $\gamma = (x_1, x_2, x_3)\colon [0, S] \to \mathbb{R}^3$ be a curve, and $\mathbf{v} = (v_1, v_2, v_3)\colon \mathbb{R}^3 \to \mathbb{R}^3$ a vector field. Then the circulation of the curve is given by

$$\begin{aligned} C &= \int_0^S \langle \mathbf{v}, \gamma' \rangle \, ds \quad \text{which can be expressed as} \\ &= \int_0^S v_1 \frac{dx_1}{ds}\, ds + v_2 \frac{dx_2}{ds}\, ds + v_3 \frac{dx_3}{ds}\, ds \quad \text{or simply} \\ &= \int_0^S \langle \mathbf{v}, d\gamma \rangle = \int_{\gamma([0,S])} v_1\, dx_1 + v_2\, dx_2 + v_3\, dx_3 \end{aligned}$$

Here $v_1\, dx_1 + v_2\, dx_2 + v_3\, dx_3$ is called a **1-form**. This quantity C with integral $\int_{\gamma([0,S])}$ generally depends on the entire path $\gamma([0, S])$, as opposed to an integral $\int_{\mathbf{a}}^{\mathbf{b}}$ depending only on the two end points $\mathbf{a} = \gamma(0), \mathbf{b} = \gamma(S)$. However, when within a certain region the one-form takes the form of

$$v_1\, dx_1 + v_2\, dx_2 + v_3\, dx_3 = d\Psi \quad \text{for some potential function } \Psi$$

that is, is an exact differential within that region, the value of C becomes

$$C = \Psi(\mathbf{b}) - \Psi(\mathbf{a})$$

and is the same for any two paths between $\mathbf{a}$ and $\mathbf{b}$, provided the path can be changed into the other by continuous motion without passing out of this region.[1]

Another important operation that involves a surface integral is the **flux** of a vector field through a surface. Let Σ be a surface, dA the surface area element,[2] and $\mathbf{n}$ the unit normal to the surface drawn towards the positive side of the surface. The flux of $\mathbf{v}$ over the surface is given and expressed by

$$\begin{aligned} \iint_\Sigma \langle \mathbf{v}, \mathbf{n} \rangle \, dA &= \iint_\Sigma v_1 n_1\, dA + \iint_\Sigma v_2 n_2\, dA + \iint_\Sigma v_3 n_3\, dA \\ &= \iint_\Sigma v_1\, dx_2\, dx_3 + \iint_\Sigma v_2\, dx_3\, dx_1 + \iint_\Sigma v_3\, dx_1\, dx_2. \end{aligned}$$

[1] The necessary condition for the 1-form $v_1\, dx_1 + v_2\, dx_2 + v_3\, dx_3$ to be an exact differential of a potential is

$$\operatorname{curl} \mathbf{v} = \left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right) = (0, 0, 0).$$

However, there are cases in which the condition $\operatorname{curl} \mathbf{v} = 0$ is fulfilled throughout a domain, and yet the line integral C from $\mathbf{a}$ to $\mathbf{b}$ may be different for two curves. This may be the case if the domain is multiply connected, *i.e.* there exists closed loop that cannot deform and shrink to a point (nontrivial **cycles**), and if the two paths from $\mathbf{a}$ and $\mathbf{b}$ follow the opposite segments of such nontrivial loops. In this case, the one path cannot be transformed into the other. Two cycles are said to be equivalent if they together form the boundary of a surface embedded in the space. In a multiply connected domain, every cycle (up to equivalence) is a multiple, or a linear combination, of finitely many basis cycles.

If $\operatorname{curl} \mathbf{v} = 0$ and the circulations along those basis cycles vanish, then $v_1\, dx + v_2\, dx_2 + v_3\, dx_3 = d\Psi$.

[2] Here dA does not mean the derivative d of a function A.

The quantity $v_1\,dx_2\,dx_3 + v_2\,dx_3\,dx_1 + v_3\,dx_1\,dx_2$ is called a **2-form**.

Here $dx_1\,dx_2$ is the area element projected to the x_1x_2-plane. The flux of a vector field through a surface is the sum of its contribution from all three projected components, each of which is the area integral of the vector component normal to the respective plane. It is important to note that these area integrals are signed integral inherited from the orientation of the space. The normal of the $dx_2\,dx_3$-plane is x_1, the normal of the $dx_3\,dx_1$-plane is x_2, *etc*. In particular, when writing flux integral in this form, the ordering of differentials matter $\iint v_1\,dx_2\,dx_3 = -\iint v_1\,dx_3\,dx_2$.

Similarly, in this context, a volumetric integral $\iiint h\,dx_1\,dx_2\,dx_3$ should have the ordering of the differential $dx_1\,dx_2\,dx_3$ positively oriented.

More generally, it turns out that the right structure for integration that evaluate circulations, fluxes, *etc.*, is infinitesimally a **skew symmetric** product of differentials $dx_{i_1}\cdots dx_{i_k}$. Later, in order not to be confused with the unsigned integrals, we shall use the notation $dx_{i_1}\wedge\cdots\wedge dx_{i_k}$ to emphasize the skew symmetry nature. We will spend Section 6.3 and Section 6.4 to study the linear algebra of these forms.

Changing Variables for Multivariable Integrals

Suppose D is a domain of in the Cartesian space $\mathbb{R}^n$ with coordinates labeled $x_1,\dots,x_n$. Let $F\colon D\to\mathbb{R}^n$ be a map into another Cartesian space with coordinates labeled $y_1,\dots,y_n$. This map parametrizes the image $F(D)$ (as a signed multiset) with $y_i = F_i(x_1,\dots,x_n)$ for $i = 1,\dots,n$. Let $h\colon F(D)\to\mathbb{R}$ be a scalar function. Then the integral

$$\int_{F(D)} h(y_1,\dots,y_n)\,dy_1\cdots dy_n$$

can be written in terms of an integral over $(x_1,\dots,x_n)\in D$ as

$$\int_{F(D)} h(y_1,\dots,y_n)\,dy_1\cdots dy_n = \int_D (h\circ F)(x_1,\dots,x_n)J\,dx_1\cdots dx_n,$$

where J is the Jacobian determinant:

$$J = \det\begin{bmatrix} \partial F_1/\partial x_1 & \cdots & \partial F_1/\partial x_n \\ \vdots & \ddots & \vdots \\ \partial F_n/\partial x_1 & \cdots & \partial F_n/\partial x_n \end{bmatrix} = \det\Big(dF(e_1),\dots,dF(e_n)\Big).$$

Here $e_1,\dots,e_n$ is the canonical basis for the Cartesian space.

In exterior calculus, we will simply write

$$\int_{F(D)}\omega = \int_D F^*\omega,$$

i.e., the integral of ω over the image $F(D)$ of a map F is the same as the integral of the **pullback** differential $F^*\omega$ via the map over the domain D. The notion of pullback in exterior calculus will be the abstraction of the routine procedure of changing variables for the integrand.

Green, Gauss and Kelvin–Stokes Theorems

- The surface integral of the flux through a closed surface may be expressed as a volume integral taken within the surface.

- The circulation taken around a closed curve may be expressed in terms of a surface integral taken over a surface bounded by the curve.

Let $\mathbf{v}\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a vector field in the 2-dimensional Cartesian space. Let $D \subset \mathbb{R}^2$ be a region, and let ∂D be its boundary curve. Then by integrating the derivatives in each variable and noting the integration orientation to assign an appropriate sign, we obtain **Green's Theorem**

$$\iint_D \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}\right) dx_1\, dx_2 = \oint_{\partial D} v_1\, dx_2 - v_2\, dx_1.$$

In general, in n-dimension, $\sum_{i=1}^{n} \partial v_i/\partial x_i$ is called the **divergence** of $\mathbf{v}$. Its integral over an n-dimensional region equals to the total (outward) flux through the boundary surface. Take 3D for example. Suppose $\mathbf{u}\colon \mathbb{R}^3 \to \mathbb{R}^3$ is a vector field in 3D (with coordinates labeled with y), $D \subset \mathbb{R}^3$ is a 3D domain, and ∂D is its boundary surface with normal vectors denoted by $\mathbf{n}\colon \partial D \to \mathbb{R}^3$. Then we have **Gauss' Divergence Theorem**

$$\begin{aligned}\iiint_D \sum_{i=1}^{3} \frac{\partial u_i}{\partial y_i}\, dy_1\, dy_2\, dy_3 &= \oiint_{\partial D} u_1\, dy_2\, dy_3 + u_2\, dy_3\, dy_1 + u_3\, dy_1\, dy_2 \\ &= \oiint_{\partial D} \langle \mathbf{v}, \mathbf{n}\rangle\, dA.\end{aligned}$$

Let us come back to Green's Theorem. Suppose : $\mathbf{D} \subset \mathbb{R}^2 \to \mathbb{R}^3$ is a parametrized surface, whose restriction to the boundary $\mathbf{f}\colon \partial D \to \mathbb{R}^3$ is a parametrized space curve. Let $\vec{u}\colon \mathbb{R}^3 \to \mathbb{R}^3$ be a vector field in 3D. Then the **circulation** of $\mathbf{u}$ along $\mathbf{f}(\partial D)$ is

$$\begin{aligned}\oint_{\mathbf{f}(\partial D)} \langle \mathbf{u}, d\mathbf{f}\rangle &= \oint_{\mathbf{f}(\partial D)} \sum_{i=1}^{3} u_i\, dy_i = \oint_{\partial D} \sum_{i=1}^{3}\sum_{j=1}^{2} u_i \frac{\partial f_i}{\partial x_j}\, dx_j \\ &\overset{\text{Green's Thm.}}{=} \sum_{i=1}^{3} \iint_D \left(\frac{\partial}{\partial x_2}\left(u_i \frac{\partial f_i}{\partial x_1}\right) - \frac{\partial}{\partial x_1}\left(u_i \frac{\partial f_i}{\partial x_2}\right)\right) dx_1\, dx_2 \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} \iint_D \left(\frac{\partial u_i}{\partial y_j}\frac{\partial f_i}{\partial x_1}\frac{\partial f_j}{\partial x_2} - \frac{\partial u_i}{\partial y_j}\frac{\partial f_i}{\partial x_2}\frac{\partial f_j}{\partial x_1}\right) dx_1\, dx_2 \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} \iint_D \left(\frac{\partial u_i}{\partial y_j} - \frac{\partial u_j}{\partial y_i}\right)\frac{\partial f_i}{\partial x_1}\frac{\partial f_j}{\partial x_2}\, dx_1\, dx_2 \\ &= \iint_D \det\left(\operatorname{curl}\mathbf{u}, d\mathbf{f}(\mathbf{e}_1), d\mathbf{f}(\mathbf{e}_2)\right)\, dx_1\, dx_2 = \iint_{\mathbf{f}(D)} \langle \operatorname{curl}\mathbf{u}, \mathbf{n}\rangle\, dA,\end{aligned}$$

which is the total flux of $\operatorname{curl}\mathbf{u} = (\partial u_2/\partial y_3 - \partial u_3/\partial y_2, \partial u_3/\partial y_1 - \partial u_1/\partial y_3, \partial u_1/\partial y_2 - \partial u_2/\partial y_1)$ through the surface $\mathbf{f}(D)$. This is the **Kelvin–Stokes Theorem**.

In exterior calculus, vector calculus operators such as grad (∇), curl ($\nabla\times$), div ($\nabla\cdot$) all coalesce into a single derivative d. Similarly, important theorems such as the Fundamental Theorem of Calculus, or Gauss' and Green's Theorem all become instances of one single general statement, the so-called Stokes' theorem

$$\int_M d\omega = \int_{\partial M} \omega,$$

i.e., the integral of the differential of ω over the domain M is the same of ω over the boundary of the domain.

6.2 An Overview of the Language

The main objects that exterior calculus deals with are **differential forms**. Functions, or scalar fields, take a given point of the domain, and return a value. Extending this idea, differential forms are to be evaluated (integrated) over given curves, surfaces, or volumes, *etc.*

Specifically, let M be a 3-dimensional domain. On M, a k-form is a differential form that is to be integrated over a k-dimensional (oriented) submanifold:

- A 3-form is to be integrated over a volumetric region. For example, the mass density ρ of physical matter should be regarded as a 3-form, denoted by $\rho \in \Omega^3(M)$. Given a region $U \subset M$, the total mass within U is denoted by $\int_U \rho$.
- A 2-form is to be integrated over an oriented surface. For example, a flux ω of a flow should be regarded as a 2-form, denoted by $\omega \in \Omega^2(M)$. It describes each total flux over a given surface Σ, denoted by $\int_\Sigma \omega$.
- A 1-form is to be line integrated over a path. For example, a force field η should be regarded as a 1-form, denoted by $\eta \in \Omega^1(M)$. The total work done by the force along a given path Γ is denoted by $\int_\Gamma \eta$.
- A 0-form is a function, or a scalar field. For example, the temperature u of a material is a 0-form, denoted by $u \in \Omega^0(M)$. It is to be evaluated at each point.

We will encounter only 5 algebraic and differential operators for differential forms:

- **Pullback operator.** If we have a differential form $\alpha \in \Omega^k(W)$, and a map $f : M \to W$, then we have a pullback differential form $f^*\alpha \in \Omega^k(M)$. It has a natural definition that its evaluation over a k-dimensional surface $\Sigma \subset M$ is given by $\int_\Sigma (f^*\alpha) = \int_{f(\Sigma)} \alpha$.
- **Exterior derivative.** Denoted by d, the exterior derivative is a differential operator that sends a k-form α to a $(k+1)$-form $d\alpha$. The evaluation of $d\alpha$ over a $(k+1)$-dimensional surface Σ is designed so that $\int_\Sigma d\alpha = \int_{\partial\Sigma} \alpha$.
- **Wedge product.** It is an algebraic operator that produces higher-degree forms. If $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M)$ then their wedge product $\alpha \wedge \beta \in \Omega^{k+\ell}(M)$.
- **Interior product.** Provided a given direction, the interior product as an algebraic operator lowers the degree of a form. Suppose X is a tangent vector of M, and α is a k-form, then the interior product $i_X\alpha$ is a $(k-1)$-form.
- **Hodge star.** Denoted by $\star$, it is an algebraic operator that turns a k-form α into an $(n-k)$-form $\star\alpha$, where $n = \dim(M)$. Among these 5 operators, this Hodge star operator is the only one that requires a Riemannian metric for M.

Macroscopic and Microscopic View

A differential 1-form η is a quantity that is waiting to be integrated over a curve γ. This is the "macroscopic" viewpoint for differentail forms. Let us take a short parametrized curve γ, which is so short that γ is well-described by a point $x = \gamma(0)$, a tangent vector $\mathbf{v} = \gamma'(0)$ at x as the velocity for the curve, and a small ε as the length of the interval of the parameter for the curve. Abusing the notation slightly, we write $\gamma = \varepsilon\mathbf{v}$. Then with continuity in η, we expect $\int_{\varepsilon\mathbf{v}} \eta \sim O(\varepsilon)$. The limit

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{\varepsilon\mathbf{v}} \eta =: \eta(\mathbf{v})$$

is a quantity depending linearly on the velocity vector $\mathbf{v}$. We say $\eta(\mathbf{v})$ is the evaluation of the 1-form on a tangent vector $\mathbf{v}$. This is the algebraic, or the "microscopic," viewpoint of differential forms.

Conversely, given a 1-form η as a scalar-valued linear function on tangent vectors, we can understand its "macroscopic" counterpart as follows. The integral of η over a curve γ can be obtained by first partioning the parameter interval of γ and then taking the sum of microscopic evaluation

$$\int_\gamma \eta := \lim_{\substack{\text{partion} \\ \text{refines}}} \sum_i \eta\big(\gamma'(t_i)\big)(t_{i+1} - t_i).$$

Similarly, a differential 2-form ω in a "macroscopic" view is a quantity to be evaluated over an oriented surface. Its "microscopic" version is a function linearly evaluating on an infinitesimal oriented parallelgram spanned by two vectors v_1, v_2. More precisely, $\omega(\mathbf{v}_1, \mathbf{v}_2)$ is bilinear (linear in each component if fixing the other) and skew symmetric $\omega(\mathbf{v}_1, \mathbf{v}_2) = -\omega(\mathbf{v}_2, \mathbf{v}_1)$. To assemble a microscopic 2-form into its integral over a surface, partition the surface into infinitesimal parallelgrams and sum the values of the bilinear forms.

As mentioned in Section 6.1, it turns out that the right structure for integration that evaluate circulations, fluxes, *etc.*, is infinitesimally a **skew symmetric multilinear form**. In Section 6.3 and Section 6.4 we study the linear algebra of these forms.

6.3 Vectors and Covectors

In this section we focus on the dual space of a vector space. The elements in dual space are called **1-forms** or **covectors**.

6.3.1 Example: Differential of Functions

Let us start with something we are familiar with—differential of a function—and try to understand the underlying structure.

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ on the n-dimensional domain $\mathbb{R}^n$. The differential df of f at a point $x \in \mathbb{R}^n$ is a linear map given as the linear term for the Taylor expansion of f at x:

$$df|_x : \mathbb{R}^n \xrightarrow{\text{linear}} \mathbb{R}$$
$$f(x + \varepsilon\mathbf{v}) = f(x) + df|_x(\mathbf{v})\varepsilon + O(\varepsilon^2).$$

That is, $df|_x$ is the linear operator which best approximates the function f at that point. This linear operator maps vectors from the domain to "amounts of change" in the range, *i.e.*, it maps the tangent space of the domain to the tangent space of the range.

More generally, the differential df of a scalar function $f : M \to \mathbb{R}$ on a manifold is the **tangent map** (*cf.* Section 5.1.2)

$$df|_x : T_xM \xrightarrow{\text{linear}} \mathbb{R}.$$

It takes a tangent vector and measures the rate of change of the value of f along that vector. In other words, we get the **directional derivative**. Later we will explain how this differential df as a linear map is "converted" into a **gradient** vector.

Definition 6.1 — Linear functional. We have a special name for linear maps from a vector space to $\mathbb{R}$. They are called **linear functionals**. In geometry, they are also called **covectors** or **1-forms**.

Objects such as $df|_x$ is a covector at each x. So we can also call them a covector field, or a 1-form field. Conventionally, people abbreviate 1-form field as just 1-form.

A proper way to illustrate df is the following. First of all, f is a real-valued function over a domain M. One can illustrate f via the level sets of f similar to topographic maps. Now, at a given point $x \in M$, the linear functional $df|_x$ is the *density* (or the spatial *frequency*) and *direction* of the level sets at that point. Imagine the level sets of f are wavefronts of a wave, then $df|_x$ is the plane wave (all wavefronts are parallel and equally spaced) that share the same frequency and orientation as f at x.

A 1-form field is to assign an orientation and frequency of "putative level sets" at each point. However, a general 1-form field as these scattered waves may not be integrated into streamlined level set (large sheets of wavefronts). The differential of a function df is a special type of 1-form field when the local linear level sets (described by the linear functional at each point) integrate into the level sets of a function.

6.3.2 Dual Vector Space

Let V be a real vector space. For example $V = T_xM$ at some $x \in M$. The collection of all linear functional on V is called the dual space of V.

Definition 6.2 The **dual space** of V is denoted and given by $V^\star = \{\alpha \mid \alpha \colon V \xrightarrow{\text{linear}} \mathbb{R}\}$.

The dual space $V^\star$ is itself a vector space and in fact has the same dimension as V.

Notation 6.1 — Dual pairing. Let $\alpha \in V^\star$, *i.e.* α is a linear functional on V. Then given any $\mathbf{v} \in V$, we can evaluate $\alpha(\mathbf{v})$ by sticking $\mathbf{v}$ into the linear functional α. A notation that we will sometimes use is **dual pairing**

$$\langle \cdot | \cdot \rangle \colon V^\star \times V \to \mathbb{R}, \quad \langle \alpha | \mathbf{v} \rangle \coloneqq \alpha(\mathbf{v}).$$

The bra–ket notation $\langle \alpha | \mathbf{v} \rangle$ is just a different way of writing $\alpha(\mathbf{v})$ to emphasize that α is also like a vector (a covector).

In elementary linear algebra, we have two types of vectors—the column vectors and row vectors. They represent vectors and covectors. One may think of V as a collection of column vectors $\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$, whereas $V^\star$ as a collection of row vectors $\begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix}$. They can be naturally paired by $\langle \alpha | v \rangle = \begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i=1}^n \alpha_i v_i$.

A geometric way of thinking of vectors and covectors are the following. The vector space V is a flat space with an origin. Each vector $\mathbf{v}$ in V is an arrow basing at the origin and ending at some location in the space. A covector $\alpha \in V^\star$ is visually represented by the level sets of the linear functional $\alpha \colon V \to \mathbb{R}$, which is a family of equally-spaced oriented hyperplanes all parallel to $\ker(\alpha) \subset V$.

Vectors are arrows, and covectors are plane waves.

The larger α is (*i.e.* the greater slope the linear functional α has), the denser (higher frequency) the parallel hyperplanes are distributed. The dual pairing $\langle \alpha | \mathbf{v} \rangle$ evaluates the number of hyperplanes the arrow $\mathbf{v}$ intersects, counted with sign.

6.3.3 Conversions between Vectors and Covectors

So far, we have seen that vectors and covectors are two quite different objects. One is a point on a vector space, and one is a linear functional. One is an arrow, and one is a plane wave. It is important to distinct them.

There are two circumstances where we can map vectors to covectors:

- When there is a metric (inner product structure) on V;
- When a basis is given on V;

each induces an linear isomorphism (an invertible map) between V and $V^\star$.

R In elementary linear algebra, we have the "transpose" operation that converts row vectors to column vectors and vice versa. Transpose is a rather sloppy notion as it may refer to one of the above isomorphisms that happen to coincide in the Euclidean and Cartesian vector space $\mathbb{R}^n$. Do keep in mind that the metric-induced and the basis-induced isomorphisms are generally different.

Metric-induced Isomorphism—The Musical Isomorphisms

Suppose $\langle \cdot, \cdot \rangle_V$ is an inner product structure on V. (See Section 5.2 for more detail.) An inner product allow us to talk about the length of an arrow and the angle between two arrows.

Using the inner product structure, we define:

Definition 6.3 — Flat $\flat$. The **flat** operator ($\flat$) is a linear map that converts vectors to covectors

$$(\cdot)^\flat \colon V \to V^\star. \tag{6.1}$$

The flat $v^\flat$ of a vector $v \in V$ is a covector (linear functional) defined by

$$\mathbf{v}^\flat(\mathbf{w}) = \langle \mathbf{v}^\flat | \mathbf{w} \rangle := \langle \mathbf{v}, \mathbf{w} \rangle_V \quad \mathbf{w} \in V. \tag{6.2}$$

Definition 6.4 — Sharp $\sharp$. The **sharp** operator ($\sharp$) is the inverse of $\flat$:

$$(\cdot)^\sharp \colon V^\star \to V, \tag{6.3}$$

that is, the sharp $\alpha^\sharp$ of a covector $\alpha \in V^\star$ is the unique vector $\alpha^\sharp = \mathbf{u} \in V$ such that

$$\alpha(\mathbf{w}) = \langle \mathbf{u}, \mathbf{w} \rangle_V, \quad \mathbf{w} \in V. \tag{6.4}$$

The existence and uniqueness of such $u \in V$ is known as the **Riesz representation theorem**.

Definition 6.5 — Dual inner product. The inner product structure $\langle \cdot, \cdot \rangle_V$ on V induces an inner product structure $\langle \cdot, \cdot \rangle_{V^\star}$ on $V^\star$, given by

$$\langle \alpha, \beta \rangle_{V^\star} := \langle \alpha^\sharp, \beta^\sharp \rangle_V, \quad \alpha, \beta \in V^\star. \tag{6.5}$$

What do all these mean geometrically?

The intuition is that, being able to measure lengths and angles of arrows, we can measure the spatial frequency (as $\frac{1}{\text{distance from the 0-level set to the 1-level set}}$) of the plane waves correspond to the

covectors. We can also measure the angles between the oriented parallel hyperplanes from the two covectors. These are the norm and angle for covectors as measured by $\langle \cdot, \cdot \rangle_{V^\star}$.

The sharp $\alpha^\sharp$ of a covector α (plane wave) is the vector (arrow) that points orthogonal (with respect to $\langle \cdot, \cdot \rangle_V$) to the hyperplane $\ker(\alpha)$ of the plane wave in the positive direction, and the arrow length equals to the frequency of α (inverse spacing between the level sets).

The flat $\mathbf{v}^\flat$ of an arrow $\mathbf{v}$ is the plane wave whose wavefronts are parallel to the orthogonal complement $\mathbf{v}^\perp$ of $\mathbf{v}$, and the frequency is given by $|\mathbf{v}|$.

▪ **Example 6.1 — Gradient.** The differential df of a function $f : M \to \mathbb{R}$ on some manifold M is a covector $df|_x \in (T_x M)^\star$ at each x. We call the dual space of a tangent space a **cotangent space**, often denoted by $T_x^\star M = (T_x M)^\star$. Now, suppose each tangent space is equipped with a metric $\langle \cdot, \cdot \rangle$, *i.e.* M is a Riemannian manifold, then we can convert the covector field df to a vector field.

The **gradient** of f, denoted by $(\operatorname{grad} f)$, is a tangent vector $(\operatorname{grad} f)_x \in T_x M$ at every $x \in M$ given by

$$(\operatorname{grad} f)_x := (df|_x)^\sharp. \tag{6.6}$$

That is, the directional derivatives are represented by the inner product between the direction and the gradient vector: $df|_x(\mathbf{v}) = \langle (\operatorname{grad} f)_x, \mathbf{v} \rangle$ for all $\mathbf{v} \in T_x M$. ▪

Basis-induced Isomorphism and Dual Basis

Suppose we only have V without any metric. Then there is no clear way of how to convert a covector in $V^\star$ to an arrow in V, or convert an arrow to a covector.

However, if we have n ($n = \dim(V)$) linearly independent covectors $\alpha_1, \ldots, \alpha_n$, *i.e.* a basis for $V^\star$. Then their 0-level sets and 1-level sets will form a parallelepiped. Incident to the origin, the n edges of this parallelepiped are n arrows. These are the n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ converted from a basis covectors: $\mathbf{v}_1$ is the intersection of the 0-level sets of $\alpha_2, \ldots, \alpha_n$, and $\mathbf{v}_2$ is the intersection of the 0-level sets of $\alpha_1, \alpha_3 \ldots, \alpha_n$, and so on.

We can also do this process in reverse. Given a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for V, we have a parallelepiped spanned by these vectors. The $(n-1)$-dimensional faces of this parallelepiped characterizes n parallel planes and their spacing, describing n covectors.

The concrete algebraic description is the following.

Definition 6.6 — Dual basis. Given a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for V, we define the basis $\mathbf{v}_1^\star, \ldots, \mathbf{v}_n^\star$ for $V^\star$ so that

$$\langle \mathbf{v}_i^\star | \mathbf{v}_j \rangle = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \tag{6.7}$$

Suppose in a scenario that each $\mathbf{v}_i$ is represented by a column vector. Then each covector in the dual basis $\mathbf{v}_j^\star$ is a row vector by virtue of representing dual pairing as row-column contraction. These column and row vectors can be stacked together as matrices. The condition for the dual basis is then given explicitly via a matrix inversion:

$$\begin{bmatrix} — & \mathbf{v}_1^\star & — \\ & \vdots & \\ — & \mathbf{v}_n^\star & — \end{bmatrix} = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{bmatrix}^{-1}. \tag{6.8}$$

Similarly, given a basis $\alpha_1, \dots, \alpha_n$ for $V^\star$, we have a dual basis $\alpha_1^\star, \dots, \alpha_n^\star$ for V.

R The dual $\mathbf{v}_i^\star$ of $\mathbf{v}_i$ only makes sense if we know the rest of $\mathbf{v}_1, \dots, \mathbf{v}_n$. The dual basis depends on the entire given basis.

Theorem 6.1 — Dual basis under a change of basis. Let $A: V \to V$ be an invertible linear map. Consider a basis transformation $\mathbf{u}_1 = A^{-1}\mathbf{v}_1, \dots, \mathbf{u}_n = A^{-1}\mathbf{v}_n$. Then the new dual basis is given by $\mathbf{u}_1^\star = \mathbf{v}_1^\star \circ A, \dots, \mathbf{u}_n^\star = \mathbf{v}_n^\star \circ A$. Here, the composition means $(\mathbf{v}_i^\star \circ A)(\mathbf{w}) = \mathbf{v}_i^\star(A\mathbf{w})$ for all $\mathbf{w} \in V$.

Proof. Check (6.7) for $\mathbf{u}$'s. □

▪ **Example 6.2 — Partial derivative.** Let V be an n-dimensional vector space, and let $\mathbf{r}: V \to \mathbb{R}^n$ be a linear map to the Cartesian space. That is, $\mathbf{r}$ represents each input $\mathbf{v} \in V$ as an n-tuple of numbers $(r_1(\mathbf{v}), \dots, r_n(\mathbf{v}))$.

We consider $\mathbf{r}$ a linear map, so each r_i is a linear functional. In some context $\mathbf{r}$ might be a linear map plus a translation (*i.e.* an affine map). Then we would extract the linear part by $d\mathbf{r}$. So, just for compatibility with more general affine coordinate $\mathbf{r}$, we say $(dr_1(\mathbf{v}), \dots, dr_n(\mathbf{v}))$ extracts the n components of $\mathbf{v}$ under this basis $\mathbf{r}$.

Suppose $f: V \to \mathbb{R}$ is a function (not necessarily linear) on V. The **partial derivative** $\partial f/\partial r_i$ in this coordinate system is given by

$$\frac{\partial f}{\partial r_i} = df((dr_i)^\star)$$

where the vectors $(dr_1)^\star, \dots, (dr_n)^\star \in V$ are the dual basis of the basis $dr_1, \dots, dr_n$ for $V^\star$. Geometrically, partial derivatives measure the directional derivative of f along the direction in which all other coordinates $r_1, \dots, r_{i-1}, r_{i+1}, r_n$ are fixed. That is, we are taking the direction of an edge of the parallelepiped bounded by the coordinate level planes, described by the covectors $dr_1, \dots, dr_n$. ▪

R The notion of partial derivative is meaningful only if we know what are the rest of dependent variables that are being fixed.

R If V has a metric $\langle \cdot, \cdot \rangle$, and $\mathbf{v}_1, \dots, \mathbf{v}_n$ is an *orthonormal basis*, then $\mathbf{v}_i^\star = \mathbf{v}_i^\flat$. Note that $\mathbf{v}_i^\star$ and $\mathbf{v}_i^\flat$ are generally different for general basis.

In fact, the column/row vector representation of elements in $V, V^\star$ uses a dual pair of basis. Let $(\mathbf{e}_i)_{i=1}^n$ be a basis for V and $(\mathbf{e}_i^\star)_{i=1}^n$ be its dual basis for $V^\star$. Then each vector $\mathbf{v} \in V$ can be decomposed into $\mathbf{v} = b_1\mathbf{e}_1 + \cdots + b_n\mathbf{e}_n$ for real numbers b_i. Similarly, for each covector $\alpha \in V$, we decompose it into $\alpha = a_1\mathbf{e}_1^\star + \cdots + a_n\mathbf{e}_n^\star$. Then we can represent v by the column vector

$\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ and α by the row vector $\begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$. Due to the consistency between the two bases, the dual pairing $\langle \alpha | \mathbf{v} \rangle$ is exactly the contraction between rows and columns:

$$\langle \alpha | \mathbf{v} \rangle = \sum_i \sum_j \langle a_i \mathbf{e}_i^\star | b_j \mathbf{e}_j \rangle = \sum_i \sum_j a_i b_j \delta_{ij} = \sum_i a_i b_i = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

6.4 Exterior Algebra

As previewed in the brief overview in Section 6.2, we will move on to k-form. Again, we will begin with a fixed n-dimensional vector space V and study k-forms on V. Later, V will be replaced by the tangent spaces $T_x M$ for all $x \in M$ of a manifold M. Assigning a k-form per tangent space describes a k-form field. A k-form field is to be put under an integral sign over a k-dimensional surface, representing the k-dimensional flux or circulation over that surface.

Let V be an n-dimensional real vector space. A **k-form** $\omega \in \bigwedge^k V^\star$ is an **alternating** (*i.e.* skew-symmetric) k-linear map

$$\omega : \underbrace{V \times V \times \cdots \times V}_{k} \to \mathbb{R}.$$

That is,

- (k-linearity): $\omega(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ is linear in each $\mathbf{v}_i$ when the rest of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are fixed. For example, $\omega(a\mathbf{u}_1 + b\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) = a\omega(\mathbf{u}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) + b\omega(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$;
- (alternating): Swapping any two entries give a minus sign $\omega(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_k) = -\omega(\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_k)$.

If $\dim(V) = n$ then $0 \le k \le n$ is the valid range for k. 1-forms are linear functionals or covectors. A 0-form, an object that take zero number of vectors, is just a real number.

▪ **Example 6.3** The determinant $\det(\cdot, \cdot, \ldots, \cdot)$ is an n-form on $\mathbb{R}^n$. It takes n vectors and computes the signed volume of the parallelepiped spanned by those n vectors. Any scalar multiple (a det) of det is also an n-form.

On a general n-dimensional vector space V without additional structure, there is no apriori given notion of volume.

A V equipped with a nonzero n-form $\det_V(\cdot, \cdot, \ldots, \cdot)$, or a **volume form**, would be called *a vector space with a volume form* or *with a determinant*. ▪

Exercise 6.1 Show that a k-form on an n-dimensional space V must be zero if $k > n$. ▪

6.4.1 Wedge Product

So how do we construct k-forms? For this we will use the **wedge product**.

Wedge Product of 1-forms

The simplest case is given by the wedge product of two 1-forms, $\alpha, \beta \in V^\star$, which is defined as

$$\alpha \wedge \beta(\mathbf{v}, \mathbf{w}) := \alpha(\mathbf{v})\beta(\mathbf{w}) - \alpha(\mathbf{w})\beta(\mathbf{v}) = \det \begin{bmatrix} \alpha(\mathbf{v}) & \beta(\mathbf{v}) \\ \alpha(\mathbf{w}) & \beta(\mathbf{w}) \end{bmatrix}. \tag{6.9}$$

Obviously $\alpha \wedge \beta$ is a 2-form since it is multi-linear (bi-linear) and alternating. More generally we have a k-fold wedge product of 1-forms to produce a k-form

$$\alpha_1 \wedge \cdots \wedge \alpha_k(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \det \begin{bmatrix} \alpha_1(\mathbf{v}_1) & \cdots & \alpha_k(\mathbf{v}_1) \\ \vdots & & \vdots \\ \alpha_1(\mathbf{v}_k) & \cdots & \alpha_k(\mathbf{v}_k) \end{bmatrix}. \tag{6.10}$$

What does it represent geometrically? If each 1-form α_i corresponds to equally-spaced parallel hyperplanes, then $\alpha_1 \wedge \cdots \wedge \alpha_k$ is their intersection, which are uniformly spaced parallel $(n-k)$-dimensional planes. For example, in $n = 3$, 1-forms are parallel planes, 2-forms are parallel lines, and 3-forms are a lattice of points. The denser they get, the larger the form is.

Note that if any $\alpha_i = \alpha_j$ in $\alpha_1 \wedge \cdots \wedge \alpha_k$, or just linearly dependent, then the product is zero.

In fact, we may talk about 1-forms that takes values in non-commutative algebra, such as quaternions, or $\mathbb{R}^3$ with cross product. For example, let α, β be $\mathbb{R}^3$-valued 1-forms. Building wedge product over the cross product $\times$, we have $\alpha \underset{\times}{\wedge} \beta(v, w) = \alpha(v) \times \beta(w) - \alpha(w) \times \beta(v)$. In this case, $\alpha \underset{\times}{\wedge} \alpha$ is not necessarily zero. In fact $\alpha \underset{\times}{\wedge} \alpha(v, w) = 2\alpha(v) \times \alpha(w)$.

Does all k-form given by these k-fold wedge product of 1-forms? No, but almost. It turns out that all k-form can be written as a linear combination of a few k-fold wedge products, which we show next. We call it a **decomposable k-form**, or sometimes a **k-blade**, for the special case that a k-form can be written as a k-fold wedge product of 1-forms.

Basis for k-forms

Suppose $\mathbf{e}_1, \ldots, \mathbf{e}_n$ forms a basis for V. Let $\mathbf{e}_1^\star, \ldots, \mathbf{e}_n^\star$ be the dual basis for $V^\star$ (Definition 6.6). Then we can construct the following $\binom{n}{2}$ 2-forms

$$\mathbf{e}_1^\star \wedge \mathbf{e}_2^\star, \quad \mathbf{e}_1^\star \wedge \mathbf{e}_3^\star, \quad \cdots, \quad \mathbf{e}_1^\star \wedge \mathbf{e}_n^\star, \quad \mathbf{e}_2^\star \wedge \mathbf{e}_3^\star, \quad \cdots, \quad \mathbf{e}_{n-1}^\star \wedge \mathbf{e}_n^\star.$$

Similarly, for k-forms we can construct the following $\binom{n}{k}$ nonzero k-blades

$$\mathbf{e}_{i_1}^\star \wedge \mathbf{e}_{i_2}^\star \wedge \cdots \wedge \mathbf{e}_{i_k}^\star, \quad 1 \le i_1 < i_2 < \cdots < i_k \le n.$$

Notation 6.2 We often abbreviate these k-blades as

$$\mathbf{e}_I^\star = \mathbf{e}_{i_1}^\star \wedge \mathbf{e}_{i_2}^\star \wedge \cdots \wedge \mathbf{e}_{i_k}^\star, \quad I = (i_1 < \ldots < i_k)$$

with the multiindex I denoting an ordered selection of k numbers between 1 and n.

The claim is that these $\binom{n}{k}$ k-blades form a basis for k-forms $\bigwedge^k V^\star$.

The linear independence between is fairly straightforward to check. For each selections $I = (i_1 < \ldots < i_k)$, and $J = (j_1 < \ldots < j_k)$, of ordered indices, we see $\mathbf{e}_{i_1}^\star \wedge \cdots \wedge \mathbf{e}_{i_k}^\star(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) = 1$ if I and J are identical, and $= 0$ otherwise. So, to show linear independence, we suppose there is a linear combination of the basis k-blades that become zero: $\sum_I c_I \mathbf{e}_I^\star = 0$. Then by plugging in $(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k})$ to the formula, we effectively extract the coefficient find $c_J = 0$. By the similar procedure we conclude $c_I = 0$ for all multiindex I.

The following theorem shows that every k-form is a linear combination of these basis k-form.

Theorem 6.2 Any k-form $\omega \in \bigwedge^k V^\star$ is a linear combination of the basis k-blades with $\omega(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) \in \mathbb{R}$ being the coefficients:

$$\omega = \sum_{1 \le j_1 < \cdots < j_k \le n} \omega(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) \left(\mathbf{e}^\star_{j_1} \wedge \cdots \wedge \mathbf{e}^\star_{j_k}\right). \tag{6.11}$$

Proof. Take any $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$, and let $\mathbf{v}_i = v_{i1}\mathbf{e}_1 + \cdots + v_{in}\mathbf{e}_n$, $i = 1, \ldots, k$. That is, $v_{ij} = \mathbf{e}^\star_j(\mathbf{v}_i)$ are the coefficients of these vectors in this basis. Then using the multi-linearity, skew-symmetry, and regrouping summations, we find

$$\begin{aligned}
\omega(\mathbf{v}_1, \ldots, \mathbf{v}_k) &= \omega\left(\sum_{j_1=1}^n v_{1j_1}\mathbf{e}_{j_1}, \sum_{j_2=1}^n v_{2j_2}\mathbf{e}_{j_2}, \ldots, \sum_{j_k=1}^n v_{kj_k}\mathbf{e}_{j_k}\right) \\
&= \sum_{j_1=1}^n \cdots \sum_{j_k=1}^n v_{1j_1} v_{2j_2} \cdots v_{kj_k} \omega(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) \\
&= \left(\sum_{1 \le j_1 < \cdots < j_k \le n}\right) \underbrace{\left(\sum_{\substack{\sigma \text{ is a permutation} \\ \text{of } \{1,\ldots,k\}}}\right) \operatorname{sgn}(\sigma) v_{1j_{\sigma(1)}} v_{2j_{\sigma(2)}} \cdots v_{kj_{\sigma(k)}}}_{\det \begin{bmatrix} v_{1j_1} & \cdots & v_{1j_k} \\ \vdots & & \vdots \\ v_{kj_1} & \cdots & v_{kj_k} \end{bmatrix}} \omega(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) \\
&= \sum_{1 \le j_1 < \cdots < j_k \le n} \omega(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) \left(\mathbf{e}^\star_{j_1} \wedge \cdots \wedge \mathbf{e}^\star_{j_k}\right)(\mathbf{v}_1, \ldots, \mathbf{v}_k).
\end{aligned}$$

□

Corollary 6.3 The space of k-forms $\bigwedge^k V^\star$ is an $\binom{n}{k}$-dimensional vector space.

Wedge Products between General Forms

We have seen wedge product between 1-forms. The wedge product between two general forms is an **associative**, **bilinear operation**

$$\wedge\colon \bigwedge^k V^\star \times \bigwedge^\ell V^\star \to \bigwedge^{k+\ell} V^\star.$$

First, for basis k- and ℓ-blades, the $(k+\ell)$-form $\mathbf{e}^\star_I \wedge \mathbf{e}^\star_J$ is given by associativity

$$\left(e^\star_{i_1} \wedge \cdots \wedge \mathbf{e}^\star_{i_k}\right) \wedge \left(\mathbf{e}^\star_{j_1} \wedge \cdots \wedge \mathbf{e}^\star_{j_\ell}\right) \coloneqq \mathbf{e}^\star_{i_1} \wedge \cdots \wedge \mathbf{e}^\star_{i_k} \wedge \mathbf{e}^\star_{j_1} \wedge \cdots \wedge \mathbf{e}^\star_{j_\ell}. \tag{6.12}$$

Given a k-form $\omega \in \bigwedge^k V^\star$ and an ℓ-form $\eta \in \bigwedge^\ell V^\star$, we can compute the $(k+\ell)$-form $\omega \wedge \eta$ by first representing them in basis forms $\omega = \sum_I \omega_I \mathbf{e}^\star_I$, $\eta = \sum_J \omega_J \mathbf{e}^\star_J$, and use distributive law (bilinearity) to expand the wedge product

$$\omega \wedge \eta = \left(\sum_I \omega_I \mathbf{e}^\star_I\right) \wedge \left(\sum_J \eta_J \mathbf{e}^\star_J\right) = \sum_I \sum_J \omega_I \eta_J \mathbf{e}^\star_I \wedge \mathbf{e}^\star_J. \tag{6.13}$$

One can check that for general vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k+\ell} \in V$,

$$\omega \wedge \eta(\mathbf{v}_1, \ldots, \mathbf{v}_{k+\ell}) = \sum_{\sigma \in \text{Shuffles}(k,\ell)} \text{sgn}(\sigma)\omega\left(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(k)}\right) \eta\left(\mathbf{v}_{\sigma(k+1)}, \ldots, \mathbf{v}_{\sigma(k+\ell)}\right) \quad (6.14)$$

where $\text{Shuffles}(k, \ell)$ is the set of all permutations σ of $(1, \ldots, k+\ell)$ that satisfies $\sigma(1) < \ldots < \sigma(k)$ and $\sigma(k+1) < \ldots \sigma(k+\ell)$. That is, the sum runs over all possible ways of grouping $\mathbf{v}_1, \ldots, \mathbf{v}_{k+\ell}$ into the a bucket of k and a bucket of ℓ.

Defining wedge products (6.13) via a basis is straightforward, but it is not clear whether wedge products depend on the choice of basis. The basis-free formula (6.14) shows that wedge products are indeed independent of any choice of basis.

Exercise 6.2 Let $V = \mathbb{R}^4$ and define the 2-form $\alpha = u_{12}\mathbf{e}_1^\star \wedge \mathbf{e}_2^\star + u_{24}\mathbf{e}_2^\star \wedge \mathbf{e}_4^\star + u_{34}\mathbf{e}_3^\star \wedge \mathbf{e}_4^\star$, and the 1-form $\beta = w_2\mathbf{e}_2^\star + w_3\mathbf{e}_3^\star$. Compute $\alpha \wedge \beta$ and $\alpha \wedge \alpha$. ▪

Exercise 6.3 Using the associativity (6.12) between basis forms $(\mathbf{e}_I^\star \wedge \mathbf{e}_J^\star) \wedge \mathbf{e}_K^\star = \mathbf{e}_I^\star \wedge (\mathbf{e}_J^\star \wedge \mathbf{e}_K^\star)$, show that the wedge product is indeed associative: Let α, β and γ be k-, ℓ-, and r-forms respectively; check that $(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$. ▪

Exercise 6.4 Let α, β be k- and ℓ-forms respectively. Show that $\alpha \wedge \beta = (-1)^{k\ell}\beta \wedge \alpha$. ▪

6.4.2 Pullback by Linear Maps

Here we introduce an important operation that allows us to transport k-forms from one vector space to another via a linear map. Let U be an m-dimensional vector space and V be an n-dimensional vector space. Let $A\colon U \to V$ be any linear map. Suppose $\omega \in \bigwedge^k V^\star$ is a k-form on V. Then, using A we have a **pullback** $A^*\omega \in \bigwedge^k U^\star$ of ω, which is now a k-form on U. It is defined by that for any vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in U$

$$(A^*\omega)(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k) \coloneqq \omega(A\mathbf{u}_1, A\mathbf{u}_2 \ldots, A\mathbf{u}_k). \quad (6.15)$$

Using formula (6.14) one sees that the pullback operator distributes over wedges:

$$A^*(\omega \wedge \eta) = (A^*\omega) \wedge (A^*\eta). \quad (6.16)$$

6.4.3 Interior Product

Let V be an n-dimensional vector space. The **interior product** contracts k-form with a vector to produce a $(k-1)$-form. Let $\mathbf{v} \in V$. Then the interior product $i_\mathbf{v}$ is the operator

$$i_\mathbf{v}\colon \bigwedge^k V^\star \to \bigwedge^{k-1} V^\star \quad (6.17)$$

given by

$$(i_\mathbf{v}\omega)(\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}) = \omega(\mathbf{v}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}) \quad (6.18)$$

that is

$$i_\mathbf{v}\omega = \omega(\mathbf{v}, \cdot, \ldots, \cdot).$$

In particular, for a 1-form α:

$$i_{\mathbf{v}}\alpha = \langle \alpha | \mathbf{v} \rangle = \alpha(\mathbf{v}). \tag{6.19}$$

Note that applying interior product with the same vector twice always gives zero

$$i_{\mathbf{v}} i_{\mathbf{v}} \omega = 0. \tag{6.20}$$

The interior product satisfies the following product rule known as **Leibniz rule** due to its resemblance to the Leibniz rule for derivatives.

Theorem 6.4 — Leibniz rule for interior products. Let ω, η be a k- and ℓ-form respectively. Then

$$i_{\mathbf{v}}(\omega \wedge \eta) = (i_{\mathbf{v}}\omega) \wedge \eta + (-1)^k \omega \wedge (i_{\mathbf{v}}\eta). \tag{6.21}$$

Proof. It is enough to check for basis k-form $\omega = \mathbf{e}^\star_{i_1} \wedge \cdots \wedge \mathbf{e}^\star_{i_k}$ and ℓ-form $\eta = \mathbf{e}^\star_{j_1} \wedge \cdots \wedge \mathbf{e}^\star_{j_\ell}$. Now, using the determinant expression (6.10), for any wedge products of 1-forms $\alpha_1 \wedge \cdots \wedge \alpha_m$, one finds

$$\begin{aligned} i_{\mathbf{v}}(\alpha_1 \wedge \cdots \wedge \alpha_m) = \langle \alpha_1 | \mathbf{v} \rangle \alpha_2 \wedge \cdots \wedge \alpha_m - \langle \alpha_2 | \mathbf{v} \rangle \alpha_1 \wedge \alpha_3 \wedge \cdots \wedge \alpha_m + - \cdots \\ + (-1)^{m-1} \langle \alpha_m | \mathbf{v} \rangle \alpha_1 \wedge \cdots \wedge \alpha_{m-1}. \end{aligned}$$

Using this formula, one checks that (6.21) holds for basis forms. □

R In fact, $i_{\mathbf{v}}$ is the unique linear operator that satisfies (6.20), (6.21), and (6.19). These are the axioms for the interior product.

6.4.4 Metric and Hodge Star

Everything defined so far requires no metric $\langle \cdot, \cdot \rangle$ on V. Here we will introduce a useful operation that will depend on a given inner product structure.

Suppose we are in $\mathbb{R}^3$ with a standard orthonormal basis. We will define an operator called Hodge star, denoted by $\star$, so that $\star(\mathbf{e}^\star_1) = \mathbf{e}^\star_2 \wedge \mathbf{e}^\star_3$, $\star(\mathbf{e}^\star_2) = \mathbf{e}^\star_3 \wedge \mathbf{e}^\star_1$, $\star(\mathbf{e}^\star_3) = \mathbf{e}^\star_1 \wedge \mathbf{e}^\star_2$, $\star(\mathbf{e}^\star_1 \wedge \mathbf{e}^\star_2) = \mathbf{e}^\star_3$, *etc.* That is, it computes the complementary basis $(n-k)$-form for a given basis k-form. Let us properly define this notion.

Suppose V is equipped with a metric $\langle \cdot, \cdot \rangle$. Then we may choose an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ for V. In that case, $\mathbf{e}^\flat_i = \mathbf{e}^\star_i$.

Definition 6.7 Given a metric $\langle \cdot, \cdot \rangle$ on V, pick an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ for V. Define the metric $\langle \cdot, \cdot \rangle$ on the space of k-forms by declaring that the basis k-forms $\mathbf{e}^\flat_{i_1} \wedge \cdots \wedge \mathbf{e}^\flat_{i_k}$, $1 \le i_1 < \cdots < i_k \le n$, are orthonormal.

For example, let $\xi, \eta \in \bigwedge^k V^\star$. Then $\langle \xi, \eta \rangle = \sum_I \xi_I \eta_I$, where $\xi = \sum_I \xi_I \mathbf{e}^\flat_I$ and $\eta = \sum_J \eta_I \mathbf{e}^\flat_J$.

Definition 6.8 — Metric-induced volume form. The metric gives rise to the standard volume form denoted by det. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be an orthonormal basis for V. Then

$$\det := \mathbf{e}^\flat_1 \wedge \cdots \wedge \mathbf{e}^\flat_n. \tag{6.22}$$

Definition 6.9 — Hodge star. The **Hodge star** $\star$ is the linear operator

$$\star_k \colon \bigwedge^k V^\star \to \bigwedge^{n-k} V^\star \tag{6.23}$$

so that for every two k-forms $\omega, \eta \in \bigwedge^k V^\star$

$$\omega \wedge \star\eta = \langle \omega, \eta \rangle \det . \tag{6.24}$$

The subscript k in $\star_k$ is often suppressed when we know the degree of the form it is applied to.

In particular, applying $\star_k$ to a basis k-form $\mathbf{e}_{i_1}^\flat \wedge \cdots \wedge \mathbf{e}_{i_k}^\flat$ yields the complementary basis $(n-k)$-form (with an appropriate sign) so that they produce the volume form when they are joined together by the wedge product.

Some useful formulas include

$$\star_0 1 = \det, \qquad \star_n \det = 1, \tag{6.25}$$

$$\star_{n-k} \star_k = (-1)^{k(n-k)}, \qquad \star_k^{-1} = (-1)^{k(n-k)} \star_{n-k} . \tag{6.26}$$

Hodge star gives a duality that relates wedge products and interior products:

$$(-1)^k \star_{k+1}^{-1} i_{\mathbf{v}} \star_k \omega = \mathbf{v}^\flat \wedge \omega. \tag{6.27}$$

That is,

$$\begin{array}{ccccccccccc}
\bigwedge^0 V^\star & \xleftarrow{i_{\mathbf{v}}} & \bigwedge^1 V^\star & \xleftarrow{i_{\mathbf{v}}} & \bigwedge^2 V^\star & \xleftarrow{i_{\mathbf{v}}} & \bigwedge^3 V^\star & \longleftarrow\cdots & \bigwedge^{n-1} V^\star & \xleftarrow{i_{\mathbf{v}}} & \bigwedge^n V^\star \\
\downarrow{\scriptstyle \star_0} & \uparrow{\scriptstyle \mathbf{v}^\flat\wedge} & \downarrow{\scriptstyle \star_1} & \uparrow{\scriptstyle \mathbf{v}^\flat\wedge} & \downarrow{\scriptstyle \star_2} & \uparrow{\scriptstyle \mathbf{v}^\flat\wedge} & \downarrow{\scriptstyle \star_3} & & \downarrow{\scriptstyle \star_{n-1}} & \uparrow{\scriptstyle \mathbf{v}^\flat\wedge} & \downarrow{\scriptstyle \star_n} \\
\bigwedge^n V^\star & \xrightarrow[i_{\mathbf{v}}]{} & \bigwedge^{n-1} V^\star & \xrightarrow[-i_{\mathbf{v}}]{} & \bigwedge^{n-2} V^\star & \xrightarrow[i_{\mathbf{v}}]{} & \bigwedge^{n-3} V^\star & \longrightarrow\cdots & \bigwedge^1 V^\star & \xrightarrow[(-1)^{n-1} i_{\mathbf{v}}]{} & \bigwedge^0 V^\star
\end{array} \tag{6.28}$$

6.4.5 $\mathbb{R}^3$ Vector Algebra

Let us look at the special case $V = \mathbb{R}^3$ with $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ being the standard orthonormal basis. In this case $\mathbf{e}_i^\star = \mathbf{e}_i^\flat$.

Let $\mathbf{u}$ be a vector $\mathbf{u} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + u_3\mathbf{e}_3 \in \mathbb{R}^3$. The 3 components $u_1, u_2, u_3 \in \mathbb{R}$ can also be arranged in a 1-form

$$\mathbf{u}^\flat = u_1\mathbf{e}_1^\flat + u_2\mathbf{e}_2^\flat + u_3\mathbf{e}_3^\flat \in \textstyle\bigwedge^1 \mathbb{R}^{3\star}$$

or a 2-form

$$\star_1 \mathbf{u}^\flat = u_1(\mathbf{e}_2^\flat \wedge \mathbf{e}_3^\flat) + u_2(\mathbf{e}_3^\flat \wedge \mathbf{e}_1^\flat) + u_3(\mathbf{e}_1^\flat \wedge \mathbf{e}_2^\flat) \in \textstyle\bigwedge^2 \mathbb{R}^{3\star}.$$

Conversely, one can convert a 1-form to a vector by taking $(\cdot)^\sharp$, and convert a 2-form ω to a vector by taking $(\star_2\omega)^\sharp$.

Then dot products and cross products in $\mathbb{R}^3$ can be understood as one of the following.

Theorem 6.5 Let $\mathbf{a}, \mathbf{b}, \mathbf{w} \in \mathbb{R}^3$, $\alpha = \mathbf{a}^\flat$, $\beta = \mathbf{b}^\flat$, and $\omega = \star \mathbf{w}^\flat$;

- Wedge between two 1-forms gives a cross product: $\alpha \wedge \beta = \star(\mathbf{a} \times \mathbf{b})^\flat$.
- Wedge between a 1-form and a 2-form gives a dot product: $\alpha \wedge \omega = \omega \wedge \alpha = \langle \mathbf{a}, \mathbf{w} \rangle \det$.
- Interior product applied to a 1-form gives a dot product: $i_{\mathbf{a}} \beta = \langle \mathbf{a}, \mathbf{b} \rangle$.
- Interior product applied to a 2-form gives a cross product: $i_{\mathbf{a}} \omega = (\mathbf{w} \times \mathbf{a})^\flat$.
- Interior product applied to the volume form gives the correspondence between vectors and 2-forms: $i_{\mathbf{w}} \det = \omega$.

Exercise 6.5 Check Theorem 6.5. ∎

Exercise 6.6 Use the Leibniz rule (6.21) and Theorem 6.5 to show the BAC-CAB formula

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}\langle \mathbf{a}, \mathbf{c} \rangle - \mathbf{c}\langle \mathbf{a}, \mathbf{b} \rangle, \quad \mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3 \tag{6.29}$$

and the Binet–Cauchy identity

$$\langle \mathbf{a} \times \mathbf{b}, \mathbf{c} \times \mathbf{d} \rangle = \langle \mathbf{a}, \mathbf{c} \rangle \langle \mathbf{b}, \mathbf{d} \rangle - \langle \mathbf{b}, \mathbf{c} \rangle \langle \mathbf{a}, \mathbf{d} \rangle, \quad \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{R}^3. \tag{6.30}$$

Hint *Treat one of the cross products as the wedge product and treat the other cross product as the interior product.* ∎

6.4.6 Cramer's Rule

Have you wondered how **Cramer's rule** work for solving linear systems? It is probably most naturally understood using exterior algebra, specifically pullbacks and interior products.

Cramer's rule is an explicit formula for the solution of a linear system using determinants. In matrix form, let $A = \begin{bmatrix} | & & | \\ \mathbf{a}_1 & \cdots & \mathbf{a}_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times n}$ be an invertible matrix, $\mathbf{b} \in \mathbb{R}^n$ be a given column vector. Solving $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x} \in \mathbb{R}^n$, Cramer's rule is

$$x_1 = \frac{\det \begin{bmatrix} | & | & & | \\ \mathbf{b} & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix}}{\det A}, \quad x_2 = \frac{\det \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{b} & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix}}{\det A}, \quad \ldots, \quad x_n = \frac{\det \begin{bmatrix} | & & | & | \\ \mathbf{a}_1 & \cdots & \mathbf{a}_{n-1} & \mathbf{b} \\ | & & | & | \end{bmatrix}}{\det A}. \tag{6.31}$$

How do we understand it geometrically or algebraically without subject to a basis that gives this matrix representation?

Let U, V be two n-dimensional vector space with volume forms $\det_U \in \bigwedge^n U^\star, \det_V \in \bigwedge^n V^\star$. Let $A \colon U \to V$ be a linear map.

When we take the determinant of a linear map such as A, what we mean is the quantity

$$\det A := \frac{A^* \det_V}{\det_U} \in \mathbb{R}, \quad \text{that is,} \quad A^* \det_V = (\det A) \det_U, \tag{6.32}$$

where A^* is the pullback operator. Here we can take the quotient of two n-forms since any n-form is a real multiple of another n-form (the dimension of $\bigwedge^n U^\star$ is $\binom{n}{n} = 1$). What this means geometrically is that the determinant of a linear map A measures how much of a volume expansion A makes; the volume measurement are made via the volume forms $\det_U, \det_V$ of the domain and range spaces.

Having a volume form $\det_U$, we can identify $\mathbf{w} \in U$ with an $(n-1)$-form $\omega \in \bigwedge^{n-1} U^\star$ by the correspondence $i_{\mathbf{w}} \det_U = \omega$. We often call an $(n-1)$-form a **flux form**. The flux form $\omega = i_{\mathbf{w}} \det_U$ evaluated as $\omega(\mathbf{u}_1, \dots, \mathbf{u}_{n-1})$ is the total flux of vector $\mathbf{w}$ through the parallelgram face spanned by $\mathbf{u}_1, \dots, \mathbf{u}_{n-1}$. It is the total volume passing through the parallelgram per unit time in a fluid flow of constant velocity $\mathbf{w}$.

Similarly, a vector $\mathbf{b} \in V$ corresponds to an $(n-1)$-form $\beta = i_{\mathbf{b}} \det_V$. The natural operation we can do to a form β on V is to pull it back to U by $A\colon U \to V$. This gives an $(n-1)$-form $A^*\beta \in \bigwedge^{n-1} U^\star$, which then corresponds to a vector in $\mathbf{u}$ so that $i_{\mathbf{u}} \det_U = A^*\beta$. This defines a linear map starting from a vector $\mathbf{b} \in V$ and landing at a vector $\mathbf{u} \in U$, called the **adjugate** of A

$$\begin{aligned} &\operatorname{adj} A\colon V \to U \\ &i_{(\operatorname{adj} A)(\mathbf{b})} \det_U := A^*(i_{\mathbf{b}} \det_V). \end{aligned} \tag{6.33}$$

You may have seen adjugate in matrix algebra: it is the transpose of the **cofactor matrix**. Here adjugate has a clear meaning: it is the pullback of fluxes written in terms of vectors.

So far, we have not assumed that A is invertible. One can always take the adjugate of A regardless the invertibility of A. Now let us suppose A is invertible. Then simply by the definition of interior product and pullback, the right-hand side of (6.33) can be written as

$$A^*(i_{\mathbf{b}} \det_V) = i_{A^{-1}\mathbf{b}}(A^* \det_V) \overset{(6.33)}{=} (\det A)\, i_{A^{-1}\mathbf{b}} \det_U . \tag{6.34}$$

Equating it to the left-hand side of (6.33) we have

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{(\operatorname{adj} A)(\mathbf{b})}{\det A}, \quad \text{or simply} \quad A^{-1} = \frac{\operatorname{adj} A}{\det A}. \tag{6.35}$$

To finally see Cramer's rule (6.31), take a basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ for U compatible with the volume form $\mathbf{e}_1^\star \wedge \cdots \wedge \mathbf{e}_n^\star = \det_U$. We are interested in the coefficients of $\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$. Note that the column vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ of A in (6.31) are precisely the image of the individual basis vectors $\mathbf{a}_j = A(\mathbf{e}_j)$. So, using (6.34), *i.e.* $i_{\mathbf{x}} \det_U = {A^*(i_{\mathbf{b}} \det_V)}/{\det A}$, we have

$$\begin{aligned} x_1 = (i_{\mathbf{x}} \det_U)(\mathbf{e}_2, \dots, \mathbf{e}_n) &= \frac{1}{\det A} (A^* i_{\mathbf{b}} \det_V)(\mathbf{e}_2, \dots, \mathbf{e}_n) \\ &= \frac{1}{\det A} (i_{\mathbf{b}} \det_V)(\mathbf{a}_2, \dots, \mathbf{a}_n) = \frac{\det_V(\mathbf{b}, \mathbf{a}_2, \dots, \mathbf{a}_n)}{\det A}. \end{aligned}$$

Similarly for other components

$$x_i = (-1)^{i-1}(i_{\mathbf{x}} \det_U)(\mathbf{e}_1, \dots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \dots, \mathbf{e}_n) = \frac{\det_V(\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{b}, \mathbf{a}_{i+1}, \dots, \mathbf{a}_n)}{\det A}.$$

6.5 Differential k-forms

In the previous section, we have studied k-form as a linear algebraic object. Now we are back to view k-forms as a to-be-integrated differential object.

We have seen that k-forms have the right structure for integration, but now the vector space V becomes the tangent space T_xM with the base point $x \in M$ varying over some region of an n-dimensional manifold M. A **differential k-form**, denoted by $\omega \in \Omega^k(M)$ or $\omega \in \Omega^k(M; \mathbb{R})$, is a **$k$-form field**, *i.e.* it is an assignment of a k-form on each T_xM for every $x \in M$ so that it is smoothly varying with respect to varying x.

Representation under a Coordinate

To understand the notion of smoothness of differential form, let us take a coordinate chart. Let $U \subset M$ denote a neighborhood in M. Let $\mathbf{r} = (r_1, \ldots, r_n) \colon U \to \mathbb{R}^n$ be a coordinate, which is a map into the Cartesian space $\mathbb{R}^n$. Then the tangent maps $dr_1, \ldots, dr_n$ are covector fields over U, each of which at a point $x \in M$ is a covector/linear functional $dr_i|_x \colon T_xM \xrightarrow{\text{linear}} \mathbb{R}$, *i.e.* $dr_i \in T_x^\star M$. In particular, due to the invertibility of $d\mathbf{r}$, these covectors $dr_1|_x, \ldots, dr_n|_x$ form a basis for $T_x^\star M$ for every x. These are the elementary coordinate differentials. One may build the basis for $\bigwedge^k T_x^\star M$ as the elementary basis k-blades $dr_{i_1} \wedge \cdots \wedge dr_{i_k}, 1 \le i_1 < \ldots < i_k \le m$. A differential k-form $\omega \in \Omega^k(M)$ is smooth in the following sense. The coefficients with respect to any coordinate chart are now differentiable functions $w_{i_1,\ldots,i_k} \colon U \to \mathbb{R}, 1 \le i_1 < \ldots < i_k \le m$:

$$\omega = \sum_{1 \le i_1 < \ldots < i_k \le m} w_{i_1,\ldots,i_k}\, dr_{i_1} \wedge \cdots \wedge dr_{i_k}. \tag{6.36}$$

Note that $w_{i_1,\ldots,i_k} = \omega(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k})$ where $\mathbf{e}_i = (dr_i)^\star = {}^{\partial \mathbf{r}^{-1}}/_{\partial r_i}$ is the basis for T_xM dual to $dr_1, \ldots, dr_n$.

6.5.1 Pullback

Pullback is the ultimate change of variable.

Suppose $\varphi \colon M \to W$ is a differentiable map from a manifold M to another manifold W. These manifolds do not need to have the same dimension. Suppose $\omega \in \Omega^k(W)$ is a k-form on W. Then we can pullback this k-form back to M using φ.

Definition 6.10 The **pullback** $\varphi^*\omega \in \Omega^k(M)$ of $\omega \in \Omega^k(W)$ by $\varphi \colon M \to W$ is given by

$$(\varphi^*\omega)(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \omega\big(d\varphi(\mathbf{v}_1), \ldots, d\varphi(\mathbf{v}_k)\big) \quad \text{for tangent vectors } \mathbf{v}_1, \ldots, \mathbf{v}_k \in TM. \tag{6.37}$$

That is, $\varphi^*\omega$ is the pullback (Section 6.4.2) by the linear map $d\varphi|_x$ at every point x.

Similar to Section 6.4.2, pullback distributes over wedges

$$\varphi^*(\alpha \wedge \beta) = (\varphi^*\alpha) \wedge (\varphi^*\beta). \tag{6.38}$$

What should be noted is that, 0-forms $h \in \Omega^0(W; \mathbb{R})$ are just scalar functions $h \colon M \to \mathbb{R}$. Its pullback is given by function composition

$$\varphi^* h = h \circ \varphi. \tag{6.39}$$

One may imagine a scalar function multiple $h\alpha$ of a form α as the wedge product $h \wedge \alpha$ between $h \in \Omega^0(W)$ and $\alpha \in \Omega^k(W)$. So, we can also distribute pullback over them

$$\varphi(h\alpha) = (\varphi^* h)(\varphi^*\alpha) = (h \circ \varphi)(\varphi^*\alpha). \tag{6.40}$$

Exercise 6.7 Suppose $\varphi \colon M \to W$. Let $j \colon M \to \mathbb{R}$ be a smooth scalar function. Show that

$$d(\varphi^* h) = \varphi^*(dh). \tag{6.41}$$

Hint *Plug in a tangent vector of M. If in doubt how to apply chain rule, represent tangent*

vector as the velocity of a curve (5.9). ▪

Exercise 6.8 Suppose $\varphi\colon M \to W$. Let $\mathbf{r}\colon W \to \mathbb{R}^n$ be some coordinate for W, representing $\omega \in \Omega^k(W)$ as $\omega = \sum_{i_1<\ldots<i_k} w_{i_1,\ldots,i_k} dr_{i_1} \wedge \cdots \wedge dr_{i_k}$ for some scalar functions $w_{i_1,\ldots,i_k}$. Show that

$$\varphi^*\omega = \sum_{i_1<\ldots<i_k} (w_{i_1,\ldots,i_k} \circ \varphi)\, d(r_{i_1} \circ \varphi) \wedge \cdots \wedge d(r_{i_k} \circ \varphi). \tag{6.42}$$

▪

6.5.2 Integration

Let us talk about integration of k-forms. One defines integration using pullbacks.

First $\omega \in \Omega^k(\Sigma)$ where Σ is a k-dimensional submanifold. For simplicity, suppose further that Σ is the image of the k-dimensional unit cube under the parametrization $\varphi\colon [0,1] \times \cdots \times [0,1] \to \Sigma$. (If not possible, partition Σ to smaller pieces and sum any resulting integrals in the end.) Then $\varphi^*\omega$ would be a k-form on the cube $[0,1]^k$, represented by a scalar function $u\colon [0,1]^k \to \mathbb{R}$ times the volume form:

$$\varphi^*\omega = u\, dx_1 \wedge \cdots \wedge dx_n.$$

Define the integral

$$\int_\Sigma \omega := \int_{[0,1]^k} \varphi^*\omega = \int_{[0,1]^k} u \det = \int_0^1 \cdots \int_0^1 u(x_1,\ldots,x_k)\, dx_1 \cdots dx_k. \tag{6.43}$$

Note that if $(r_1,\ldots,r_k)\colon \Sigma \to \mathbb{R}^k$ is another coordinate system on Σ, giving another scalar function representation $\omega = w\, dr_1 \wedge \cdots \wedge dr_k$. Then $u = w \det(d\varphi)$, where $\det(d\varphi) = {\varphi^*(dr_1\wedge\cdots\wedge dr_k)}/{dx_1\wedge\cdots\wedge dx_k}$ is the familiar Jacobian determinant for change of integration variable. See (6.32) for quotients of volume forms.

Next, let us consider a k-form $\omega \in \Omega^k(M)$ in an n-dimensional manifold M, $n > k$. Then we can integrate ω over a k-dimensiona surface. Suppose $f\colon \Sigma \to M$ is a k-dimensional surface, forming an oriented multi-set $f(\Sigma) \subset M$. Then we define

$$\int_{f(\Sigma)} \omega := \int_\Sigma f^*\omega. \tag{6.44}$$

The latter is an integration of a k-form on a k-dimensional manifold, which is the case we have already discussed.

6.5.3 Exterior Derivative

Now we are ready for the differential as it applies to k-forms,

$$d\colon \Omega^k(M) \to \Omega^{k+1}(M). \tag{6.45}$$

This **exterior derivative** d

- acts on functions h (0-forms) just as the usual differential dh, which we already saw.
- It also satisfies $d(dh) = 0$ and more generally $d \circ d = 0$.

- Finally, it satisfies the product rule with respect to the wedge product, called the **Leibniz rule**,

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta \tag{6.46}$$

where $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M)$.

(Note its similarity in the latter two properties to interior products (6.20), (6.21).)

One can start with the above axiomatic properties of the exterior derivative to get the following properties, or one can define d via the following formulas. Let $\mathbf{r}$ be a coordinate.

- Let $\omega = \sum_I w_I dr_I$, define d by

$$d\omega = \sum_I d(w_I) \wedge dr_I. \tag{6.47}$$

- d applied on a 0-form h is defined by

$$dh = \sum_{i=1}^{n} \frac{\partial f}{\partial r_i} dr_i. \tag{6.48}$$

Exercise 6.9 Using (6.47) and (6.48), and the fact that $\frac{\partial^2 h}{\partial r_i \partial r_j} = \frac{\partial^2 h}{\partial r_j \partial r_i}$, show that $dd\omega = 0$. ∎

Vector Calculus Operators

Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a 0-form and $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^3$ a vector field. We can arrange $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3$ into a 1-form $\mathbf{v}^\flat = v_1\, dr_1 + v_2\, dr_2 + v_3\, dr_3$ or a 2-form $\star\mathbf{v}^\flat = i_{\mathbf{v}} \det = v_1\, dr_2 \wedge dr_3 + v_2\, dr_3 \wedge dr_1 + v_3\, dr_1 \wedge dr_2$. Then

- d on 0-form gives gradient.
- d on 1-form gives curl.
- d on 2-form gives divergence.

$$\operatorname{grad} f = \nabla f = (df)^\sharp \tag{6.49}$$

$$(\operatorname{curl} \mathbf{v})^\flat = (\nabla \times \mathbf{v})^\flat = \star d\mathbf{v}^\flat \tag{6.50}$$

$$\operatorname{div} \mathbf{v} \det = (\nabla \cdot \mathbf{v}) \det = d \star \mathbf{v}^\flat = d i_{\mathbf{v}} \det . \tag{6.51}$$

Exercise 6.10 Verify (6.50) and (6.51) using (6.47) and (6.48). ∎

In fact, the relation between ∇, $\nabla\times$, $\nabla\cdot$ and d's are not limited to $\mathbb{R}^3$, but any 3-dimensional Riemannian manifold. One defines grad, curl and div through Hodge stars and exterior derivatives.

A diagram summarizing the vector calculus operators on a 3-dimensional manifold is

$$\begin{array}{ccccccccc} & & & & \Gamma(TM) & & & & \\ & & & & \flat \downarrow \uparrow \sharp & & & & \\ \Omega^0(M) & \xrightarrow[(\text{grad})]{d_0} & \Omega^1(M) & \xrightarrow[(\text{curl})]{d_1} & \Omega^2(M) & \xrightarrow[(\text{div})]{d_2} & \Omega^3(M) & & \\ \star_0 \downarrow \uparrow \star_3 & & \star_1 \downarrow \uparrow \star_2 & & \star_2 \downarrow \uparrow \star_1 & & \star_3 \downarrow \uparrow \star_0 & & \\ \Omega^3(M) & \xleftarrow[d_2]{(\text{div})} & \Omega^2(M) & \xleftarrow[d_1]{(\text{curl})} & \Omega^1(M) & \xleftarrow[d_0]{(\text{grad})} & \Omega^0(M) & & \\ & & & & \sharp \downarrow \uparrow \flat & & & & \\ & & & & \Gamma(TM) & & & & \end{array} \tag{6.52}$$

Recall that $d \circ d = 0$, which corresponds to the vector calculus identities

$$\operatorname{curl} \circ \operatorname{grad} = 0, \quad \operatorname{div} \circ \operatorname{curl} = 0. \tag{6.53}$$

Exercise 6.11 In a later chapter, we will focus on the Hodge Laplacian for k-forms. Here we just write the Laplacian for functions, *i.e.*, 0-forms, and 1- respectively 2-forms using exterior calculus. Let $f : \mathbb{R}^3 \to \mathbb{R}$. Then

$$\Delta f = \frac{\partial^2 f}{\partial r_1^2} + \frac{\partial^2 f}{\partial r_2^2} + \frac{\partial^2 f}{\partial r_3^2} = \operatorname{div}(\operatorname{grad} f) = \star d \star df. \tag{6.54}$$

This follows immediately from the definitions of div and grad in terms of exterior calculus. More interesting is the case of Laplacian of a vector field. Let $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^3$ be a vector field then the standard vector Laplacian reads as

$$\Delta \mathbf{v} = (\Delta v_1 + \Delta v_2, \Delta v_3), \tag{6.55}$$

i.e., each components of $\mathbf{v} = (v_1, v_2, v_3)$ is treated as a function. Now let $\alpha = \mathbf{v}^\flat$ and show that

$$(\Delta \mathbf{v})^\flat = (\star d \star d - d \star d \star)\alpha. \tag{6.56}$$

■

Exercise 6.12 Let $f, g : \mathbb{R}^3 \to \mathbb{R}$ be a scalar field and $\mathbf{a}, \mathbf{b} : \mathbb{R}^3 \to \mathbb{R}^3$ be vector fields in 3D. Use Leibniz rules (for d) and Section 6.4.5 to show

(a) $\nabla \cdot (\mathbf{a} \times \mathbf{b}) = (\nabla \times \mathbf{a}) \cdot \mathbf{b} - \mathbf{a} \cdot (\nabla \times \mathbf{b})$.
(b) $\nabla \cdot (f\mathbf{a}) = (\nabla f) \cdot \mathbf{a} + f \nabla \cdot \mathbf{a}$.
(c) $\nabla \times (f\mathbf{a}) = \nabla f \times \mathbf{a} + f \nabla \times \mathbf{a}$.
(d) $\nabla \times (f \nabla g) = \nabla f \times \nabla g$.

These formulae hold in general 3-dimensional Riemannian manifold, where the cross product is defined in the sense of Section 6.4.5. ■

Pullback and Exterior Derivative Commute

We have seen that a useful property of the pullback operator is that it distributes over the wedge product

$$\varphi^*(\alpha \wedge \beta) = (\varphi^*\alpha) \wedge (\varphi^*\beta).$$

Even more useful is the fact that pullback commutes with the exterior derivative

Theorem 6.6 — Naturality of exterior derivative. $d(\varphi^*\alpha) = \varphi^*(d\alpha)$.

Proof. It follows from (6.41) for 0-form, (6.47), and that pullback distributes over wedges. □

An intuitive picture behind Theorem 6.6 is via the Stokes Theorem, $\int_\Sigma d\omega = \int_{\partial\Sigma} \omega$, which we will see later. Combining Stokes Theorem and $\int_{\varphi(\Sigma)} \omega = \int_\Sigma \varphi^*\omega$, what Theorem 6.6 really says is that $\partial(\varphi(\Sigma)) = \varphi(\partial\Sigma)$, that is, the boundary of the image of a map (treated as a multiset) is the image of the boundary.

Corollary 6.7 The implications of this almost-trivial fact (Theorem 6.6) to 3D vector calculus are less obvious, or even mysterious. After translated into vector calculus identities, they are useful formulas for Continuum Mechanics as Theorem 6.6 is ultimately coordinate changes for derivatives. Unlike those in Exercise 6.12, these identities are not easily seen using the traditional multivariable calculus or index notation machinery.

For example, let $\mathbf{v}\colon \mathbb{R}^3 \to \mathbb{R}^3$ be a vector field, and $\varphi\colon \mathbb{R}^3 \to \mathbb{R}^3$ be a map, typically representing a flow map or change of coordinates. Consider $d\varphi$ written as a matrix field called **deformation gradient** $\mathbf{F}$, $F_{ij} = \partial\varphi_j/\partial x_i$. Let $J = \det \mathbf{F}$. Then

$$J\mathbf{F}^{-1}(\nabla \times \mathbf{v}) = \nabla \times (\mathbf{F}^\intercal \mathbf{v}), \tag{6.57}$$

$$\nabla \cdot [J\mathbf{F}^{-1}\mathbf{v}] = J\nabla \cdot \mathbf{v}. \tag{6.58}$$

One can also write $J\mathbf{F}^{-1} = \operatorname{adj} \mathbf{F}$, which is the adjugate matrix of $\mathbf{F}$, *i.e.* the transpose of the cofactor matrix $(\operatorname{adj} \mathbf{F}) = (\operatorname{cof} \mathbf{F})^\intercal$.

Taking $\mathbf{v} = \mathbf{e}_i$ for (6.58), one concludes interesting statements such as *each row of the cofactor matrix of any deformation gradient is divergence-free.*

Proof. (6.57) is obtained by the vector version of $\varphi^* d\mathbf{v}^\flat = d\varphi^* \mathbf{v}^\flat$, and noting that pullback of flux forms (2-forms) is represented by the adjugate matrix (Section 6.4.6). (6.57) is the vector version of $d\varphi^*(i_{\mathbf{v}} \det) = \varphi^* d(i_{\mathbf{v}} \det)$. □

6.5.4 Stokes Theorem

The central theorem and workhorse of exterior calculus is **Stokes Theorem**.

Theorem 6.8 Let ω be a $(k-1)$-form and Σ be a compact k-dimensional manifold with boundary $\partial\Sigma$. Then

$$\int_\Sigma d\omega = \oint_{\partial\Sigma} \omega. \tag{6.59}$$

Or in words, integrating the exterior derivative of a form over a region is the same as integrating the form itself over the boundary of the region. If we think of integration as a bilinear pairing $\langle\!\langle \cdot | \cdot \rangle\!\rangle$ between a domain and a differential form this can also be expressed as

$$\langle\!\langle \Sigma | d\omega \rangle\!\rangle = \langle\!\langle \partial\Sigma | \omega \rangle\!\rangle. \tag{6.60}$$

In other words, the topological operator ∂ is the adjoint of the differential operator d under this pairing. This way of thinking about it will become important later on when we *define* the discrete exterior derivative using Stokes Theorem and the boundary operator.

Stokes Theorem unifies (and extends to higher dimension) a number of theorems from calculus. For example, the **Fundamental Theorem of Calculus** is the 1D version

$$\int_{[a,b]} df = \int_{\partial[a,b]} f = f(b) - f(a) \tag{6.61}$$

since $\partial[a,b] = b - a$ as two points with opposite signs. Another example is the **divergence theorem**, which states that the divergence of a vector field $\mathbf{v}\colon \mathbb{R}^3 \to \mathbb{R}^3$ over a region $C \subset \mathbb{R}^3$ is equal to the net in-/out-flow across the boundary

$$\iiint_C \nabla \cdot \mathbf{v} \det = \oiint_{\partial C} \mathbf{v} \cdot \mathbf{n} \det_{\partial C} \tag{6.62}$$

where det is the standard volume form, $\mathbf{n}$ the outward pointing normal of ∂C, and $\det_{\partial C} (= i_{\mathbf{n}} \det)$ the area form on ∂C. Stokes tells us that

$$\int_C d \star \mathbf{v}^\flat = \oint_{\partial C} \star \mathbf{v}^\flat \tag{6.63}$$

and $\star\, \mathbf{v}^\flat = (i_{\mathbf{v}} \det)$ measures the flux $\mathbf{v} \cdot \mathbf{n} \det_{\partial C}$ at the boundary (for each tangent vector $\mathbf{u}_1, \mathbf{u}_2$ of the boundary surface, $i_{\mathbf{v}} \det(\mathbf{u}_1, \mathbf{u}_2) = \det(\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2) = (\mathbf{v} \cdot \mathbf{n}) \det(\mathbf{n}, \mathbf{u}_1, \mathbf{u}_2) = (\mathbf{v} \cdot \mathbf{n}) \det_{\partial C}(\mathbf{u}_1, \mathbf{u}_2)$).

For 1-form, we have the **Kelvin–Stokes Theorem** for surfaces $\Sigma \subset \mathbb{R}^3$

$$\iint_\Sigma (\nabla \times \mathbf{v}) \cdot \mathbf{n} \det_\Sigma = \oint_{\partial\Sigma} \mathbf{v} \cdot d\mathbf{l} \tag{6.64}$$

corresponding to

$$\int_\Sigma d\mathbf{v}^\flat = \oint_{\partial\Sigma} \mathbf{v}^\flat. \tag{6.65}$$

Proof of Stokes Theorem

We will show Stokes Theorem for k-dimensional cubes $[0,1]^k$ only. All other integrals can be pullback to this domain under suitable partitions and parametrizations. Also, since the formula of Stokes Theorem is linear in the $(k-1)$-form $\omega \in \Omega^{k-1}([0,1]^k)$, we only need to show the formula for $(k-1)$-blades

$$\omega = f\, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k. \tag{6.66}$$

Here $\widehat{dx_i}$ denotes skipping this term. With this choice of direction of the $(k-1)$-blade, we particularly have

$$\oint_{\partial[0,1]^k} \omega = \left(\int_{\{x_i=1\}} - \int_{\{x_i=0\}} \right) f \, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k, \tag{6.67}$$

that is, it is only the flux through the faces normal to x_i-axis. We will see that it matches with the other side of the Stokes formula.

Now, on the other hand

$$\begin{aligned} d\omega &= \sum_{j=1}^{k} \frac{\partial f}{\partial x_j} \, dx_j \wedge dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k \\ &= (-1)^{i-1} \frac{\partial f}{\partial x_i} \, dx_1 \wedge \cdots \wedge dx_k. \end{aligned}$$

Thus

$$\begin{aligned} \int_{[0,1]^k} d\omega &= \int_{[0,1]^3} (-1)^{i-1} \frac{\partial f}{\partial x_i} \, dx_1 \wedge \cdots \wedge dx_k \\ &= \int_{[0,1]^k} \frac{\partial f}{\partial x_i} \, dx_i \wedge dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k \\ &= \int_{[0,1]^{k-1}} \left(\int_0^1 \frac{\partial f}{\partial x_i} \, dx_i \right) dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k \\ &= \int_{[0,1]^{k-1}} \left(f|_{x_i=1} - f|_{x_i=0} \right) dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k \\ &= \left(\int_{\{x_i=1\}} - \int_{\{x_i=0\}} \right) f \, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_k, \end{aligned}$$

which indeed agrees with $\oint_{\partial[0,1]^k} \omega$. Therefore

$$\int_{[0,1]^k} d\omega = \oint_{\partial[0,1]^k} \omega.$$

6.6 Codifferential and Laplacian

Let us talk about codifferential, $\delta \colon \Omega^k(M) \to \Omega^{k-1}(M)$. It is defined as

$$\delta_k := (-1)^k \star_{k-1}^{-1} d_{n-k} \star_k . \tag{6.68}$$

That is,

$$\begin{array}{ccccccccccccc} \Omega^0(M) & \xrightarrow{d_0} & \Omega^1(M) & \xrightarrow{d_1} & \Omega^2(M) & \xrightarrow{d_2} & \Omega^3(M) & \longrightarrow\cdots\longrightarrow & \Omega^{n-1}(M) & \xrightarrow{d_{n-1}} & \Omega^n(M) \\ \downarrow{\scriptstyle \star_0} & \overset{\delta_1}{\nwarrow} & \downarrow{\scriptstyle \star_1} & \overset{\delta_2}{\nwarrow} & \downarrow{\scriptstyle \star_2} & \overset{\delta_3}{\nwarrow} & \downarrow{\scriptstyle \star_3} & & \downarrow{\scriptstyle \star_{n-1}} & \overset{\delta_n}{\nwarrow} & \downarrow{\scriptstyle \star_n} \\ \Omega^n(M) & \xleftarrow[-d_{n-1}]{} & \Omega^{n-1}(M) & \xleftarrow[d_{n-2}]{} & \Omega^{n-2}(M) & \xleftarrow[-d_{n-3}]{} & \Omega^{n-3}(M) & \longleftarrow\cdots\longleftarrow & \Omega^1(M) & \xleftarrow[(-1)^n d_0]{} & \Omega^0(M) \end{array} \tag{6.69}$$

As the name suggests it is something of an adjoint for the exterior derivative d with respect to the inner product on forms (*cf.* Section 6.4.4). Let $\alpha \in \Omega^{k-1}(M)$, $\beta \in \Omega^k(M)$, and $\star 1 = \det \in \Omega^n(M)$ the volume form. By Leibniz rule,

$$d(\alpha \wedge \star\beta) = (d\alpha) \wedge \star\beta - (-1)^k \alpha \wedge d \star \beta = (d\alpha) \wedge \star\beta + \alpha \wedge \star(\delta\beta),$$

or

$$d(\alpha \wedge \star\beta) = (\langle d\alpha, \beta\rangle - \langle \alpha, \delta\beta\rangle) \det . \tag{6.70}$$

If we integrate both sides over a manifold *without a boundary* $\partial M = \varnothing$, then they are indeed adjoints of one another due to Stokes theorem.

Definition 6.11 — L^2 inner product. Define an inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle : \Omega^k(M) \times \Omega^k(M) \to \mathbb{R}$ for k-form fields by

$$\langle\!\langle \omega, \eta \rangle\!\rangle := \int_M \omega \wedge \star\eta = \int_M \langle \omega, \eta \rangle \det, \quad \omega, \eta \in \Omega^k(M). \tag{6.71}$$

Theorem 6.9 — Integration by parts (without boundary). Suppose $\partial M = \varnothing$. Then

$$\langle\!\langle d\alpha, \beta \rangle\!\rangle = \langle\!\langle \alpha, \delta\beta \rangle\!\rangle, \quad \alpha \in \Omega^{k-1}(M), \beta \in \Omega^k(M). \tag{6.72}$$

Now that we have the codfferential we can define the **Hodge Laplacian** for k-forms as

$$-\Delta = d\delta + \delta d. \tag{6.73}$$

Notice that for 0-form this amounts to $\Delta f = -\delta df$ since $\delta f = 0$. Similarly for n-forms μ we get $\Delta\mu = -d\delta\mu$ since $d\mu = 0$.

Exercise 6.13 Check that $\Delta f = \sum_{i=1}^n \partial^2 f / \partial x_i^2$ for $f \in \Omega^0(\mathbb{R}^n)$. ∎

Exercise 6.14 Show that for a vector field $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^3$, $(\Delta \mathbf{v}^\flat)^\sharp$ is the vector Laplacian $\Delta \mathbf{v} = (\Delta v_1, \Delta v_2, \Delta v_3)$. Conclude as an immediate corollary that the 2-form Hodge Laplacian in $\mathbb{R}^3$ also amounts to the standard vector Laplacian. ∎

6.7 Interior Products and Lie Derivatives of Differential Forms

The machinery we will introduce in this section is important when it comes to many vector identities involved in fluid dynamics. It is not immediately useful for later chapters. But we include this section for completeness.

We saw similarity between the exterior derivative d and interior products $i_{\mathbf{v}}$: If α is a k-form, then

- $d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^k \alpha \wedge d\beta$.
- $i_{\mathbf{v}}(\alpha \wedge \beta) = (i_{\mathbf{v}}\alpha) \wedge \beta + (-1)^k \alpha \wedge i_{\mathbf{v}}\beta$.
- $dd\alpha = 0$.
- $i_{\mathbf{v}} i_{\mathbf{v}} \alpha = 0$.

We also saw that d can be understood as the adjoint of the boundary operator (Stokes theorem)

$$\int_{\partial\Sigma} \alpha = \int_{\Sigma} d\alpha.$$

Is there a Stokes-like theorem for the interior products? In addition, we saw that pullback operator give the integral formula $\int_{\mathbf{f}(\Sigma)} \alpha = \int_{\Sigma} \mathbf{f}^* \alpha$ that is somewhat Stokes-style. How does pullback operator interplays with d and $i_{\mathbf{v}}$?

It turns out that $i_{\mathbf{v}}$ does admit a Stokes-like formula. The interplay between pullback, d and $i_{\mathbf{v}}$ lies in the chain rule for pullback and the so-called Cartan's magic formula (a.k.a. Reynolds transport theorem).

6.7.1 Flow Map

Recall that a vector $\mathbf{v} \in T_x M$ at a point $x \in M$ is viewed as the velocity $\mathbf{v} = {}^{d\gamma}/_{dt}|_{t=0}$ of some path γ passing through x. Similarly, each *vector field* $\mathbf{v} \in \Gamma(TM)$ is the velocity field of a **flow map**. A flow map is parametrized by time, and at each fixed time it is a map on M:

$$\phi\colon \underbrace{[0,\varepsilon]}_{\text{time}} \times M \to M, \quad \phi_t\colon M \to M \text{ for each } t \in [0,\varepsilon] \tag{6.74}$$

representing the velocity field $\mathbf{v} \in T_x M$ by

$$\dot{\phi} = \left.\frac{\partial \phi}{\partial t}\right|_{t=0} = \mathbf{v}. \tag{6.75}$$

Later we will just write $\dot{\phi}$ and $\mathbf{v}$ interchangeably (just like how $\dot{\gamma}$ and $\mathbf{v}$ was used interchangeably previously).

6.7.2 Lie Derivative of Differential Forms

Suppose $\alpha \in \Omega^k(M)$ is a k-form on M. Let ϕ_t be a flow map. Consider $\phi_t^* \alpha \in \Omega^k(M)$ which is time dependent. What is the time derivative of $\phi_t^* \alpha \in \Omega^k(M)$? That is, how do we apply chain rule for it? The time derivative of this pullback form is defined as the **Lie derivative**.

Definition 6.12 — Lie derivative of differential forms. Let $\mathbf{v} \in \Gamma(TM)$ be a vector field. The Lie derivative $\mathscr{L}_{\mathbf{v}}\colon \Omega^k(M) \to \Omega^k(M)$ along a vector field $\mathbf{v}$ is defined by

$$\mathscr{L}_{\mathbf{v}}\, \alpha = \left.\frac{\partial}{\partial t}\right|_{t=0} \phi_t^* \alpha \quad \text{where } \phi \text{ is a flow map with } \dot{\phi} = \left.\frac{\partial \phi}{\partial t}\right|_{t=0} = \mathbf{v}. \tag{6.76}$$

In other words,

$$\left.\frac{\partial}{\partial t}\right|_{t=0} \phi_t^* \alpha = \mathscr{L}_{\left.\partial\phi/\partial t\right|_{t=0}} \alpha. \tag{6.77}$$

Using $\int_{\Sigma} \phi_t^* \alpha = \int_{\phi_t(\Sigma)} \alpha$ we can talk about the integral form of Lie derivative:

$$\left.\frac{d}{dt}\right|_{t=0} \int_{\phi_t(\Sigma)} \alpha = \left.\frac{d}{dt}\right|_{t=0} \int_{\Sigma} \phi_t^* \alpha = \int_{\Sigma} \left.\frac{\partial}{\partial t}\right|_{t=0} \phi_t^* \alpha = \int_{\Sigma} \mathscr{L}_{\dot{\phi}}\, \alpha. \tag{6.78}$$

That is, the evaluation of a differential form α on a flowing surface $\phi_t(\Sigma)$ has a rate of change given by the evaluation of the Lie derivative of α along the flow velocity on the surface.

R In many applications, the k-form field α may represent a physical quantity that also change over time. In that case,

$$\left.\frac{\partial}{\partial t}\right|_{t=0} \phi_t^* \alpha_t = \frac{\partial}{\partial t}\alpha + \mathscr{L}_\phi \alpha. \tag{6.79}$$

We call the operator $\partial/\partial t + \mathscr{L}_\phi$ **Lie material derivative**. If α is a 0-form, it is known as the **material derivative** $\partial/\partial t + \mathbf{v}\cdot\nabla$ which measures the rate of change of a function α evaluated at a point flowing along the flow.

6.7.3 Extrusion

The exterior derivative d (differential form operation) is adjoint to the boundary operator ∂ (geometric operation) with respect to the pairing $\int_{(\text{geometry})}(\text{differential form})$. The interior product $i_{\mathbf{v}}$ is adjoint to—the extrusion.

Let ϕ_t be a flow map associated to the vector field $\mathbf{v}$. For each k-dimensional surface Σ define the $(k+1)$-dimensional submanifold by extruding Σ along the flow:

$$\text{ext}_{\mathbf{v}}^{\varepsilon}(\Sigma) = \bigsqcup_{0\le t\le\varepsilon} \phi_t(\Sigma) \tag{6.80}$$

or more precisely if $\Sigma = \mathbf{f}(\hat{\Sigma})$ for some parametrization $\mathbf{f}$, then $\text{ext}_{\mathbf{v}}^{\varepsilon}(\Sigma) = \tilde{\mathbf{f}}([0,\varepsilon]\times\Sigma)$ is parametrized by

$$\tilde{\mathbf{f}}\colon [0,\varepsilon]\times\Sigma \to M, \quad \tilde{\mathbf{f}}(t,\mathbf{x}) = \phi_t(\mathbf{f}(\mathbf{x})). \tag{6.81}$$

Then for each $\beta \in \Omega^{k+1}(M)$,

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_{\text{ext}_{\mathbf{v}}^{\varepsilon}(\Sigma)} \beta = \int_{\Sigma} i_{\mathbf{v}}\beta. \tag{6.82}$$

6.7.4 Cartan's Magic Formula

As geometric operations, it is not hard to picture that

$$\phi_\varepsilon(\Sigma) - \Sigma = \partial\left(\text{ext}_{\mathbf{v}}^{\varepsilon}(\Sigma)\right) + \text{ext}_{\mathbf{v}}^{\varepsilon}\left(\partial(\Sigma)\right). \tag{6.83}$$

Using (6.82) and (6.78), and applying $\partial/\partial\varepsilon|_{\varepsilon=0}$ we turn this geometric formula into a differential form formula

$$\mathscr{L}_{\mathbf{v}}\,\alpha = i_{\mathbf{v}}(d\alpha) + d(i_{\mathbf{v}}\alpha). \tag{6.84}$$

This formula is called **Cartan's Magic Formula** (or **Cartan's homotopy formula**). Its integral form is given by

$$\left.\frac{d}{dt}\right|_{t=0} \int_{\phi_t(\Sigma)} \alpha = \int_{\partial\Sigma} i_{\mathbf{v}}\alpha + \int_{\Sigma} i_{\mathbf{v}}(d\alpha) + \underbrace{\int_{\Sigma}\frac{\partial\alpha}{\partial t}}_{\substack{\text{if } \alpha \text{ is also}\\ \text{time dependent.}}} \tag{6.85}$$

Eq. (6.85) is known as the **Reynolds Transport Theorem**.

R In 3D, the Lie material derivatives $\partial/\partial t + \mathscr{L}_{\mathbf{v}}$ of various k-forms look like:

- f is a 0-form. Then $(\partial/\partial t + \mathscr{L}_{\mathbf{v}})f = \partial f/\partial t + \mathbf{v}\cdot\nabla f$, called the **material derivative** of f.
- $\alpha = \mathbf{a}^\flat$ is a 1-form. Then $((\partial/\partial t + \mathscr{L}_{\mathbf{v}})\alpha)^\sharp = \partial \mathbf{a}/\partial t + \mathbf{v}\cdot\nabla\mathbf{a} + (\nabla\mathbf{v})\cdot\mathbf{a}$, called the **lower convected derivative** of $\mathbf{a}$.
- $\omega = \star\mathbf{w}^\flat$ is a 2-form. Then $(\star((\partial/\partial t + \mathscr{L}_{\mathbf{v}})\omega))^\sharp = \partial\mathbf{w}/\partial t + \mathbf{v}\cdot\nabla\mathbf{w} - \mathbf{w}\cdot\nabla\mathbf{v} + (\nabla\cdot\mathbf{v})\mathbf{w}$, known as the material derivative with **stretching** and **compression**.
- $\rho = q\det$ is a 3-form. Then $\star((\partial/\partial t + \mathscr{L}_{\mathbf{v}})\rho) = \partial q/\partial t + \mathbf{v}\cdot\nabla q + (\nabla\cdot\mathbf{v})q = \partial q/\partial t + \nabla\cdot(q\mathbf{v})$, known as the conservative material derivative with compression, or the derivative in the conservative form.

These various "material derivatives with additional terms" show up frequently in continuum mechanics. It is a fun exercise identifying them as the Lie material derivative of differential form of a particular degree. Note that this list is just a glimpse into various Lie derivatives. We have only talked about Lie derivatives of differential forms. We have not talked about Lie derivatives on other tensors such as vectors, vector-valued forms, *etc*.

6.8 Summary

Important measurements such as line- and surface-integrals for circulations and fluxes motivate us to study differential forms. Infinitesimally they should be treated as alternating multilinear forms. Macroscopically, they are objects waiting to be integrated along a geometry of a certain dimension.

A 1-form is a covector. k-forms are linear combinations of k-fold wedge product of 1-forms.

Covectors and vectors can be treated like the same objects only when there is a metric or there is a full set of basis. The metric-dependent conversion between vectors and covectors is the musical isomorphism ($\sharp$ and $\flat$). The basis-related conversion is that give a full set of basis $(\mathbf{e}_j)_j$ for a vector space, then there is a canonical dual basis $(\mathbf{e}_i^\star)_i$ for the covector space that pairs with the primal basis into identity: $\mathbf{e}_i^\star(\mathbf{e}_j) = \delta_{ij}$.

The common operations for k-forms are wedge $\wedge$, pullback A^* by a map, and interior product $i_{\mathbf{v}}$ with a vector $\mathbf{v}$. When there is a metric, we can map a k-form to a $(k-1)$-form by the Hodge star $\star$. When it comes to k-form field, we can take exterior derivative d.

We can build codifferential δ and Laplacian Δ using d and $\star$.

Facts involving d and $i_{\mathbf{v}}$: suppose α is a k-form

- $d(\alpha\wedge\beta) = (d\alpha)\wedge\beta + (-1)^k\alpha\wedge d\beta$.
- $i_{\mathbf{v}}(\alpha\wedge\beta) = (i_{\mathbf{v}}\alpha)\wedge\beta + (-1)^k\alpha\wedge i_{\mathbf{v}}\beta$.
- $dd\alpha = 0$.
- $i_{\mathbf{v}}i_{\mathbf{v}}\alpha = 0$.
- $(-1)^k \star^{-1} d\star\alpha = \delta\alpha$.
- $(-1)^k \star^{-1} i_{\mathbf{v}}\star\alpha = \mathbf{v}^\flat\wedge\alpha$.

Duality between maps and pullbacks, and between boundary and exterior derivative:

- $\int_\Sigma d\alpha = \int_{\partial\Sigma}\alpha$.
- $\int_\Sigma \mathbf{f}^*\alpha = \int_{\mathbf{f}(\Sigma)}\alpha$.

For a time-dependent flow map $\phi_t : M\to M$

- $\int_\Sigma i_{\dot\phi}\alpha = \frac{d}{dt}\int_{\mathrm{ext}^t_\phi(\Sigma)}\alpha$;
- $\int_\Sigma \mathscr{L}_{\dot\phi}\alpha = \frac{d}{dt}\int_{\phi_t(\Sigma)}\alpha$.

Here the Lie derivative satisfies Cartan's magic formula

- $\mathscr{L}_{\mathbf{v}} = i_{\mathbf{v}} \circ d + d \circ i_{\mathbf{v}}$.

7. Discrete Exterior Calculus

In this chapter we will talk about taking the smooth concepts we learned about so far and applying them to the discrete setting of meshes. We will also talk about interpolation of discrete k-forms and conversely the sampling of smooth k-forms for the discrete setting.

The following treatment will be independent of dimension, though the reader is invited to visualize surface meshes (oriented triangle mesh) and/or volume meshes (tetrahedral meshes) as the most common case.

7.1 Discrete k-forms

Differential k-forms are to be integrated along a k-dimensional surface. In the discrete setting, we represent a k-form by its integral along each k-dimensional face. For example,

- a discrete 0-form is a function defined over each vertex. The value on a vertex represent the 0-form evaluated at the point.
- a discrete 1-form is a function defined over each oriented edge. The value on an edge represent the 1-form integrated along the edge.
- a discrete 2-form is a function defined over each oriented face. The value on a face represent the 2-form integrated over the face.

In general an oriented mesh $M = (V, E, F,$ or higher dimensional cells) is made of vertices V, oriented edges E, oriented faces F, *etc.*, with a certain incidence relation. A more general notation is $M = (M_0, M_1, M_2, \ldots, M_n)$ where M_k is called the **k-skeleton**. We assume that when all these various-dimensional cells are stitched together they form a manifold (*cf.* Section 4.2.2). A discrete k-form is a function $\alpha \colon M_k \to \mathbb{R}$; since it is also a assigning a real number per element in M_k we also view it as an array $\alpha \in \mathbb{R}^{|M_k|}$. Here $|M_k|$ is the number of elements in M_k.

7.1.1 Chains and Cochains

In the smooth setting, an integral such as $\int_\Sigma \alpha$ involves two type of objects: a k-dimensional shape Σ and a k-form α. As mentioned, k-forms are values assigned on each k cell. What about a k-dimensional set Σ?

One can view Σ as a subset of $\mathcal{M}_k$: some k-faces are turned on and the rest are turned off, displaying as Boolean values over $\mathcal{M}_k$. Due to possibly opposite orientation, one may see that it is more appropriate to view Σ as some $\{-1, 0, 1\}$-valued function over $\mathcal{M}_k$. But there are also situations where Σ is given by the sum of many overlapping surfaces. In that case Σ is ultimately an integer-valued function defined over $\mathcal{M}_k$. We further relax the integer values and replace it by real values (think of rasterization with anti-aliasing). So, finally, we take Σ as a real-valued function over $\mathcal{M}_k$; that is, $\Sigma \in \mathbb{R}^{|\mathcal{M}_k|}$.

Therefore, in the discrete setting, there is no actual distinction between a k-dimensional shape Σ and a k-form. However, they do represent different concepts in the smooth setting. So we distinguish them by calling a k-dimensional surface a **k-chain** and calling a k-form a **k-cochain**.

Definition 7.1 A ($\mathbb{R}$-valued) **k-chain** Σ is a real-valued function defined over the k-skeleton $\mathcal{M}_k$ indicating the (oriented) occupation of Σ at each k-cell. The space of all k-chains is a linear space denoted by

$$C_k(\mathcal{M}) = \left\{ \sum_{e_i \in \mathcal{M}_k} c_i e_i \,\middle|\, c_i \in \mathbb{R} \right\} \cong \mathbb{R}^{|\mathcal{M}_k|}. \tag{7.1}$$

Here each e_i is a k-cell. They form a natural basis for $C_k(\mathcal{M})$.

Definition 7.2 The space of ($\mathbb{R}$-valued) **k-cochain** $C^k(\mathcal{M})$ is the dual space of the space of k-chain

$$C^k(\mathcal{M}) = C_k(\mathcal{M})^\star = \left\{ \alpha \colon C_k(\mathcal{M}) \xrightarrow{\text{linear}} \mathbb{R} \right\} = \left\{ \sum_{e_i \in \mathcal{M}_k} \alpha_i \mathbb{1}_{e_i} \,\middle|\, \alpha_i \in \mathbb{R} \right\} \cong \mathbb{R}^{|\mathcal{M}_k|}. \tag{7.2}$$

A k-cochain $\alpha \in C^k(\mathcal{M})$ being a linear functional $\alpha \colon C_k(\mathcal{M}) \xrightarrow{\text{linear}} \mathbb{R}$ means that it is to-be-evaluated along a k-chain. We denote

$$\alpha(\Sigma) = \langle\!\langle \alpha | \Sigma \rangle\!\rangle = \int_\Sigma \alpha, \quad \alpha \in C^k(\mathcal{M}), \Sigma \in C_k(\mathcal{M}). \tag{7.3}$$

In (7.2), $\mathbb{1}_{e_i}$ denote the dual basis from the k-cell $(e_i)_i$ (the basis of $C_k(\mathcal{M})$). That is,

$$\mathbb{1}_{e_i}(e_j) = \int_{e_j} \mathbb{1}_{e_i} = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

A k-cochain is also called a **discrete k-form**.

7.2 Discrete Differentials

Our main task is the appropriate definition of the differential. We employ Stokes Theorem to *define* the differential in terms of the boundary operator

$$\int_{\Sigma} d\omega := \int_{\partial\Sigma} \omega. \tag{7.4}$$

Thinking of integration as the bilinear pairing of k-domains ($C_k(M)$) and k-forms ($C^k(M)$) the differential is the adjoint of the boundary operator:

$$\langle\!\langle \Sigma | d\omega \rangle\!\rangle := \langle\!\langle \partial\Sigma | \omega \rangle\!\rangle. \tag{7.5}$$

Thus the **discrete exterior derivative** is also known as the **co-boundary operator** under the integral pairing between chains and cochains.

The **boundary operator** ∂ is a linear map between chains

$$\partial_k \colon C_k(M) \to C_{k-1}(M). \tag{7.6}$$

It sends each basis element $\sigma \in M_k$, *i.e.* each k-cell, of $C_k(M)$ to the sum of its boundary $(k-1)$-cells:

$$\partial_k \sigma = \sum_{\substack{\tau \in M_{k-1} \\ \tau < \sigma}} \pm\tau, \tag{7.7}$$

where the sign $\pm$ is chosen so that $\pm\tau$ is consistent with the orientation of σ.

R The boundary of an oriented simplex (2-simplex is a triangle, 3-simplex is a tetrahedron, 4-simplex is a 5-cell, *etc.*) has an explicit formula. Let $\sigma = [i_0, \ldots, i_k] \in M_k$ be a k-simplex, given by its $k+1$ vertices $i_0, \ldots, i_k \in M_0$. Then

$$\partial\sigma := \sum_{j=0}^{k} (-1)^j [i_0, \ldots, \widehat{i_j}, \ldots, i_k]. \tag{7.8}$$

In other words, the boundary of a k-simplex consists of the $(k-1)$-simplices one gets by dropping (indicated by the hat notation) one of the indices and adjusting the sign.

The boundary operator $\partial_k \colon C_k(M) \to C_{k-1}(M)$ can thus be represented by a $|M_{k-1}|$-by-$|M_k|$ matrix $\partial_k \colon \mathbb{R}^{|M_k|} \to \mathbb{R}^{|M_{k-1}|}$ with entries taking values in $\{-1, 0, 1\}$. The nonzero entries indicate the incidence relation between the corresponding k-cell and $(k-1)$-cell.

The co-boundary operator $d_{k-1} \colon C^{k-1}(M) \to C^k(M)$ in its matrix representation is a $|M_k|$-by-$|M_{k-1}|$ matrix given by

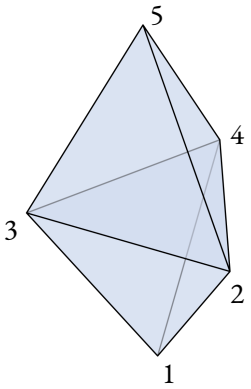
$$d_{k-1} = \partial_k^{\mathsf{T}} \colon \mathbb{R}^{|M_{k-1}|} \to \mathbb{R}^{|M_k|}. \tag{7.9}$$

To get a better sense as to what this all means imagine an implementation. Assume all cells in the mesh are simplices. Each k-form is a vector of coefficients, each coefficient corresponding to the value associated with a particular k-cell according to some arbitrary but fixed numbering scheme. This setup turns the discrete d operators into a sequence of sparse matrices d_0, d_1, *etc.*,

which are transposes of incidence matrices ∂'s. For example, the boundary operator for edges, ∂_1, has dimension $|V| \times |E|$ with a single -1 and $+1$ in each column denoting the beginning resp. end of a given edge. The differential of discrete 0-forms is then $d_0 = \partial_1^\mathsf{T}$. Applying d_0 to a function now amounts to taking the difference of values of the function along each oriented edge. This difference is not an approximation of the differential, but is exact by virtue of us using *integrated forms* and the fundamental theorem of calculus (Stokes theorem in 1D). Similarly, the boundary operator ∂_2 for facets (triangles) is of dimension $|E| \times |F|$ with each column containing three non-zero entries of ± 1 according to the intrinsic orientation of each edge relative to the orientation of the containing triangle

$$(\partial_2)_{[jk],[ijk]} = +1, \quad (\partial_2)_{[ik],[ijk]} = -1, \quad (\partial_2)_{[ij],[ijk]} = +1, \tag{7.10}$$

leading to $d_1 = \partial_2^\mathsf{T}$. Here we used simpllices to index the rows and columns of the associated matrix. This pattern of construction continues as we go up in dimension. The important part is that the d matrices are sparse and their entries are easily derived from incidence relations between the elements of the mesh.



$$V = \begin{pmatrix} [1] \\ [2] \\ [3] \\ [4] \\ [5] \end{pmatrix} \quad E = \begin{pmatrix} [12] \\ [13] \\ [14] \\ [23] \\ [24] \\ [25] \\ [34] \\ [35] \\ [45] \end{pmatrix} \quad F = \begin{pmatrix} [123] \\ [124] \\ [134] \\ [234] \\ [235] \\ [245] \\ [345] \end{pmatrix} \quad T = \begin{pmatrix} [1234] \\ [2345] \end{pmatrix}$$

Figure 7.1 A mesh consisting of two tetrahedra together with the enumeration of all its simplices.

As an example consider the tetrahedral mesh in Figure 7.1. There we used the convention of sorted index order as the canonical representative of a given simplex. Since discrete k-forms are just maps from a simplex to a number we can store them as vectors with the same order as the arrays of simplices. Using this ordering we get the incidencce matrices, and hence their transposes, the discrete d operators, as

$$\partial_3 = \begin{bmatrix} -1 & 0 \\ +1 & 0 \\ -1 & 0 \\ +1 & -1 \\ 0 & +1 \\ 0 & -1 \\ 0 & +1 \end{bmatrix}, \quad \partial_2 = \begin{bmatrix} +1 & +1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & +1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & +1 & +1 & 0 & 0 \\ 0 & +1 & 0 & -1 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & +1 & +1 & 0 & 0 & +1 \\ 0 & 0 & 0 & 0 & +1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & +1 & +1 \end{bmatrix}$$

$$\partial_1 = \begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & +1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & +1 & 0 & +1 & 0 & +1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & +1 & 0 & +1 & +1 \end{bmatrix}$$

▪ **Example 7.1 — Building d_0.** One practical construction of d_0 is the following. Let $\{0, \ldots, |V|-1\}$ be the index set of vertices and $\{0, \ldots, |E|-1\}$ be the index set of edges. Suppose for each edge we have access to its source and destination vertices

$$\text{src}\colon \{0, \ldots, |E|-1\} \to \{0, \ldots, |V|-1\}, \tag{7.11}$$
$$\text{dst}\colon \{0, \ldots, |E|-1\} \to \{0, \ldots, |V|-1\}. \tag{7.12}$$

Let us take this information as arrays

$$\text{src} = \begin{bmatrix} \text{src}_1 \\ \vdots \\ \text{src}_{|E|-1} \end{bmatrix}, \quad \text{dst} = \begin{bmatrix} \text{dst}_1 \\ \vdots \\ \text{dst}_{|E|-1} \end{bmatrix} \tag{7.13}$$

Then, for each edge $j \in \{0, \ldots, |E|-1\}$ we assign $(d_0)_{j,\text{src}_j} = -1$ and $(d_0)_{j,\text{src}_j} = 1$.

This for loop for matrix element assignment can often be replaced by a function call that constructs a sparse matrix:

$$d_0 = \textsc{Sparse}\left(\textsc{row} = \begin{bmatrix} 0 \\ \vdots \\ |E|-1 \\ 0 \\ \vdots \\ |E|-1 \end{bmatrix}, \textsc{col} = \begin{bmatrix} | \\ \text{src} \\ | \\ | \\ \text{dst} \\ | \end{bmatrix}, \textsc{val} = \begin{bmatrix} -1 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \textsc{size} = (|E|, |V|) \right).$$

▪

7.3 Discrete Hodge Star

Another discrete operator we need is the Hodge star. Recall that the Hodge star maps a k-form to an $(n-k)$-form. While the former is evaluated on a k-dimensional domain, the latter is evaluated on an $(n-k)$-dimensional domain. In the discrete setting we begin by creating this pairing between k-dimensional and $(n-k)$-dimensional pieces of domain. This is accomplished via the dual mesh.

7.3.1 Dual Mesh

Given a mesh $M = (M_0, M_1, \ldots, M_n)$ we consider a dual mesh $M^\star = (M_0^\star, M_1^\star, \ldots, M_n^\star)$. Let us first take an example where M is a triangle mesh $M = (V, E, F)$ with no boundary. Take each triangle $f \in F$ as the vertex of $M^\star$, *i.e.* $V^\star := F$. Then we connect vertices in $V^\star$ together if they correspond to adjacent triangles. Since each adjacent pair of triangles also correspond to an edge separating them, the dual edge set equals to the primal edge set $E^\star = E$. Note that the dual edge $e^\star \in E^\star$ is pointing from one triangle to another, while the corresponding primal edge $e \in E$ is pointing from one vertex to another vertex in that neighborhood. We will choose the orientation of $e^\star$ in such a way that $(e, e^\star)$ form a positively oriented basis in that order. Finally, each vertex in V is associated with a polygon surrounded by the dual edges. That is, $F^\star := V$.

More generally, for a mesh $M = (M_0, \dots, M_n)$ with no boundary, we let $M^\star = (M_0^\star, \dots, M_n^\star)$ where $M_k^\star = M_{n-k}$, assigned with an appropriate orientation for each element.

R If M is a mesh with boundary, then there are different ways one takes the dual mesh depending on the boundary condition of the problem. For example, one may take $M_k^\star \subseteq M_{n-k}$ by throwing away all the primal $(n-k)$-cells that lie completely in ∂M.

Definition 7.3 The space of **dual k-chains** is denoted by

$$C_k(M^\star) = \left\{ \sum_{e_i^\star \in M_k^\star} c_i e_i^\star \;\middle|\; c_i \in \mathbb{R} \right\} \cong C_{n-k}(M) \cong \mathbb{R}^{|M_k^\star|} = \mathbb{R}^{|M_{n-k}|}. \tag{7.14}$$

The space of **discrete dual k-forms** (**dual k-cochains**) $C^k(M^\star)$ is given by

$$C^k(M^\star) = C_k(M^\star)^\star = \left\{ \sum_{e_i^\star \in M_k^\star} \alpha_i \mathbb{1}_{e_i^\star} \;\middle|\; \alpha_i \in \mathbb{R} \right\} \cong C^{n-k}(M) \cong \mathbb{R}^{|M_k^\star|} = \mathbb{R}^{|M_{n-k}|}. \tag{7.15}$$

The incidence matrices (boundary operators) of the dual mesh, $\partial^\star$, are given as

$$\begin{aligned} &\partial_k^\star \colon C_k(M^\star) \to C_{k-1}(M^\star) \\ &\partial_k^\star = (-1)^{n-k+1} \partial_{n-k+1}^\intercal, \quad k = 1, \dots, n. \end{aligned} \tag{7.16}$$

Consequently, the dual d operators are given by

$$\begin{aligned} &d_k^\star \colon C^k(M^\star) \to C^{k+1}(M^\star) \\ &d_k^\star = (\partial_{k+1}^\star)^\intercal = (-1)^{n-k} \partial_{n-k} = (-1)^{n-k-1} d_{n-k}^\intercal. \end{aligned} \tag{7.17}$$

In other words, the incidence matrices of the primal mesh are all that is needed to define the discrete differential both on the primal mesh and the dual mesh.

R The differential for the dual forms will be used when defining codifferentials $\delta_k = (-1)^k \star^{-1} d \star$. The rather complicated signs here will be resolved in that context.

7.3.2 Diagonal Hodge Star

So far we have only defined the dual mesh combinatorially, and not given it an embedding. The dual mesh is not ways required to be embedded. If we need an embedding there are two common choices for simplicial meshes. One places the dual vertex at the barycenter of its containing highest-dimensional simplex, the other places it at the center of the circumsphere of the containing simplex.

Having the correspondence $C^k(M) \cong C^{n-k}(M^\star)$ between the domain for the primal k-forms and dual $(n-k)$-forms, we now describe how the values of discrete forms transform under the **discrete Hodge star**

$$\star_k \colon C^k(M) \to C^{n-k}(M^\star). \tag{7.18}$$

There are a number of different possible approaches and we use something called the **diagonal Hodge star**. Given an embedding of the dual mesh, the diagonal Hodge star ensures that the integral over the respective pieces of domain agree on average:

$$\frac{1}{|\sigma|}\int_{\sigma}\alpha = \frac{1}{|P_{\sigma^\perp}\sigma^\star|}\int_{\sigma^\star}\star\alpha, \tag{7.19}$$

where $|\cdot|$ denotes the intrinsic volume, σ an appropriate cell, $\sigma^\star$ its dual cell, and $P_{\sigma^\perp}$ denotes the projection to the orthogonal complement $\sigma^\perp$ of σ. Here the orientation $(\sigma, \sigma^\star)$ (concatenating their bases) is positive. The intrinsic volume

- of a vertex is 1;
- of an edge is its length;
- of a facet is its area;
- of a tetrahedron is its volume; *etc.*

Correspondingly the components α_σ of a discrete form $\alpha \in C^k(\mathcal{M})$ that represents the integral $\int_\sigma \alpha$, are mapped to the components $(\star\alpha)_{\sigma^\star}$ as

$$(\star\alpha)_{\sigma^\star} := \frac{|P_{\sigma^\perp}\sigma^\star|}{|\sigma|}\alpha_\sigma. \tag{7.20}$$

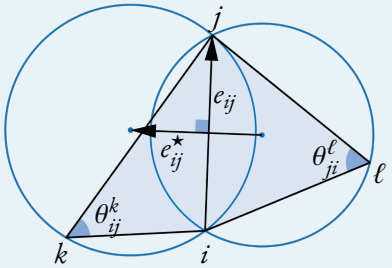
This relationship corresponds to a diagonal matrix whose entries are the quotients of intrinsic volumes of k-dimensional cells and their duals

$$\begin{aligned}\star_k &: \mathbb{R}^{|\mathcal{M}_k|} \to \mathbb{R}^{|\mathcal{M}^\star_{n-k}|} = \mathbb{R}^{|\mathcal{M}_k|}\\ (\star_k)_{\sigma^\star,\sigma} &= \frac{|P_{\sigma^\perp}\sigma^\star|}{|\sigma|}, \quad (\star_k^{-1})_{\sigma,\sigma^\star} = \frac{|\sigma|}{|P_{\sigma^\perp}\sigma^\star|}.\end{aligned} \tag{7.21}$$

Exercise 7.1 — The cotangent Hodge star. Consider an embedded surface triangle mesh and use circumcentric duals, *i.e.* place the dual vertex of a triangle at the center of its circumcircle. Note that the circumcenter of a triangle is the intersection of all three edge perpendicular bisectors, so the dual edges is already orthogonal to the primal edges ($P_{e_{ij}^\perp}(e_{ij}^\star) = e_{ij}^\star$). Show that in that case

$$\frac{|e_{ij}^\star|}{|e_{ij}|} = \frac{1}{2}\left(\cot\theta_{ij}^k + \cot\theta_{ji}^\ell\right) \tag{7.22}$$

for edge e_{ij} shared by facets f_{ijk} and $f_{ji\ell}$. The angle $\alpha_{ij}^k = {}_k\angle_i^j$ is across from e_{ij} at vertex k, and correspondingly for α_{ji}^ℓ. Under what circumstances is the sum of the two cotans negative?



7.4 Discrete Codifferential and Discrete Laplacian

Given that we now have both a discrete d and a discrete $\star$ we can also build the **discrete codifferential**. Recall its definition for a k-form in the smooth setting

$$\delta_k = (-1)^k \star_{k-1}^{-1} d_{n-k} \star_k . \tag{7.23}$$

In the discrete setting this amounts to the product

$$\delta_k := (-1)^k \star_{k-1}^{-1} d_{n-k}^{\star} \star_k = \star_{k-1}^{-1} d_k^{\mathsf{T}} \star_k . \tag{7.24}$$

That is,

$$\begin{array}{ccccccccccccc}
C^0(M) & \xrightarrow{d_0} & C^1(M) & \xrightarrow{d_1} & C^2(M) & \xrightarrow{d_2} & C^3(M) & \longrightarrow\cdots\longrightarrow & C^{n-1}(M) & \xrightarrow{d_{n-1}} & C^n(M) \\
\downarrow{\scriptstyle \star_0} & \overset{\delta_1}{\nearrow} & \downarrow{\scriptstyle \star_1} & \overset{\delta_2}{\nearrow} & \downarrow{\scriptstyle \star_2} & \overset{\delta_3}{\nearrow} & \downarrow{\scriptstyle \star_3} & & \downarrow{\scriptstyle \star_{n-1}} & \overset{\delta_n}{\nearrow} & \downarrow{\scriptstyle \star_n} \\
C^n(M^\star) & \xleftarrow[(=-d_{n-1}^\star)]{d_0^{\mathsf{T}}} & C^{n-1}(M^\star) & \xleftarrow[(=d_{n-2}^\star)]{d_1^{\mathsf{T}}} & C^{n-2}(M^\star) & \xleftarrow[(=-d_{n-3}^\star)]{d_2^{\mathsf{T}}} & C^{n-3}(M^\star) & \longleftarrow\cdots & C^1(M^\star) & \xleftarrow[(=(-1)^n d_0^\star)]{d_{n-1}^{\mathsf{T}}} & C^0(M^\star)
\end{array} \tag{7.25}$$

The smooth Laplacian

$$\Delta_k = -\delta_{k+1} d_k - d_{k-1} \delta_k \tag{7.26}$$

amounts to the discrete operator $\Delta_k \colon C^k(M) \to C^k(M)$,

$$\Delta_k = - \star_k^{-1} d_k^{\mathsf{T}} \star_{k+1} d_k - d_{n-k}^{\mathsf{T}} \star_{k-1}^{-1} d_{n-k} \star_k . \tag{7.27}$$

Definition 7.4 — Discrete (symmetric positive-semidefinite) Laplace matrix. By multiplying both sides of (7.27) by $-\star_k$ we define

$$\begin{aligned} &L_k \colon C^k(M) \to C^{n-k}(M), \\ &L_k = - \star_k \Delta_k = d_k^{\mathsf{T}} \star_{k+1} d_k + \star_k d_{n-k}^{\mathsf{T}} \star_{k-1}^{-1} d_{n-k} \star_k . \end{aligned} \tag{7.28}$$

Note that this Laplace matrix L_k is symmetric $L_k^{\mathsf{T}} = L_k$ and positive (semi)definite.

7.4.1 An Example: the Laplacian for 0-fomrs

To get a little practice with these discrete operators, let us consider the Laplacian for 0-forms on a triangle mesh. Suppose we are given an orientable 2-manifold simplicial complex $M = (V, E, F)$ (without boundary) together with an embedding (vertex position) $(\mathbf{p}_i \in \mathbb{R}^3)_{i \in V}$. Let us place the dual vertices at the circumcenter of their primal triangle circumcircles. This is all we need to define the discrete differential on the primal and dual mesh as well as the diagonal Hodge star operators. In the smooth setting the so-called **Poisson equation** is written as

$$-\Delta u = g, \quad \text{that is} \quad \delta d u = g, \tag{7.29}$$

for $u, g \in \Omega^0(M)$. (The Poisson problem is to solve u for a given g.) The corresponding discrete linear system is given by

$$L_0 u = \star_0 g \quad \Longleftrightarrow \quad d_0^\mathsf{T} \star_1 d_0 = \star_0 g. \tag{7.30}$$

Note that we moved the leftmost Hodge star in the definition of Δ to the right hand side so as to get a symmetric linear system, which admits much faster solution methods than a general matrix. In the discrete setting, let $(\star g_i)_{i \in V}$ be the array of integrals of g over the dual 2-cells (dual to each primal vertex), *i.e.* $\star g_i = \int_{[i]^\star} \star g$, and $(u_i)_{i \in V}$ represent the discrete primal 0-form to be solved for. We then solve the linear algebra problem

$$\underbrace{d_0^\mathsf{T} \star_1 d_0}_{L_0} u = \star g \tag{7.31}$$

where $\star_1$ is the diagonal matrix with the entry corresponding to edge e given by

$$\frac{|e_{ij}^\star|}{|e_{ij}|} = \frac{1}{2}\left(\cot \theta_{ij}^k + \cot \theta_{ji}^\ell\right). \tag{7.32}$$

Here θ_{ij}^k and θ_{ji}^ℓ are the interior angle at the opposite vertex for each of the two triangles adjacent to e_{ij}. See Exercise 7.1.

This Laplace matrix $L_0 = d_0^\mathsf{T} \star_1 d_0$ is called the **cotan Laplacian**.

8. Differential Equations on Manifolds

In this chapter we look at a few classical linear **partial differential equations** (**PDEs**) formulated using exterior calculus. They include the heat equation, the Poisson equation, and the Maxwell's equations.

These differential equations originate from classical physics and are often formulated in coordinates. For example, the Laplace operator Δ, which we will see frequently in the equations we consider, can be expressed as $\Delta u = \frac{\partial^2}{\partial x^2}u + \frac{\partial^2}{\partial y^2}u + \frac{\partial^2}{\partial z^2}u$ for scalar functions $u\colon \mathbb{R}^3 \to \mathbb{R}$. However, when the coordinate basis $\left(\frac{\partial}{\partial x_i}\right)_{i=1}^n$ one uses is not orthonormal everywhere, which is often the case on a manifold, the coordinate expression for fundamental operators such as the Laplacian becomes impenetrable:[1]

$$\Delta u = \frac{1}{\sqrt{\det g}} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial}{\partial x_i}\left(\sqrt{\det g}(g^{-1})_{ij}\frac{\partial}{\partial x_j}u\right), \quad g_{ij} = \left\langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right\rangle,$$

to say nothing of its discretization for computation. Compare it with the coordinate-free exterior calculus expression

$$\Delta u = -\delta du,$$

whose simplicity is not influenced by the choice of any coordinate system. We use this latter geometric language to go through the physical modeling, analysis, to numerical methods (via exterior calculus) aspects of a few PDEs.

Physical Units

Many of the equations are derived from physics. We will write the precise physical unit in these cases. A differential k-form is regarded as an *oriented k-density*, namely it describes some physical

[1] Here $\partial/\partial x_i$ is the shorthand for the tangent vector $\mathbf{e}_i = \partial/\partial x_i$ so that the directional derivative $df(\mathbf{e}) = \partial f/\partial x_i$.

quantity per unit k-dimensional volume. For example, in $\mathbb{R}^3$, the mass density ρ is a 3-form; it becomes a mass after integrated over a volume. We denote

$$\rho \in \Omega^3(M; \mathbb{R}\,\mathrm{kg}), \qquad \int_V \rho \in \mathbb{R}\,\mathrm{kg}$$

and read that ρ is a *mass*-valued (or kg-valued) 3-form. The actual physical dimension of ρ as a mass density is

$$[\rho] = [\mathrm{kg}/\mathrm{m}^3]$$

since it has to be integrate over a 3-dimensional region to become an element of $\mathbb{R}$ kg. Infinitesimally, one has to take 3 tangent vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ to plug in into the alternating triple linear form ρ to evaluate the mass, and tangent vectors (viewed as displacements) have units of meter.

Another example is that a mass flux $J \in \Omega^2(M; \mathbb{R}\,\mathrm{kg/s})$ is a 2-form modeling the amount of mass passing through a surface per unit area per unit time. The unit of J is $[J] = [\mathrm{kg}/\mathrm{m}^2\mathrm{s}]$, and its evaluation after integrating over an area is a number in $\mathbb{R}\,\mathrm{kg/s}$.

In general, a differential k-form $\omega \in \Omega^k(M; \mathbb{R}\,\mathrm{unit})$ has a unit $[\omega] = [\mathrm{unit}/\mathrm{m}^k]$.

The exterior derivative d maps $\Omega^k(M; \mathbb{R}\,\mathrm{unit})$ to $\Omega^{k+1}(M; \mathbb{R}\,\mathrm{unit})$. In other words, the exterior derivative gives an extra "length" in the denominator, $[d\omega] = [\omega][1/\mathrm{m}]$. For this being said, in $\mathbb{R}^n$ the coordinate functions $x_1, \dots, x_n$, which are meter-valued functions $: M \to \mathbb{R}\,\mathrm{m}$, gives a basis for k-form $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_k}$ that belongs to $\Omega^k(M; \mathbb{R}\,\mathrm{m}^k)$. Therefore $[dx_I] = [1]$ has no unit. The unit of $\omega = \sum_I w_I\, dx_I$ is the same as that of the coefficient functions w_I.

The standard Hodge star $\star$ has no unit, since $\star\omega = \sum_I w_I(\star dx_I)$ has the same unit as w_I and thus the same as the unit of ω. Therefore

$$\star_k \colon \Omega^k(M; \mathbb{R}\,\mathrm{unit}) \to \Omega^{n-k}(M; \mathbb{R}\,\mathrm{unit} \cdot \mathrm{m}^{n-2k})$$

in order to give $[\omega] = [\star\omega] = [\mathrm{unit}/\mathrm{m}^k]$.

In many physical examples, we will take the Hodge star as a more general operator that can map differential forms of different unit of integrals:

$$\star_k \colon \Omega^k(M; \mathbb{R}\,\mathrm{unit}_1) \to \Omega^{n-k}(M; \mathbb{R}\,\mathrm{unit}_2)$$

encoding a material property that relates $\mathrm{unit}_1/\mathrm{m}^k$ with $\mathrm{unit}_2/\mathrm{m}^{n-k}$.

8.1 Heat Equation

The heat equation, a.k.a. diffusion equation, models how temperature diffuses in a material. Consider the domain to be an n-dimensional manifold M, and at each $t \geq 0$ let $u(t) \colon M \to \mathbb{R}^+\,\mathrm{K}$ be the temperature (in Kelvin) at each point on M.

Temperature Difference Induces Heat Flux

The 1-form $du \in \Omega^1(M; \mathbb{R}\,\mathrm{K})$ describes the temperature difference: $\int_\gamma du$ is the temperature difference between the two ends of the path γ. **Fourier's Law of thermal conductance** (or **Newton's Law of cooling**) states that the time rate of heat transfer is proportional to the (negative) temperature difference. Namely, the **heat flux**, the energy flowing through a surface per unit area per unit time, is a $(n-1)$-form $Q(t) \in \Omega^{n-1}(M; \mathbb{R}\,\mathrm{J/s})$ that is related to du through

$$Q = -\star du. \tag{8.1}$$

Here the Hodge dual $\star = \star_1$ maps a *temperature*-valued 1-form to an *energy-per-second*-valued $(n-1)$-form. In particular $\star_1$ describes the **thermal conductivity**.

To illustrate this, we consider the case of isotropic and uniform thermal conductivity in the material. In this case, we look at a primal edge e and its corresponding dual facet $e^\star$ in a discrete setup, and discover

$$\int_{e^\star} Q = -\int_{e^\star} \star du = -\frac{|P_{e^\perp} e^\star|}{|e|} \int_e du.$$

On the left hand side we have the total heat transfer per unit time, which is negative proportional to the temperature difference on the right hand side. The conductance rate is $|P_{e^\perp} e^\star|/|e|$, which is proportional to the cross section area $|P_{e^\perp} e^\star|$ and inverse proportional to the distance $|e|$. This is the familiar empirical thermal conductivity described in terms of the material geometry.

The above example uses $\star = \star_1$ derived from the Riemannian metric on M. In general one can take $\star = \star_1$ in $Q = -\star\, du$ at each point p an arbitrary linear map

$$\star_1 \colon \Omega^1_p(M; \mathbb{R}\,\mathrm{K}) \to \Omega^{n-1}_p(M; \mathbb{R}\,\mathrm{J/s})$$

with positive orientation, *i.e.* $(\alpha \wedge \star_1 \alpha)$ is a positive n-form. It encodes a general thermal conductivity in a material.

Temperature Changes Due to Heat Flux

With a heat flux Q given, the change in temperature over time is given by

$$\star \frac{\partial}{\partial t} u = -dQ. \tag{8.2}$$

Here $\star = \star_0$ maps changes in temperature to heat flux. That is,

$$\star_0 \colon \Omega^0(M; \mathbb{R}\,\mathrm{K/s}) \to \Omega^n(M; \mathbb{R}\,\mathrm{J/s})$$

is the **heat capacity**. In a material with uniform specific heat, on a small region U

$$-\int_{\partial U} Q = \int_U \star \frac{\partial}{\partial t} u \approx |U| \frac{\partial}{\partial t} u$$

which states that the total in-flowing energy (LHS) results in a rise in temperature (RHS) with a rate proportional to the volume $|U|$. In general, at each $p \in M$, $\star_0$ can be an arbitrary linear map $\star_0 \colon \Omega^0(M; \mathbb{R}\,\mathrm{K/s}) \to \Omega^n(M; \mathbb{R}\,\mathrm{J/s})$ with positive $\star_0 1$, modeling a general local heat capacity in a material.

The Heat Equation

Putting (8.1) (Newton–Fourier's Law) and (8.2) (conservation of energy) together, one obtains the **heat equation**

$$\frac{\partial}{\partial t} u = \star_0^{-1} d \star_1 du = \Delta u. \tag{8.3}$$

Exercise 8.1 — Conservation of energy. Here is a quick exercise for using Stokes' Theorem. Suppose M has no boundary. Show that the total internal (thermal) energy $\int_M \star_0 u$ is constant over time when u satisfies (8.3). ▪

If there is a heat source, namely there is $\beta(t) \in \Omega^n(M; \mathbb{R}\,\mathrm{J/s})$ being the rate of heat transferred into the domain per unit volume and per unit time, then (8.2) is replaced with

$$\star \frac{\partial}{\partial t} u = -dQ + \beta$$

and (8.3) becomes

$$\frac{\partial}{\partial t} u = \Delta u + \star_0^{-1} \beta. \tag{8.4}$$

Analytic Aspects

How should the temperature u behave subject to the heat equation? Intuitively, the heat keeps flowing from high temperature to low temperature, which effectively keeps averaging the temperature. This process is also called a **diffusion process**. As the averaging process suggests, we expect the following properties for the heat equation, which will be shown as the discussion goes on.

- **Maximum principle**. For u satisfying the heat equation for $t \geq 0$, the maximum and minimum temperature occurs at $t = 0$ or on the boundary ∂M.
- **Time irreversibility**. The total entropy increases in time.
- **Dirichlet energy** $\int_M \star |du|^2 = \int_M du \wedge \star du$, which measures the roughness of the function u, decreases under heat flow.

8.1.1 Maximum Principle

The maximum principle for the heat equation states that the maximum of the temperature occurs initially or on the boundary, as expected in an averaging process. With the maximum/minimum principle, we justify that the temperature in Kelvin $u \geq 0$ is not violated if $u \geq 0$ initially and on the boundary.

Theorem 8.1 — Maximum principle. Suppose u satisfies (8.3) for $0 < t \leq T$ on a compact manifold M. Then the maximum of u occurs on $t = 0$ or on ∂M.

Proof. First of all, by compactness of $M \times [0, T]$ the maximum is always attained. Let us first prove a slight variation of the statement: if $\tilde{u}$ satisfies

$$\frac{\partial}{\partial t} \tilde{u} < \Delta \tilde{u}, \quad \text{for } 0 < t \leq T,$$

then its maximum must occurs on $t = 0$ or on ∂M. To show this, assume that there is an interior point $p_0 \in M$ and $0 < t_0 \leq T$ so that $\tilde{u}_{(p_0,t_0)}$ is a maximum. Then $\frac{\partial \tilde{u}}{\partial t}_{(p_0,t_0)} \geq 0, d\tilde{u}_{(p_0,t_0)} = 0$ and $\Delta u_{(p_0,t_0)} \leq 0$ (Can you see why? If they do not hold, take a small neighborhood and integrate over it to show that u reaches a larger value somewhere else). We obtain a contradiction $0 \leq \frac{\partial \tilde{u}}{\partial t}_{(p_0,t_0)} < \Delta \tilde{u}_{(p_0,t_0)} \leq 0$.

Now to show the maximum principle for u solving $\frac{\partial u}{\partial t} = \Delta u$ for $0 < t \leq T$, take $\tilde{u}_\varepsilon := u - \varepsilon t$ for each $\varepsilon > 0$. Note that $\frac{\partial \tilde{u}_\varepsilon}{\partial t} < \Delta \tilde{u}_\varepsilon$ hence its maximum occurs only on $t = 0$ or ∂M. Take

$\varepsilon \to 0$, we obtain $\tilde{u}_\varepsilon \to u$ and

$$\max_{M\times[0,T]} u = \max_{\partial M\cup\{t=0\}} u.$$

□

Corollary 8.2 By a similar proof, the minimum occurs also on $t = 0$ or on ∂M.

Exercise 8.2 — Uniqueness. On a compact manifold M without boundary, using the maximum (and minimum) principle, show that the solution to (8.3) is uniquely determined by its initial condition $u_{t=0}$.
Hint *Consider u, v both satisfying the heat equation with common initial condition. What happens to their difference?* ▪

8.1.2 Entropy

An important concept in the thermodynamics is the **entropy**. An entropy is a quantity of a system that only increases (or decreases depending on convention) in time, indicating **time-irreversibility** of the system. A heat diffusion process governed by (8.3) does have (a lot of different definitions of, see Exercise 8.3,) entropy that asserts the time-irreversibility.

The standard entropy introduced in the classical thermodynamics is defined through the following. The entropy density is an n-form $\sigma \in \Omega^n(M; \mathbb{R}\,\mathrm{J/K})$ whose changes in time is given by the quotient between total heat input and the temperature

$$\frac{\partial}{\partial t}\sigma := -\frac{1}{u}dQ.$$

Recall that $Q \in \Omega^{n-1}(M; \mathbb{R}\,\mathrm{J/s})$ is the heat flux (energy per area per time), and $-dQ$ is the total in-flowing energy rate. From (8.2) we have

$$\frac{\partial}{\partial t}\sigma = \star\frac{1}{u}\frac{\partial}{\partial t}u = \star\frac{\partial}{\partial t}\left(\log u\right),$$

which suggests that

$$\sigma = \star \log u$$

up to an additive constant.[2] The **total entropy** is given by $S = \int_M \sigma$. Assuming that M has no boundary and there is no heat source, the change of total entropy is nonnegative (**Second Law of thermodynamics**):

$$\begin{aligned}\frac{d}{dt}S = \int_M \frac{\partial}{\partial t}\sigma = \int_M -\frac{1}{u}dQ &\overset{(8.1)}{=} \int_M \frac{1}{u}d\star du \\ = -\int_M d\left(\frac{1}{u}\right)\wedge\star du = \int_M \frac{1}{u^2}du\wedge\star du &= \int_M \star\frac{1}{u^2}|du|^2 \geq 0\end{aligned}$$

and $\frac{d}{dt}S = 0$ if and only if $du \equiv 0$, *i.e.* u is a constant (at thermal equilibrium). Here we have used integration by parts derived from $d\left(\frac{1}{u}\star du\right) = d\left(\frac{1}{u}\right)\wedge\star du + \frac{1}{u}d\star du$ and Stokes' Theorem that removes the boundary term.

[2] To write it more carefully, including the additive constant we write $\sigma = \star(\log u - \log u_0)$, that is $\sigma = \star\log(u/u_0)$. In this way, the quantity in the logarithm is unit-less.

Exercise 8.3 — Second Law of Thermodynamics. Let $f : \mathbb{R}^+ \to \mathbb{R}$ be any smooth strictly convex function, *i.e.* $f'' > 0$. Such functions are also known as *mathematicians' entropy function*, whereas the special case $f = -\log$ is *minus the physical entropy function*. Show that the total entropy

$$S(u) := \int_M \star f(u)$$

is strictly *decreasing* in time when u is non-constant and satisfies (8.3).

Hint *Integration by parts.* ∎

Exercise 8.4 — L^2 function. The space of all square-integrable functions is denoted by $L^2(M) = \left\{v : M \to \mathbb{R} \,\middle|\, \int_M \star v^2 < \infty\right\}$. Show that if u satisfies (8.3) on M without boundary, and if $u \in L^2(M)$ (*i.e.* u is square-integrable) at $t = 0$, then $u \in L^2(M)$ for all later time.

Hint *Exercise 8.3.* ∎

8.1.3 Dirichlet Energy

Here is a whole different way of looking at the heat equation. For each function $u : M \to \mathbb{R}$, the **Dirichlet energy**

$$E_D(u) = \frac{1}{2}\int_M du \wedge \star du = \frac{1}{2}\|du\|_{L^2}^2$$

is a common way to measure the "roughness" of the function u. When $E_D(u) = 0$, we have u being constant (in each connected component of M). In a diffusion process, *e.g.* governed by the heat flow (8.3), we expect the function u becomes smoother in time. This is indeed true measured with the Dirichlet energy: assuming $\partial M = \varnothing$,

$$\frac{d}{dt}E_D(u) = \int_M d\left(\frac{\partial u}{\partial t}\right) \wedge \star du = \int_M \frac{\partial u}{\partial t} \star (-\Delta u) = -\int_M \star \left(\frac{\partial u}{\partial t}\right)^2 \leq 0$$

and E_D is constant if and only if u is stationary.

Gradient Descent of Dirichlet Energy

A more interesting fact between Dirichlet energy and the heat equation is that the heat flow is the **steepest descent** (**gradient flow**) of the Dirichlet energy. Given a function $u : M \to \mathbb{R}$ and consider a variation $\mathring{u}$ with $\mathring{u}|_{\partial M} = 0$. Then the variation of its Dirichlet energy is

$$\mathring{E}_D = \int_M d\mathring{u} \wedge \star du = \int_M - \star \mathring{u}\Delta u = \langle\!\langle -\Delta u, \mathring{u} \rangle\!\rangle.$$

That is, Δu is the negative gradient of the Dirichlet energy (in the space of all L^2 functions), and (8.3) is understood as the gradient flow of the Dirichlet energy.

One can similarly show that the heat equation with heat source (8.4) is the gradient flow of

$$E_D(u) + \int_M u\beta.$$

8.1.4 Weak formulation

For u to satisfy $\frac{\partial}{\partial t}u = \Delta u$, seemingly u requires twice differentiability (due to the Laplacian). Nonetheless, there is an alternative way to "describe" the heat equation with at most one differentiation. Multiply the heat equation by any differentiable function $\varphi\colon M \to \mathbb{R}$ and integrate the result over M:

$$\int_M \star\varphi \frac{\partial}{\partial t}u = \int_M \star\varphi\Delta u = -\int_M d\varphi \wedge \star du.$$

We again used the integration by parts, assuming M has no boundary. If φ are time-independent, we may pull out $\frac{\partial}{\partial t}$ from the integral and obtain

$$\frac{d}{dt}\langle\!\langle \varphi, u\rangle\!\rangle = -\langle\!\langle d\varphi, du\rangle\!\rangle, \quad \varphi \in \Omega^0(M;\mathbb{R}). \tag{8.5}$$

We say u is a **weak solution** to the heat equation if u satisfies the **heat equation in the weak form** (8.5) for all **test functions** $\varphi \in \Omega^0(M;\mathbb{R})$. If u satisfies (8.3), *i.e.* u is a **strong solution**, then u must also be a weak solution (as just derived). If u is a weak solution, and if one could show that u is twice differentiable (**regularity**), then u is also a strong solution (by reversing the integration by parts).

The weak formulation is important not only in PDE analysis but also in numerical approaches (as the base of finite element method). With the knowledge that the strong solution exists and uniquely agrees with the weak solution, we are safe to adopt numerical approaches which solves the weak form, say, using piecewise linear test functions which are only once differentiable.

8.2 Discrete heat equation

Using Discrete Exterior Calculus (DEC) it is rather straightforward to write down a heat equation on a mesh. In fact, since the entire heat equation theory involves only primal 0-forms (temperature u), 1-forms (du), dual $(n-1)$-forms (energy flux) and dual n-forms (energy density), the heat equation requires only points and edges, such as a graph. But here we consider a discrete complex M such as a triangulated surface.

Let the temperature u sit on points (vertices). Hence $u \in C^0(M;\mathbb{R}\,\mathrm{K})$, where C^k denotes the k-cochain. The temperature difference $d_0u \in C^1(M;\mathbb{R}\,\mathrm{K})$ sits on edges, where $d_0 = \partial_1^\mathsf{T}$ is the incidence matrix. Now, consider the diagonal Hodge star $\star_1$ be a diagonal matrix with positive entries. These entries are the **thermal conductivity**, or the **edge weight** in general. Namely,

$$Q = -\star_1 d_0 u \in C^{n-1}(M^\star;\mathbb{R}\,\mathrm{J/s})$$

is the heat flux sitting on the dual $(n-1)$-cells across the primal edges. When considering a homogeneous material with uniform isotropic heat conductivity, we take the diagonal Hodge star

$$\star_1 = \operatorname{diag}\left(\frac{|P_{e^\perp}e^\star|}{|e|}\right),$$

i.e., the cotangent weight in case of a triangulated surface. The total in-flowing flux into a dual n-cell is given by

$$-d_0^\intercal \star_1 d_0 u.$$

Finally, take the primal Hodge star on the 0-cell $\star_0$, which is the mass matrix. In the context of heat equation, $\star_0$ is the **heat capacitance**. With this final piece we arrive at the discrete heat equation

$$\star_0 \frac{\partial}{\partial t} u = -d_0^\intercal \star_1 d_0 u. \tag{8.6}$$

Similar to the smooth setting, we may also consider a heat source $\beta \in C^n(M^\star; \mathbb{R}\,\mathrm{J/s})$ sitting on the dual n-cell as the amount of additional heat poured into the cell per unit time:

$$\star_0 \frac{\partial}{\partial t} u = -d_0^\intercal \star_1 d_0 u + \beta. \tag{8.7}$$

8.2.1 Discrete Dirichlet Energy

Analogous to the smooth theory, the discrete heat equation is the gradient flow to the Dirichlet energy. Here one has to be careful that on $C^0(M)$, which is the vector space u belongs to, uses an inner product respecting $\star_0$. That is, for $f, g \in C^0(M)$, we have $\langle\!\langle f, g \rangle\!\rangle_{C^0(M)} = f^\intercal \star_0 g$. Similarly, for $\xi, \eta \in C^1(M)$, $\langle\!\langle \xi, \eta \rangle\!\rangle_{C^1(M)} = \xi^\intercal \star_1 \eta$. Now, the Dirichlet energy is the L^2-norm of $d_0 u$, which is given by

$$\begin{aligned} E_D &= \frac{1}{2} \langle\!\langle d_0 u, d_0 u \rangle\!\rangle_{C^1(M)} \\ &= \frac{1}{2} u^\intercal d_0^\intercal \star_1 d_0 u. \end{aligned}$$

Its variation is then

$$\mathring{E}_D = \mathring{u}^\intercal d_0^\intercal \star_1 d_0 u = \langle\!\langle \mathring{u}, \star_0^{-1} d_0^\intercal \star_1 d_0 u \rangle\!\rangle_{C^0(M)}.$$

Hence

$$\operatorname{grad}_{C^0(M)} E_D = \star_0^{-1} d_0^\intercal \star_1 d_0 u$$

which justifies that (8.6) is the gradient flow of the discrete Dirichlet energy.

8.2.2 Finite Element Approach

The discrete heat equation (8.6) and (8.7) coincides with the discretization obtained from the piecewise linear (PL) finite element method (FEM) on a triangulated surface. The PL-FEM approximates the solution using a piecewise linear function. To derive a finite element scheme, we begin with the heat equation in the weak form (8.5). To be precise, the weak heat equation (8.5) is a statement for u, φ in the function space

$$W^{1,2}\Omega^0(M; \mathbb{R}) = \left\{ v \in \Omega^0(M; \mathbb{R}) \,\middle|\, \|v\|^2 < \infty \text{ and } \|dv\|^2 < \infty \right\}.$$

(The **Sobolev space** notation $W^{k,p}$ indicates k-times differentiability and $\int |\cdot|^p < \infty$ integrability.) Observe that the space $\mathrm{PL}\Omega^0(M; \mathbb{R})$ of all functions which are continuous and piecewise

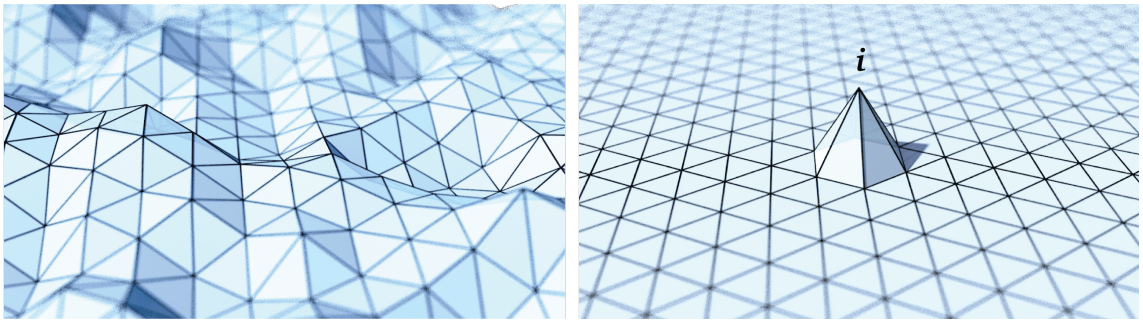


Figure 8.1 Left: a generic piecewise linear function whose linear pieces are the triangle faces of a given mesh. Right: the hat function ϕ_i.

linear over the faces in a triangular mesh is a *finite dimensional* subspace of this infinite dimensional space $W^{1,2}\Omega^0(M;\mathbb{R})$. A canonical basis for $\mathrm{PL}\Omega^0(M;\mathbb{R})$ is the "hat functions" $\{\phi_i\}_i$ which are piecewise linear and $\phi_i(p_j) = \delta_{ij}$ on vertex p_j (Figure 8.1).

Any function $\hat{u} \in \mathrm{PL}\Omega^0(M;\mathbb{R})$ is written as

$$\hat{u} = \sum_i u_i \phi_i$$

where each $u_i \in \mathbb{R}$ is a number sitting on a vertex, *i.e.* $u = (u_i)_i \in C^0(M;\mathbb{R})$ zeroth cochain. Notice that $\hat{u}(p_i) = u_i$. Now, to best approximate the heat equation, we *orthogonally project* the heat equation onto the subspace $\mathrm{PL}\Omega^0(M;\mathbb{R})$:

$$\frac{d}{dt}\langle\!\langle \phi_i, \hat{u} \rangle\!\rangle = -\langle\!\langle d\phi_i, d\hat{u} \rangle\!\rangle, \quad \forall \text{ basis element } \phi_i.$$

Note that this is the orthogonal projection of the heat equation in the sense that the remainder error $r =$ "$\frac{\partial}{\partial t}\hat{u} - \Delta\hat{u}$" is L^2-orthogonal to all basis element ϕ_i.

Now, with $\hat{u} = \sum_j u_j \phi_j$, we have

$$\sum_j \langle\!\langle \phi_i, \phi_j \rangle\!\rangle \frac{d}{dt} u_j = \sum_j -\langle\!\langle d\phi_i, d\phi_j \rangle\!\rangle u_j, \qquad \text{for all } i. \tag{8.8}$$

By working out the L^2 inner product (in the smooth sense) for $\langle\!\langle \phi_i, \phi_j \rangle\!\rangle$ and $\langle\!\langle d\phi_i, d\phi_j \rangle\!\rangle$, define matrices $(\star_0^{\mathrm{FEM}})_{ij} = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle$ (the **mass matrix**) and $L_{ij} = -\langle\!\langle d\phi_i, d\phi_j \rangle\!\rangle$ (the **stiffness matrix**). Thus (8.8) becomes

$$\star_0^{\mathrm{FEM}} \frac{d}{dt} u = Lu. \tag{8.9}$$

If we can compute the matrix elements of L and $\star_0^{\mathrm{FEM}}$, we can solve (8.9) for the coefficients u.

Exercise 8.5 — The stiffness matrix. On a discrete triangular mesh M, the stiffness matrix L is given by

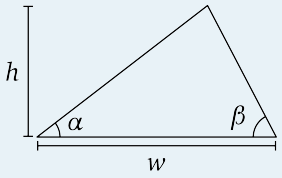
$$L_{ij} = -\int_M d\varphi_i \wedge \star d\varphi_j = -\sum_{t_{ijk} \succ e_{ij}} \langle \mathrm{grad}\, \varphi_i, \mathrm{grad}\, \varphi_j \rangle A_{ijk}$$

where A_{ijk} is the area of triangle t_{ijk}. Here we used the fact that the hat functions φ_i, φ_j

are piecewise linear and hence their gradients are piecewise constant on the triangles. We also used the fact that grad φ_i and grad φ_j have common support (the region on which the function is non-zero) only on the triangles incident to both i and j.

(a) Show that the aspect ratio of a triangle can be expressed as the sum of the cotangents of the interior angles at its base, *i.e.*,

$$\frac{w}{h} = \cot\alpha + \cot\beta.$$



(b) Show that the gradient of the hat function on triangle t_{ijk} is given by

$$\operatorname{grad}\varphi_i = \frac{e_{jk}^{\perp}}{2A_{ijk}}$$

where $e_{jk}^{\perp}$ is the vector e_{jk} rotated 90° counterclockwise within t_{ijk}.

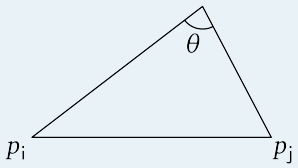
(c) Show that for any hat function φ_i associated with vertex p_i of triangle t_{ijk},

$$\langle \operatorname{grad}\varphi_i, \operatorname{grad}\varphi_i \rangle A_{ijk} = \frac{1}{2}(\cot\alpha + \cot\beta).$$

(d) Show that for the hat functions φ_i and φ_j associated with vertices p_i and p_j of triangle t_{ijk}, we have

$$\langle \operatorname{grad}\varphi_i, \operatorname{grad}\varphi_j \rangle A_{ijk} = -\frac{1}{2}\cot\theta$$

where θ is the angle between the opposite edge vectors.



Putting all these facts together, we have the infamous *cotan formula*

$$(Lu)_i = \frac{1}{2}\sum_{e_{ij} > p_i} (\cot\alpha_{ij} + \cot\beta_{ij})(u_j - u_i).$$

where α_{ij} and β_{ij} are the angles of opposite vertices across from e_{ij} in the two adjacent triangles. ∎

On a discrete surface, it happens that the finite element stiffness matrix and the discrete exterior calculus Laplace matrix agree

$$L = -d_0^\intercal \star_1 d_0.$$

Since the finite element mass matrix $\star_0^{\text{FEM}}$ is not a diagonal matrix, it does not agree with the diagonal discrete Hodge star $\star_0^{\text{diag}}$. One can show that $\star_0^{\text{diag}}$ and $\star_0^{\text{FEM}}$ has the same row sum. The diagonal entries of $\star_0^{\text{diag}}$ is often referred as the **lumped mass**, and it has been shown that replacing $\star_0^{\text{FEM}}$ with $\star_0^{\text{diag}}$ does not affect the order of accuracy.

8.2.3 Time discretization for the heat flow

So far we have seen the heat equation discretized in space, but is still continuous in the time variable:

$$\star_0 \frac{d}{dt} u = Lu, \quad u \in C^0(M;\mathbb{R}), \quad L = -d_0^\intercal \star_1 d_0.$$

One sure can plug this system ODE to any ODE solver. Here we note on a few popular ODE schemes and comment on their stability.

Denote $0 = t_0 < t_1 < t_2 < \dots$ be the discretization of the time $\mathbb{R}^+ = \{0 \le t < \infty\}$. Assume uniform time step $t_n - t_{n-1} = \Delta t$. Let $u^{(n)} \in C^0(M;\mathbb{R})$ be the approximated solution at $t = t_n$.

One of the most straightforward schemes is the **forward Euler scheme**, which directly replaces $\frac{d}{dt}$ by the forward difference:

$$\star_0 \frac{1}{\Delta t}\left(u^{(n+1)} - u^{(n)}\right) = Lu^{(n)};$$

after rearrangement,

$$u^{(n+1)} = u^{(n)} + \Delta t \star_0^{-1} Lu^{(n)}. \tag{8.10}$$

This is an **explicit** iteration: one may update $u^{(n+1)}$ directly from $u^{(n)}$ by operations involving at most matrix multiplication. (When $\star_0$ is not diagonal, you may need to solve a linear system with a symmetric positive definite matrix $\star_0$).

The **backward Euler scheme** is similar but **implicit**:

$$\star_0 \frac{1}{\Delta t}\left(u^{(n+1)} - u^{(n)}\right) = Lu^{(n+1)};$$

after rearrangement,

$$(\star_0 - \Delta t L)\, u^{(n+1)} = \star_0 u^{(n)}. \tag{8.11}$$

To calculate $u^{(n+1)}$ from a current $u^{(n)}$, one has to solve a linear system with a symmetric positive definite matrix $(\star_0 - \Delta t L)$. (Recall $L = -d_0^\intercal \star_1 d_0$, which is symmetric negative semi-definite).

Both the forward and backward Euler time stepping are first-order accurate in time. However, in terms of stability, the forward Euler scheme is unstable unless Δt is extremely small; on the other hand the backward Euler scheme is unconditionally stable (stability is independent of the choice of Δt).

Stability

The most elementary way to analyze stability is by the **von Neumann stability analysis**, where one investigates the stability of the time discretization for each eigenvector of the Laplacian. Let λ_k, v_k be the eigenvalues and eigenvectors of the (generalized) eigenvalue problem

$$-Lv_k = \lambda_k \star_0 v_k$$

for $k = 1, \ldots, N$ when L is N-by-N (*i.e.* there are N vertices). Note that $\lambda_k \geq 0$ since L is negative semi-definite.

Now, plugging a single eigen-mode $u^{(n)} = c_k^{(n)} v_k$ (where $c_k^{(n)}$ is the time-dependent amplitude) into the forward Euler scheme (8.10), we obtain a difference equation

$$c_k^{(n+1)} = (1 - \Delta t \lambda_k) c_k^{(n)}.$$

This first order difference equation (yielding a geometric progression) is stable *only when* $|1 - \Delta t \lambda_k| \leq 1$. This implies that stability requires

$$\Delta t \leq \frac{2}{\lambda_k}$$

for all eigenvalues λ_k. A fact about the eigenvalues $-Lv_k = \lambda_k \star_0 v_k$ is that the maximum eigenvalue λ_N is of order $\frac{1}{\ell^2}$ where ℓ is a typical edge length. That is, in order to simulate the heat equation using the forward Euler method, you have to pick

$$\Delta t \leq O(\ell^2)$$

otherwise the solution will blow up on you.

Using the same technique let us investigate the backward Euler method. Plugging in a single eigen-mode $u^{(n)} = c_k^{(n)} v_k$, (8.11) reads

$$c_k^{(n+1)} = \frac{1}{1 + \Delta t \lambda_k} c_k^{(n)}.$$

Note that for any $\Delta t > 0$ the amplification factor $\frac{1}{1+\Delta t \lambda_k} \leq 1$ and $= 1$ only when $\lambda = 0$. (One needs to check the eigenspace for $\lambda = 0$ is at most 1-dimensional). The unconditional stability follows.

Exercise 8.6 — Unconditional instability. The **leapfrog scheme** for the heat equation is given by

$$\star_0 \frac{1}{2\Delta t} \left(u^{(n+1)} - u^{(n-1)} \right) = L u^{(n)}$$

which allows one to calculate $u^{(n+1)}$ using $u^{(n-1)}$ and $u^{(n)}$ with second order accuracy in time. Show that the leapfrog scheme for the heat equation is however *unstable* for *any* $\Delta t > 0$.
Hint *A linear second order difference equation* $a_2 c^{(n+2)} + a_1 c^{(n+1)} + a_0 c^{(n)} = 0$ *is unstable if there is a root* $z \in \mathbb{C}$ *of the polynomial* $a_2 x^2 + a_1 x + a_0$ *that* $|z| > 1$. ∎

Exercise 8.7 — Crank-Nicolson. The **Crank-Nicolson method** for the heat equation is given by

$$\star_0 \frac{1}{\Delta t}\left(u^{(n+1)} - u^{(n)}\right) = L\frac{u^{(n)} + u^{(n+1)}}{2},$$

which is an implicit method with second order accuracy in time. Show that it is stable for all choice of $\Delta t > 0$ (*i.e.* unconditionally stable). ▪

8.3 Poisson Equation

On a manifold M, given a function $f \in \Omega^0(M;\mathbb{R})$, the **Poisson equation** for $u \in \Omega^0(M;\mathbb{R})$ is given by

$$\Delta u = f. \tag{8.12}$$

You can think of it as the steady solution of the heat equation

$$\frac{\partial}{\partial t}u = \Delta u + \star_0^{-1}\beta$$

with $\frac{\partial}{\partial t}u = 0$ and $\beta = -\star_0 f$. For $\partial M = \varnothing$, the weak formulation of the Poisson equation is that

$$-\langle\!\langle d\varphi, du\rangle\!\rangle = \langle\!\langle \varphi, f\rangle\!\rangle$$

for all test functions $\varphi \in \Omega^0(M;\mathbb{R})$. The solution to (8.12) is also the minimizer to the energy

$$E(u) = \frac{1}{2}\|du\|^2 + \langle\!\langle u, f\rangle\!\rangle.$$

Think of it as the final destination of a heat flow as the gradient flow for this convex functional.

The discrete Poisson equation is written down straightforward by DEC: given $f \in C^0(M;\mathbb{R})$, solve for $u \in C^0(M;\mathbb{R})$ satisfying

$$-d_0^\intercal \star_1 d_0 u = \star_0 f \quad \text{or simply} \quad Lu = \star_0 f.$$

Notice that we write $d_0^\intercal \star_1 d_0$ on the LHS instead of $\star_0^{-1} d_0^\intercal \star_1 d_0$ in order to obtain a linear system with a symmetric system matrix.

8.3.1 Fredholm Alternative

The equation (8.12) should be thought of as a linear algebra problem like $Ax = b$. A natural question to ask here is that, which choices of b admit the existence of the solution x? To answer this type of questions we recall the following basic linear algebra fact.

Theorem 8.3 Let $A\colon V \to W$ be a linear operator from space V to W. Let $A^\star\colon W \to V$ be the adjoint of A, which is the operator that satisfies $\langle Ax, y\rangle_W = \langle x, A^\star y\rangle_V$ for all $x \in V$ and $y \in W$. Then $\ker(A)^\perp = \mathrm{im}(A^\star)$, and $\mathrm{im}(A)^\perp = \ker(A^\star)$.

Back to the Poisson equation. On a manifold M without boundary, we have that Δ is self-adjoint ($\Delta = \Delta^\star$) since $\langle\!\langle \Delta u, v\rangle\!\rangle = -\langle\!\langle du, dv\rangle\!\rangle = \langle\!\langle u, \Delta v\rangle\!\rangle$. With the following exercise, you will show that $\ker(\Delta) = \ker(\Delta^\star)$ consists of functions constant on connected component of M.

Exercise 8.8 — Harmonic functions on M. Suppose M has no boundary. Show that $\Delta u = 0$ if and only if u is constant on each connected component of M.
Hint: Show $\|du\|^2 = 0$. ▪

The Poisson equation (8.12) admits a solution if and only if $f \in \operatorname{im}(\Delta)$. Knowing $\ker(\Delta)$ we conclude the following theorem.

Theorem 8.4 Suppose M is a manifold without boundary, and $M = M_1 \sqcup \ldots \sqcup M_m$ where each M_k is a connected component of M. Then

$$\Delta u = f$$

has a solution if and only if $\int_{M_k} \star f = 0$ for each $k = 1, \ldots, m$.

In most cases the domain M is connected. In this case $\Delta u = f$ has a solution if and only if f has zero mean.

How does this translate into the discrete setting? In the discrete setting one solves the linear system

$$Lu = \star_0 f$$

where $u \in C^0(M; \mathbb{R})$ and the matrix $L = -d_0^\mathsf{T} \star_1 d_0$. Assume that the diagonal Hodge star $\star_1$ is a diagonal matrix with positive diagonal entries. Then the only kernel of L is the subspace spanned by constant function (assuming one connected component). That is, the right-hand side array $\star_0 f$ must have zero sum of entries; in other words, zero sum of f weighted by point area encoded in $\star_0$ analogous to $\int_M \star f = 0$.

The Poisson equation $\Delta u = f$ (and so is the discrete $Lu = \star_0 f$) is not well-posed if f has constant component. In some situation it makes sense to remove the constant component. That is, instead of trying to solve $\Delta u = f$ one can solve $\Delta u = f - \overline{f}$ where $\overline{f} = \frac{1}{\int_M \star 1} \int_M \star f$. Similarly in the discrete version where one attempts to solve $Lu = b$ with the right-hand side array $b = \star_0 f$ being area weighted values of f on vertices, one can instead solve $Lu = b - \overline{b}$ where $\overline{b} = \frac{1}{n} \sum_{i=0}^{n-1} b_i$ and n is the number of vertices. Of course you must be certain that this trick makes sense in the context of your particular problem.

8.4 Classical Maxwell's Equations

The classical Maxwell's equations describe the time evolution of the electric field (1-form) and magnetic field (2-form) interacting with electrically charged matter. In this section we consider the domain being a manifold M representing the space, with time being a separate dimension. In the next section we discuss the relativistic electromagnetism, in which case we consider the domain M being the space and time together as a manifold called a spacetime.

8.4.1 Charge and current

Let M be an n-dimensional manifold. The **electric charge density** is a (coulomb-valued) n-form

$$\rho \in \Omega^n(M; \mathbb{R}\,\mathrm{C})$$

whose sign (compared with the ambient orientation $\star 1$) represent the sign of the charge. Here $\mathbb{R}\,\mathrm{C}$ is the real numbers with unit *coulomb*. On a region $V \subset M$, $\int_V \rho$ represent the total charge in the volume V. The **electric current** j is a (coulomb-per-second-valued) $(n-1)$-form

$$j \in \Omega^{n-1}(M; \mathbb{R}\,\mathrm{C/s})$$

describing the rate electric charge passing through a section per unit area per time. The **conservation of electric charge** states that

$$\frac{\partial}{\partial t}\rho + dj = 0. \tag{8.13}$$

Integrating (8.13) over a volume, using Stokes' Theorem, one has

$$\frac{d}{dt}\int_V \rho = -\int_{\partial V} j$$

which reads that the change of total charge is contributed only from the in-flowing current.

8.4.2 Electric and magnetic field

To introduce the electric field and magnetic field, we consider a single point mass (particle) of charge q located at $p \in M$ and moving with velocity $\mathbf{v} \in T_pM$. A force applied on the particle is written as an *energy*-valued 1-form $f \in \Omega^1_p(M; \mathbb{R}\,\mathrm{J})$. (In particular the unit $[f]$ is Newton and $\int_\gamma f$ along a particle path γ is the total work.) The force due to electromagnetism is given by the **Lorentz force law**

$$f = q\,(E - i_v B)\,.$$

Here $i_{\mathbf{v}}$ is the interior product. By measuring f at every $p \in M$ and $\mathbf{v} \in T_pM$ we determine a 1-form $E \in \Omega^1(M; \mathbb{R}\,\mathrm{J/C})$ called the **electric field**, and a 2-form $B \in \Omega^2(M; \mathbb{R}\,\mathrm{J\cdot s/C})$ called the **magnetic field**.

The Lorentz force law for moving charged particles can be generalized for charged fluids. Let $\rho \in \Omega^n(M; \mathbb{R}\,\mathrm{C})$ be the charge density, which flows according to a velocity vector field $\mathbf{v} \in \Gamma(TM)$. Then the force field $f \in \Omega^1(M; \mathbb{R}\,\mathrm{J})$ on the fluid is given by

$$f = (\star_0^{-1}\rho)\,(E - i_{\mathbf{v}} B)\,.$$

8.4.3 Maxwell's Equations

The (classical) Maxwell's equation is the differential equation for the electric field and magnetic field in the presence of charged density and current and time variation of the fields themselves. The classical Maxwell's equations consist of the four equations: the absence of magnetic charge, Faraday's law, Gauss' law, and Ampere-Maxwell law.

The absence of magnetic charge

This law states that $B \in \Omega^2(M; \mathbb{R}\,\mathrm{J\cdot s/C})$ is **closed**, *i.e.*

$$dB = 0. \tag{8.14}$$

When $M = \mathbb{R}^3$ and write the magnetic field as a vector $\mathbf{B} = (\star B)^\sharp$, then $dB = 0$ amounts to the divergence free condition $\nabla \cdot \mathbf{B} = 0$.

Faraday's law

Micheal Faraday discovered that a time variation in a magnetic field induces a circulation of electric field

$$dE = -\frac{\partial}{\partial t} B. \tag{8.15}$$

The above two equations (8.14) and (8.15) do not require metric or $\star$. They are just topological properties. The remaining equations require Hodge stars, which require either a metric on M, or some general weights as the electric permittivity and magnetic permeability as the Hodge stars.

Gauss' law

Gauss' law states that

$$d \star E = \rho. \tag{8.16}$$

To understand this equation, take the integral of it over a volume V:

$$\int_{\partial V} \star E = \int_V \rho$$

which reads that the total flux of electric field over the boundary surface is related to the total charge enclosed. Here the Hodge star $\star = \star_1$ is regarded as a map from primal 1-forms to dual $(n-1)$-forms

$$\star_1 \colon \Omega^1(M; \mathbb{R}\,\mathrm{J/C}) \to \Omega^{n-1}(M; \mathbb{R}\,\mathrm{C})$$

encoding the physical relation called the **electric permittivity** (often denoted with ϵ). In the vacuum in $\mathbb{R}^3$ without distinguishing primal and dual forms, one has $\star_1 dx = \epsilon_0 dy \wedge dz$ with $\epsilon_0 \approx 8.85 \times 10^{-12}\ \mathrm{C^2/N{\cdot}m^2}$. In a general material, at each point $p \in M$ the permittivity $\star_1$ is a general linear map from $\Omega^1(M; \mathbb{R}\,\mathrm{J/C})$ to $\Omega^{n-1}(M; \mathbb{R}\,\mathrm{C})$. The form $D = \star E$ is also known as **electric displacement field**.

Ampere-Maxwell law

Ampere-Maxwell law states that

$$d \star_2 B = j + \star_1 \frac{\partial}{\partial t} E. \tag{8.17}$$

Here

$$\star_2 \colon \Omega^2(M; \mathbb{R}\,\mathrm{J{\cdot}s/C}) \to \Omega^{n-2}(M; \mathbb{R}\,\mathrm{C/s})$$

is called the **magnetic permeability** often denoted as $\frac{1}{\mu}$. In the vacuum $\mathbb{R}^3$, $\star_2 dx \wedge dy = \frac{1}{\mu_0} dz$ where $\mu_0 \approx 1.26 \times 10^{-6}\ \mathrm{N{\cdot}s^2/C^2}$. The $(n-2)$-form $H = \star_2 B \in \Omega^{n-2}(M; \mathbb{R}\,\mathrm{C/s})$ is also known as the **auxiliary magnetic field**. Note that $\epsilon_0 \mu_0 = \frac{1}{c^2}$ where c is the vacuum speed of light (as would be seen later from the electromagnetic wave section.)

A summary

In summary, the classical Maxwell's equations are given by the following. Given a permittivity $\star_1 \colon \Omega^1(M; \mathbb{R}\,\mathrm{J/C}) \to \Omega^{n-1}(M; \mathbb{R}\,\mathrm{C})$, a permeability $\star_2 \colon \Omega^2(M; \mathbb{R}\,\mathrm{J\cdot s/C}) \to \Omega^{n-2}(M; \mathbb{R}\,\mathrm{C/s})$, and a charge density $\rho(t) \in \Omega^n(M; \mathbb{R}\,\mathrm{C})$ and current $j(t) \in \Omega^{n-1}(M; \mathbb{R}\,\mathrm{C})$, then the electric field $E \in \Omega^1(M; \mathbb{R}\,\mathrm{J/C})$ and the magnetic field $B \in \Omega^1(M; \mathbb{R}\,\mathrm{J\cdot s/C})$ satisfy

$$\begin{cases} d \star_1 E = \rho \\ dB = 0 \\ dE = -\frac{\partial}{\partial t} B \\ d \star_2 B = j + \star_1 \frac{\partial}{\partial t} E. \end{cases}$$

8.4.4 Electromagnetic wave

In the absence of ρ and j, *i.e.* with no free charge and free current (not necessarily vacuum), the electromagnetic fields E, M inducts each other and form electromagnetic waves.

Take a time derivative to Ampere-Maxwell law (8.17), which gives

$$d \star_2 \frac{\partial}{\partial t} B = \star_1 \frac{\partial^2}{\partial t^2} E.$$

Then replace $\frac{\partial}{\partial t} B$ by $-dE$ using Faraday's law (8.15), and obtain

$$-d \star_2 dE = \frac{\partial}{\partial t} j + \star_1 \frac{\partial^2}{\partial t^2} E. \tag{8.18}$$

Recall that the codifferential

$$\delta_k = (-1)^k (\star_{k-1})^{-1} d_{n-k} \star_k$$

and the Laplacian is given by

$$\Delta_k = -d_{k-1}\delta_k - \delta_{k+1} d_k.$$

After rearranging (8.18) we have

$$\frac{\partial^2}{\partial t^2} E = - \star_1^{-1} d \star_2 dE = -\delta dE.$$

From the Gauss' law (8.16) $d \star_1 E = \rho$ with $\rho = 0$, we have that $\delta E = - \star_0^{-1} d \star_1 E = 0$. Therefore $d\delta E = 0$ and $\Delta E = (-d\delta - \delta d)E = -\delta dE$. Therefore

$$\frac{\partial^2}{\partial t^2} E = \Delta E. \tag{8.19}$$

This is the **wave equation** for an 1-form E.

Exercise 8.9 In the absence of ρ and j, show that B satisfies

$$\frac{\partial^2}{\partial t^2} B = \Delta B = d \star_1^{-1} d \star_2 B. \tag{8.20}$$

∎

The **wave speed** of the wave equations (8.19) and (8.20) are encoded by Δ_1 and Δ_2 respectively. Namely, in both cases the wave speed is determined by the product of the factor contributed from $\star_1^{-1}$ and $\star_2$. In the $\mathbb{R}^3$ vacuum $\star_1 = \epsilon_0 \star_1^{\mathbb{R}^3}$ and $\star_2 = \frac{1}{\mu_0}\star_2^{\mathbb{R}^3}$, where $\star^{\mathbb{R}^3}$ is the usual dimension-less Hodge star induced from the standard metric of $\mathbb{R}^3$. So $\Delta_1 = c^2\Delta_1^{\mathbb{R}^3}$ and $\Delta_2 = c^2\Delta_2^{\mathbb{R}^3}$ where $c^2 = \frac{1}{\epsilon_0\mu_0}$. In a more general material one has a more general $\star_1$ and $\star_2$ that steer the propagation of electromagnetic waves.

8.5 Relativistic Maxwell's Equations

In the previous section, the electromagnetism is formulated with time-dependent 1-form E and 2-form B on a manifold M where M is usually 3-dimensional which represents space. In this section, we consider M as a 4-dimensional manifold including the information of space and time, and the electric and magnetic fields are united as 2-form defined on spacetime.

8.5.1 Electromagnetism on $\mathbb{R} \times M$

Before going to the full relativistic version, let us look at a simpler setup. A straightforward way of putting both time and space together as a domain manifold is by taking the Cartesian product

$$\widetilde{M} \coloneqq \mathbb{R} \times M$$

where the $\mathbb{R}$ component represents the time, and the n-dimensional (usually 3-dimensional) manifold M represents the space. Let t be the coordinate of the time component. Then $dt \in \Omega^1(\widetilde{M}; \mathbb{R}\,\mathrm{s})$ is a natural time-valued 1-form. Its corresponding vector field $dt^\sharp$ is the time direction in $\widetilde{M}$.

Note that this setup is only a quick way to build a spacetime $\widetilde{M}$. In the theory of relativity, there is no preferred time direction, so there is no canonical way to decompose $\widetilde{M}$ into $\mathbb{R} \times M$.

On $\widetilde{M}$ we introduce a pseudo-metric so that

$$\begin{aligned} &\langle dt^\sharp, dt^\sharp \rangle = -c^2, \\ &\langle X, X \rangle = \langle X, X \rangle_M, \quad \text{for } X \in T_pM \subset T_{(t,p)}\widetilde{M} \end{aligned}$$

where $\langle \cdot, \cdot \rangle_M$ is the metric from M. The reason for this choice of metric will explained in the next subsection on special relativity. We also have a volume form for $\widetilde{M}$. If μ_M is the volume n-form for M, then the volume $(n+1)$-form μ for $\widetilde{M}$ is given by

$$\mu = dt \wedge \mu_M.$$

Suppose $\star^M$ is the Hodge star for M, then the Hodge star $\star$ for $\widetilde{M}$ is given by the following. If ω is a k-form *tangent to* M, *i.e.* $\omega(dt^\sharp) = 0$, then

$$\star\omega = (-1)^k dt \wedge \star_M \omega, \quad \star(dt \wedge \omega) = -c^2 \star_M \omega.$$

One should check $\alpha \wedge \star\alpha = \langle \alpha, \alpha \rangle \mu$.

Now, electric charge density $\rho \in \Omega^n(M; \mathbb{R}\,\mathrm{C})$ and flux $j \in \Omega^{n-1}(M; \mathbb{R}\,\mathrm{C/s})$ can be unified as the **4-current** (4 is the case when $n = 3$ and $\widetilde{M}$ is 4-dimensional)

$$J \coloneqq \rho - dt \wedge j \in \Omega^n(\widetilde{M}; \mathbb{R}\,\mathrm{C}).$$

The conservation of charge $\frac{\partial}{\partial t}\rho + d_M j = 0$ is then equivalent to

$$dJ = 0.$$

One checks that

$$dJ = dt \wedge \frac{\partial}{\partial t}\rho + dt \wedge d_M j = 0$$

where d_M is the exterior derivative on M.

The electric field and magnetic field on M can be unified as a 2-form called the **Faraday 2-form** defined on $\widetilde{M}$:

$$F := -dt \wedge E + B \in \Omega^2(M; \mathbb{R}^{\mathrm{J\cdot s/C}}).$$

Suppose at time t_0 there is a moving charged particle at $p \in M$ with velocity $v_M \in T_p M$. Then the particle is located at $(t_0, p) \in \widetilde{M}$ with velocity $v = dt^\sharp + v_M \in T_{(t_0,p)}\widetilde{M}$. Suppose the particle has electric charge q. Then its Lorentz force is given by

$$f = -q i_v F.$$

Finally, the Maxwell's equations in terms of the Faraday 2-form is concisely written as

$$\begin{cases} dF = 0 \\ d \star F = J. \end{cases} \tag{8.21}$$

The first equation $dF = 0$ in expansion is

$$dt \wedge d_M E + dt \wedge \frac{\partial}{\partial t} B + d_M B = 0$$

which unifies $d_M B = 0$ (by taking another $dt\wedge$ to the equation) and the Faraday's Law $d_M E + \frac{\partial}{\partial t} B = 0$ (by applying $i_{dt^\sharp}$ to the equation). The second equation $J = d \star F$ is

$$\begin{aligned} \rho - dt \wedge j &= d(-c^2 \star_M E + dt \wedge \star_M B) \\ &= -c^2 d_M \star_M E - c^2 dt \wedge \star_M \frac{\partial}{\partial t} E - dt \wedge d_M \star_M B \end{aligned}$$

unifying the Gauss' Law and Ampere-Maxwell's Law.

One can almost immediately write down the corresponding DEC formulation for the discrete version of (8.21). The corresponding algorithm for the case when the mesh being regular was also known as the **Finite Difference Time Domain (FDTD)** method, or **Yee's Scheme**. DEC approach generalizes Yee's Scheme to arbitrary mesh for $\widetilde{M}$, even with non-synchronous time grid (EM-fields are solved at different time resolution in different cells).

8.5.2 Special Relativity

In order to understand the spacetime domain on which we are going to define relativistic electromagnetic fields independent of the choice of time-axis, we need a bit background on Relativity. In particular Special Relativity provides a model for space and time.

Albert Einstein postulated in one of his 1905 papers that

1. the laws of physics are the same in all **inertial frames of reference**.
2. the speed of light in free space has the same value c in all inertial frames of reference.

The first postulate says that if a physical law is formulated in terms of a coordinate system (t, x, y, z), then it must take the same form when written in terms of another coordinate system $(\tilde{t}, \tilde{x}, \tilde{y}, \tilde{z})$ which is a uniform translating motion with respect to (t, x, y, z). This postulate has been known since Galileo Galilei and is applicable to all classical physics. The second postulate is more counterintuitive in the classical sense, and it is just based on the surprising result of 1887 Michelson-Morley experiment: the measurement of the speed of light is the same in all seasons (when earth is moving in different directions!)

To have the two postulates consistently fit together, one puts space and time together as a geometric entity $\mathcal{M}$ called **spacetime**, a 4-dimensional (or in general $(n+1)$-dimensional) manifold. Each point p in spacetime is called an **event**. A curve $\gamma : I \to \mathcal{M}$ in $\mathcal{M}$ is a smoothly varying points on $\mathcal{M}$, with the direction of $\gamma' \in T_\gamma \mathcal{M}$ encodes how "fast" it moves through spacetime.

For example, if $T_p\mathcal{M} = \mathbb{R}^4$ with an orthonormal (in the Euclidean sense) basis $\{e_0, e_1, e_2, e_3\}$ and suppose e_0 is the time direction for some observer, then a vector $v = v_0 e_0 + \cdots + v_3 e_3 \in \mathbb{R}^4$ has a classical velocity $\left(\frac{v_1}{v_0}, \frac{v_2}{v_0}, \frac{v_3}{v_0}\right)$.

In this example, the coefficients v_0 measures the time displacement and $v_{1,2,3}$ measures the spacial displacement of a vector v (a geometric object). The values of the coefficients depends on the basis $\{e_0, e_1, e_2, e_3\}$ (how fast the observer moves and how the observer is oriented, *i.e.* the inertial frames of reference). Although it is more intuitive (in the classical sense) to formulate laws of physics in terms of the values $v_{0,1,2,3}$, the first postulate of Einstein means that the laws of physics should depend only on the geometric object v rather than the coefficients $v_{0,1,2,3}$. Hence if we write everything "geometrically" on spacetime, we arrive at formulation that is relativistic.

Now, the second postulate of Einstein says that the measurement of speed of light is also geometric, *i.e.* independent of coordinate. How to come up with a framework so that it is true?

Before revealing the sequence of definitions, let us look at the example of $\mathbb{R}^4$. If a vector $v \in \mathbb{R}^4$ is tangent to a light trajectory, we must have $c^2 = \left(\frac{v_1}{v_0}\right)^2 + \left(\frac{v_2}{v_0}\right)^2 + \left(\frac{v_3}{v_0}\right)^2$. In other words,

$$-c^2 (v_0)^2 + (v_1)^2 + (v_2)^2 + (v_3)^2 = 0.$$

If we define for each $u, v \in T_p\mathcal{M}$ the "inner product"

$$\langle u, v \rangle_L := -c^2 u_0 v_0 + u_1 v_1 + u_2 v_2 + u_3 v_3$$

then we may use $\langle v, v \rangle_L$ as a measuring device for testing whether a tangent direction v is a light direction. We call v **light-like** if $\langle v, v \rangle_L = 0$. The collection of all light-like vectors in $T_p\mathcal{M}$, as the solutions of the quadratic polynomial $\langle v, v \rangle_L = 0$, forms a cone called **light cone**. We call v **time-like** if $\langle v, v \rangle_L < 0$ (slower than speed of light, *i.e.* v lies in the solid interior of the light cone); and we call v **space-like** if $\langle v, v \rangle_L > 0$. See Figure 8.2.

It turns out $\langle ., . \rangle_L$ can be thought of as a "metric" for the spacetime manifold $\mathcal{M}$. Using such a metric we may talk about geometries of embedded curves (which are particle path/history) analogous to space curves (curves embedded in $\mathbb{R}^3$); in terms of geometry of curves in spacetime the motion law will be made relativistic. With the notion of metric, we may also talk about the Hodge star $\star$ for differential forms on $\mathcal{M}$ which will bring us relativistic Maxwell's equations.

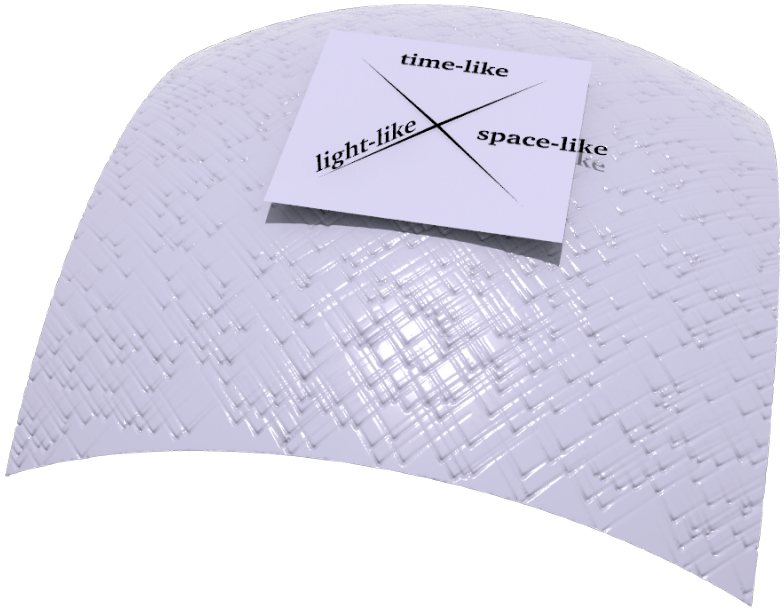


Figure 8.2 A spacetime manifold of dimension 1 + 1. Each tangent space has a distinct **light cone** (collection of light-like vectors) that separates time-like vectors and space-like vectors. The texture of the surface shows a few trajectories of light in this spacetime.

Definition 8.1 — Lorentzian manifold. Let $\mathcal{M}$ be an $(n+1)$-dimensional smooth manifold. It is **Lorentzian** if it is equipped with $\langle .,. \rangle_L$ defined (smoothly) on every tangent space

$$\langle .,. \rangle_L\big|_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$$

such that

- $\langle .,. \rangle_L$ is bilinear and symmetric ($\langle u, v \rangle_L = \langle v, u \rangle_L$).
- $\langle .,. \rangle_L$ is non-degenerate; *i.e.* there is no $u \in T_p\mathcal{M}$ such that $\langle u, v \rangle_L = 0$ for all $v \in T_p\mathcal{M}$.
- $\langle .,. \rangle_L$ has **signature** $(1, n)$. That is for any basis $\{e_0, \ldots, e_n\}$ for $T_p\mathcal{M}$, the matrix $[\langle e_i, e_j \rangle_L]_{ij}$ has exactly one negative eigenvalue and n positive eigenvalues.

$\langle .,. \rangle_L$ is called a **Lorentzian metric** or **Lorentzian inner product**.

Exercise 8.10 Check that the signature of the above definition is independent of the choice of basis. ∎

Definition 8.2 Let $\mathcal{M}$ be an $(n+1)$-dimensional Lorentzian manifold. Then $u \in T_p\mathcal{M}$ is said to be

- **time-like** if $\langle u, u \rangle_L < 0$;
- **light-like** if $\langle u, u \rangle_L = 0$;
- **space-like** if $\langle u, u \rangle_L > 0$.

A vector $u \in T_p\mathcal{M}$ is a **unit vector** if $\langle u, u \rangle_L = -c^2$ or $\langle u, u \rangle_L = 1$. A curve $\gamma : I \to \mathcal{M}$ is said to be a **world line** if γ' is time-like (a path of moving object never exceeding speed

of light). A world line $\gamma\colon \tau \mapsto \gamma(\tau)$ is **parameterized by arclength** if $\langle \gamma', \gamma' \rangle_L = -c^2$, in which case the parameter τ is called the **proper time**.

Similar to the Curve Chapter, we shall assume all world lines are parameterized by arclength (parameterized by proper time). An observer is a (arclength-parameterized) world line $\gamma\colon [0, \tau_{\max}] \to M$, and the proper time τ is his or her biological clock (or any other clock he or she brings along). Let $T = \gamma' \in T_\gamma M$, which is the time direction of the observer γ. Let $T^\perp = \{w \in T_\gamma M \mid \langle w, T \rangle_L = 0\}$ be the n-dimensional orthogonal complement. Note that $T^\perp$ consists of space-like vectors (can you see why?) which is the *space* experienced by the observer γ. Define for each $u \in T_\gamma M$

$$u_T = \frac{\langle T, u \rangle_L}{\langle T, T \rangle_L} = -\frac{1}{c^2}\langle T, u \rangle_L \quad \text{(Lorentz factor)}$$
$$u_{T^\perp} = u - u_T T \in T^\perp.$$

Then every $u \in T_\gamma M$ can be orthogonally decomposed into

$$u = u_T T + u_{T^\perp}.$$

(The Lorentz factor is often denoted by "γ" in special relativity. Since the letter "γ" coincides with the notation for curves, and that the Lorentz factor totally depends on both u and the observer's direction T, and that it is the projection factor from u to T, we shall use the notation "u_T".)

Definition 8.3 Let γ be an observer, $T = \gamma'$, and suppose $u \in T_\gamma M$. Then the **classical velocity** of u observed by γ is given by

$$\frac{u_{T^\perp}}{u_T} \in T^\perp,$$

which is the space displacement divided by the time displacement.

Now we can check that Einstein's 2nd postulate is satisfied. Suppose u is a light-like vector ($\langle u, u \rangle_L = 0$). For any observer γ with $T = \gamma'$, one has

$$u = u_T T + u_{T^\perp}.$$

From $\langle u, u \rangle_L = 0$, using $\langle T, u_{T^\perp} \rangle_L = 0$ we get

$$0 = u_T^2 \langle T, T \rangle_L + \langle u_{T^\perp}, u_{T^\perp} \rangle_L = -c^2 u_T^2 + |u_{T^\perp}|^2$$

Hence the classical speed of u is $\frac{|u_{T^\perp}|}{|u_T|} = c$. Therefore, the value of the speed of light measured by any observer is the same constant c.

Note

Classically velocity or speed has the unit of $^{\mathrm{m}}/_{\mathrm{s}}$. The constancy of c allows us to think of the unit s be just another unit of length through the canonical identification

$$c\colon \mathbb{R}\,\mathrm{s} \xrightarrow{\cong} \mathbb{R}\,\mathrm{m} \quad \text{linear},$$
$$c \cdot 1\,\mathrm{s} := 299792458\,\mathrm{m} \quad \text{in our universe}.$$

Many authors would choose the time unit as $\tilde{\mathrm{s}} = \frac{1}{299792458}\,\mathrm{s}$ so that $c = 1$. However we would like to keep c in our formulas to give the right sense of scales. Hence, let us agree on the following convention: "second" is a unit of length with $1\,\mathrm{s} = 299792458\,\mathrm{m}$, and c is the number $c = 299792458 \in \mathbb{R}$. In particular *velocities are dimensionless*.

8.5.3 Momentum and Force

Suppose a world line $\gamma : I \to M$ is a particle trajectory with tangent $u = \gamma'$ (assuming arclength parameterization $\langle u, u\rangle_L = -c^2$). The vector $u \in T_{\gamma(\tau)}M$ is also called the **4-velocity**.

Let $m_0 \in \mathbb{R}\,\mathrm{kg}$, $m_0 > 0$, be the **rest mass** of the particle. Then the **4-momentum** is given by

$$P := m_0 u.$$

(The physical unit of P is $[P] = [\mathrm{kg}]$) And the **force** is given by

$$f := P' = m_0 u'.$$

($[f] = [\mathrm{kg/m}]$) As a consequence of $\langle u, u\rangle_L = -c^2$ being constant, we always have $\langle f, u\rangle_L = 0$. In fact u', the **acceleration**, is exactly the curvature normal for the curve γ. The meaning of u' is clear when $M = \mathbb{R}^4$ analogous to space curves. When M is curved, the curvature of γ is referred as the **geodesic curvature** of the curve on M.

How does momentum and force "looks like" classically? Suppose there is an observer who travels at a direction T, $\langle T, T\rangle_L = -c^2$. Then

$$\begin{aligned} P &= m_0 u_T T + m_0 u_{T^\perp} \\ &= m_T T + m_T \frac{u_{T^\perp}}{u_T} \end{aligned}$$

where $m_T := m_0 u_T$ is the **classical mass** and $m_T \frac{u_{T^\perp}}{u_T} = P_{T^\perp}$ is the **classical momentum** (classical mass times classical velocity). Similarly,

$$f = m_T' T + f_{T^\perp}$$

assuming $T' = 0$. Here $f_{T^\perp} = P'_{T^\perp}$, so the **classical force** is given by $f_c := \frac{f_{T^\perp}}{u_T}$ (change of classical momentum per unit time displacement). Now, using $\langle f, u\rangle_L = 0$, one has

$$\begin{aligned} 0 &= m_T' \langle T, u\rangle_L + u_T \langle f_c, u\rangle_L \\ &= -m_T' c^2 u_T + u_T \langle f_c, u_T T + u_{T^\perp}\rangle_L \\ &= -m_T' c^2 u_T + u_T^2 \left\langle f_c, \frac{u_{T^\perp}}{u_T} \right\rangle. \end{aligned}$$

This implies the **classical work rate** $\left\langle f_c, \frac{u_{T^\perp}}{u_T} \right\rangle$ equals to $\frac{m_T'}{u_T} c^2$ where $\frac{m_T'}{u_T}$ classical rate of mass gain. Since work is the energy added into the particle, it associates an energy E_T to a mass m_T via

$$E_T = m_T c^2.$$

Thus, in the observer's expansion of the momentum $P = m_T T + m_T \frac{u_{T^\perp}}{u_T}$, while the space component $m_T \frac{u_{T^\perp}}{u_T}$ is understood as the classical momentum, the time component $m_T = \frac{E_T}{c^2}$ is understood as the energy.

The value observed energy $E_T = m_T c^2$ can be written in terms of m_0 and the information $u_T, u_{T^\perp}$ of their relative speed. From $\langle P, P\rangle_L = -m_0^2 c^2$, one has

$$-c^2 m_T^2 + \left| m_T \frac{u_{T^\perp}}{u_T} \right|^2 = -m_0^2 c^2$$

and thus obtains

$$E_T^2 = m_0^2 c^4 + c^2 \left| m_T \frac{u_{T^\perp}}{u_T} \right|^2$$

i.e. the observed energy is contributed from the rest mass and the observed kinetic energy.

8.5.4 Maxwell's Equations

On the spacetime manifold M, suppose γ is a world line representing the trajectory of a charged particle with charge q. Let $u = \gamma'$ (again, $\langle u, u \rangle_L = -c^2$). Assume that the particle feels no other forces than the electromagnetic force. Then the force $f = m_0 u'$ is proportional to q and it is orthogonal to u. The latter statement can be rephrased by that $i_u f^\flat = 0$. Thus, $f^\flat$ *must be* the result of i_u applied on some 2-form. (Can you see the image of i_u equals to the kernel of i_u?) Therefore, we conclude

$$f^\flat = -q i_u F$$

where F is a 2-form called the **Faraday 2-form**. ($F \in \Omega^2(M; \mathbb{R}\,\mathrm{kg/C})$). If T is a unit time-like vector as the direction of an observer, then

$$E = \frac{1}{c^2} i_T F, \quad B = J_{T^\perp}^* F$$

where $J_{T^\perp}^*$ is the pull-back for the inclusion $J_{T^\perp} : T^\perp \hookrightarrow T_p M$. That is, B is the projection of F on $T^\perp$.[3] See Example 8.1 in Appendix **??** for more detail. Note that in fact the observer T decomposes F into

$$F = -T^\flat \wedge E + B.$$

[3] It turns out that

$$(v^\flat \wedge) \circ i_v + i_v \circ (v^\flat \wedge) = |v|^2 \,\mathrm{id}\,.$$

If $|v|^2 = 1$, then the operators $i_v \circ (v^\flat \wedge)$ and $(v^\flat \wedge) \circ i_v$ are projection operators. One sees that in the evaluation $(i_v(v^\flat \wedge \omega))(X_1, \ldots, X_k)$, any v component of X_j will not contribute. If $J : v^\perp \hookrightarrow T_p M$ is the inclusion map, then one would have that

$$i_v(v^\flat \wedge \omega)(X_1, \ldots, X_k) = (J^\star \omega)(X_1, \ldots, X_k)$$

for all $X_1, \ldots, X_k \in v^\perp$.

■ **Example 8.1** In Section 8.5.4, if T is a unit time-like vector field ($\langle T, T \rangle_L = -c^2$), and F is the Faraday 2-form. Then

$$\begin{aligned} F &= -\frac{1}{c^2}\Big((T^\flat \wedge) \circ i_T + i_T \circ (T^\flat \wedge)\Big) F \\ &= -T^\flat \wedge \underbrace{\frac{1}{c^2} i_T F}_{E} + \underbrace{J_{T^\perp}^\star F}_{B} \end{aligned}$$

where E and B are the electric and magnetic fields measured by an observer traveling along a world line tangent to T. ■

Now we have F concisely represents both electric field and magnetic field. In terms of F, Maxwell's Equations are given by

$$\begin{cases} dF = 0 \\ d \star F = J \end{cases}$$

where $J \in \Omega^n(M; \mathbb{R}\,\mathrm{C})$ (M is $(n+1)$-dimensional) necessarily satisfying $dJ = 0$ with

$$\rho = \frac{1}{c^2} i_T J, \quad j = J^*_{T^\perp} F$$

being the classical charge density and charge current observed by T.

9. Integral Geometry

In this chapter, we make a slight digression back to extrinsic geometry of surfaces embedded in $\mathbb{R}^3$. With the tools of exterior calculus, we revisit the area and curvature measurements of surfaces. It turns out that the volume enclosed by a closed surface, the total area of a surface, the total mean curvature and Gaussian curvature are all related in a coherent pattern.

9.1 Variational Formulas of Curvatures

Let M be a 2D oriented manifold, and let $\mathbf{f}\colon M \to \mathbb{R}^3$ be its realization as a surface in 3D.

Since $\mathbf{f}$ is a function over M that takes values in $\mathbb{R}^3$, we say $\mathbf{f}$ is an $\mathbb{R}^3$-valued 0-form

$$\mathbf{f} \in \Omega^0(M;\mathbb{R}^3).$$

Let

$$\mathbf{n} \in \Omega^0(M;\mathbb{R}^3)$$

denote the unit normal vector (Gauss map).

9.1.1 Expressions Involving $\mathbb{R}^3$-valued Forms

There is the cross product $\times\colon \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ and the dot product $\langle\cdot,\cdot\rangle\colon \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ for $\mathbb{R}^3$ vectors. Building on top of them we have the wedge products for $\mathbb{R}^3$-valued forms. They are defined the same as the usual wedge products, only that the products between the values of the forms are replaced by the cross/dot products. For example, for $\mathbb{R}^3$-valued 1-forms $\alpha, \beta \in \Omega^1(M;\mathbb{R}^3)$,

$$\alpha \underset{\times}{\wedge} \beta \in \Omega^2(M;\mathbb{R}^3), \quad (\alpha \underset{\times}{\wedge} \beta)(X,Y) \coloneqq \alpha(X) \times \beta(Y) - \alpha(Y) \times \alpha(X) \tag{9.1}$$

$$\langle \alpha \underset{,}{\wedge} \beta \rangle \in \Omega^2(M;\mathbb{R}), \quad \langle \alpha \underset{,}{\wedge} \beta \rangle(X,Y) \coloneqq \langle \alpha(X), \beta(Y) \rangle - \langle \alpha(Y), \alpha(X) \rangle \tag{9.2}$$

for all $X, Y \in TM$.

Area Form

Using these wedge products for $\mathbb{R}^3$-valued forms, we can write the area element $\sigma \in \Omega^2(M;\mathbb{R})$ on the surface as

$$\sigma = \frac{1}{2}\langle \mathbf{n}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle \tag{9.3}$$

since

$$\begin{aligned}\sigma(X,Y) &= \frac{1}{2}\langle \mathbf{n}, (d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f})(X,Y)\rangle &(9.4)\\ &= \frac{1}{2}\big\langle \mathbf{n}, d\mathbf{f}(X), d\mathbf{f}(Y)\big\rangle - \frac{1}{2}\big\langle \mathbf{n}, d\mathbf{f}(Y), d\mathbf{f}(X)\big\rangle &(9.5)\\ &= \det\big(\mathbf{n}, d\mathbf{f}(X), d\mathbf{f}(Y)\big). &(9.6)\end{aligned}$$

Similarly, the area normal 2-form has the expression

$$\mathbf{n}\sigma = \frac{1}{2} d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}. \tag{9.7}$$

Curvature Forms

Using principal curvature coordinates allows us to simplify expression involving $d\mathbf{n}$ (since the result is independent of the coordinate system used, we might as well use the most convenient one). Let X and Y be the tangent vectors along the principal curvature directions, *i.e.*, $d\mathbf{n}(X) = \kappa_1 d\mathbf{f}(X)$ and $d\mathbf{n}(Y) = \kappa_2 d\mathbf{f}(Y)$. Then

$$\begin{aligned}(d\mathbf{n} \underset{\times}{\wedge} d\mathbf{f})(X,Y) &= d\mathbf{n}(X) \times d\mathbf{f}(Y) - d\mathbf{n}(Y) \times d\mathbf{f}(X)\\ &= \kappa_1 d\mathbf{f}(X) \times d\mathbf{f}(Y) - \kappa_2 d\mathbf{f}(Y) \times d\mathbf{f}(X)\\ &= (\kappa_1 + \kappa_2)\, \mathbf{n}\sigma(X,Y) &(9.8)\end{aligned}$$

and

$$\begin{aligned}(d\mathbf{n} \underset{\times}{\wedge} d\mathbf{n})(X,Y) &= d\mathbf{n}(X) \times d\mathbf{n}(Y) - d\mathbf{n}(Y) \times d\mathbf{n}(X)\\ &= \kappa_1\kappa_2 d\mathbf{f}(X) \times d\mathbf{f}(Y) - \kappa_1\kappa_2 d\mathbf{f}(Y) \times d\mathbf{f}(X)\\ 2\kappa_1\kappa_2\, \mathbf{n}\sigma(X,Y). &(9.9)\end{aligned}$$

Therefore, recalling that the mean curvature $H = {}^{(\kappa_1+\kappa_2)}/_2$ and Gaussian curvature $K = \kappa_1\kappa_2$, we have the following expression for the mean curvature normal 2-form and the Gaussian curvature normal 2-form:

$$\frac{1}{2} d\mathbf{n} \underset{\times}{\wedge} d\mathbf{f} = H\mathbf{n}\sigma \tag{9.10}$$

$$\frac{1}{2} d\mathbf{n} \underset{\times}{\wedge} d\mathbf{n} = K\mathbf{n}\sigma. \tag{9.11}$$

Volume

For a closed surface $\partial M = \varnothing$ the enclosed volume (counted with sign) is given by

$$\mathrm{Vol} = \frac{1}{6}\int_M \langle \mathbf{f}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle. \tag{9.12}$$

To see this, recognize that $\mathbf{f}(M)$ enclosing a volume means that $\mathbf{f}$ admits an extension defined on some interior U with $M = \partial U$. Use Stokes Theorem

$$\frac{1}{6}\int_{\partial U}\langle \mathbf{f}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle = \frac{1}{6}\int_{U}\langle d\mathbf{f} \underset{,}{\wedge} d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle = \mathrm{Vol}\,.$$

The last equality uses the fact that $\frac{1}{6}\langle d\mathbf{f} \underset{,}{\wedge} d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle$ is the $\mathbb{R}^3$ volume form.[1]

Exercise 9.1 — Conservation laws of curvature normals. When a surface is not sufficiently smooth, pointwise evaluation of $\mathbf{n}$ or curvatures H, K may not be available. In those cases it is often instructive to look at the integral of the quantity of interest over a small region $R \subset M$. Use Stokes' Theorem to show that they can be written as boundary integrals:

$$\int_R \mathbf{n}\sigma = \frac{1}{2}\oint_{\partial R} \mathbf{f} \times d\mathbf{f}, \tag{9.13}$$

$$\int_R H\mathbf{n}\sigma = \frac{1}{2}\oint_{\partial R} \mathbf{n} \times d\mathbf{f}, \tag{9.14}$$

$$\int_R K\mathbf{n}\sigma = \frac{1}{2}\oint_{\partial R} \mathbf{n} \times d\mathbf{n}. \tag{9.15}$$

Conclude that if M is a closed surface then

$$\int_M \mathbf{n}\sigma = \int_M H\mathbf{n}\sigma = \int_M K\mathbf{n}\sigma = \mathbf{0}. \tag{9.16}$$

▪

9.1.2 Functional Gradients

The volume Vol is a functional that takes the whole function $\mathbf{f}$ and returns a real number. In general, suppose $\mathcal{F}$ is a functional that takes the whole function $\mathbf{f}$ and returns a real number. If we vary $\mathbf{f}$ by $\mathring{\mathbf{f}}$, which leads to the functional value $\mathcal{F}$ to vary by $\mathring{\mathcal{F}}$ in response taking the following expression for some $\mathbb{R}^3$-valued 2-form $\omega \in \Omega^2(M;\mathbb{R}^3)$

$$\mathring{\mathcal{F}} = \int_M \langle \mathring{\mathbf{f}}, \omega\rangle,$$

then we call ω the **functional gradient** of $\mathcal{F}$ with respect to the pairing $\int_M\langle\cdot,\cdot\rangle$.

[1] This can be seen by taking three vectors $X, Y, Z \in TU$

$$\begin{aligned}
&\frac{1}{6}\langle d\mathbf{f} \underset{,}{\wedge} d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle(X,Y,Z)\\
&= \frac{1}{6}\langle d\mathbf{f}(X), d\mathbf{f}(Y)\times d\mathbf{f}(Z)\rangle - \frac{1}{6}\langle d\mathbf{f}(X), d\mathbf{f}(Z)\times d\mathbf{f}(Y)\rangle\\
&+ \frac{1}{6}\langle d\mathbf{f}(Y), d\mathbf{f}(Z)\times d\mathbf{f}(X)\rangle - \frac{1}{6}\langle d\mathbf{f}(Y), d\mathbf{f}(X)\times d\mathbf{f}(Z)\rangle\\
&+ \frac{1}{6}\langle d\mathbf{f}(Z), d\mathbf{f}(X)\times d\mathbf{f}(Y)\rangle - \frac{1}{6}\langle d\mathbf{f}(Z), d\mathbf{f}(Y)\times d\mathbf{f}(X)\rangle\\
&= \det(d\mathbf{f}(X), d\mathbf{f}(Y), d\mathbf{f}(Z)).
\end{aligned}$$

9.1.3 Variation of Volume

Let us start with the variation of volume:

$$\begin{aligned}\mathring{\mathrm{Vol}} &= \left(\frac{1}{6}\int_M \langle \mathbf{f}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle\right)^{\circ} = \frac{1}{6}\int_M \langle \mathbf{f}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle^{\circ} \\ &= \frac{1}{6}\int \langle \mathring{\mathbf{f}}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle + \frac{1}{3}\int \langle \mathbf{f}, d\mathbf{f} \underset{\times}{\wedge} d\mathring{\mathbf{f}}\rangle \\ &= \frac{1}{2}\int_M \langle \mathring{\mathbf{f}}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}\rangle = \int_M \langle \mathring{\mathbf{f}}, \mathbf{n}\sigma\rangle.\end{aligned}$$

The second line is due to the product rule and $d\mathbf{f} \underset{\times}{\wedge} d\mathring{\mathbf{f}} = d\mathring{\mathbf{f}} \underset{\times}{\wedge} d\mathbf{f}$. The third line follows from integration by parts:

$$d\langle \mathbf{f}, d\mathbf{f} \times \mathring{\mathbf{f}}\rangle = \langle d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f}, \mathring{\mathbf{f}}\rangle - \langle \mathbf{f}, d\mathbf{f} \underset{\times}{\wedge} d\mathring{\mathbf{f}}\rangle$$

and subsequent application of Stokes' theorem (with $\partial M = \varnothing$).

Theorem 9.1 The functional gradient of the enclosed volume Vol is the area normal $\mathbf{n}\sigma$:

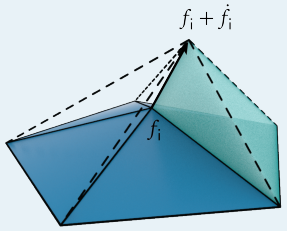
$$\mathring{\mathrm{Vol}} = \int_M \langle \mathring{\mathbf{f}}, \mathbf{n}\sigma\rangle. \tag{9.17}$$

An intuitive interpretation is that a perturbation of the position $\mathbf{f}$ that most rapidly increase the enclosed volume is to move in the normal direction. The resulting perturbation of the volume per unit area is the normal component $\langle \mathring{\mathbf{f}}, \mathbf{n}\rangle$ of the position variation $\mathring{\mathbf{f}}$.

Exercise 9.2 — Discrete volume gradient. Let $\mathbf{f}\colon \{p_0, \ldots, p_{n-1}\} \to \mathbb{R}^3$ be the point (vertex) positions of a closed triangulated surface. Let V be the total polytope volume enclosed by the discrete surface. Let A_{ijk} and $\mathbf{n}_{ijk}$ be the area and the outer normal of the triangle t_{ijk}. Show that under a variation $\dot{\mathbf{f}}\colon \{p_0, \ldots, p_{n-1}\} \to \mathbb{R}^3$, the variation of the total volume is

$$\dot{V} = \sum_{p_i} \sum_{t_{ijk} \succ p_i} \left\langle \left(\frac{1}{3} A_{ijk} \mathbf{n}_{ijk}\right), \dot{\mathbf{f}}_i \right\rangle,$$

where $t_{ijk} \succ p_i$ denotes p_i is incident to t_{ijk}. That is, the point normal derived from the volume gradient is the area-weighted averaging of the neighboring face normals.



Hint *What is the extra volume gained by moving one vertex?* ∎

9.1.4 Variation of Area

From now on we will allow $\partial M \neq \varnothing$.

The total area of the surface is given by

$$\text{Area} = \int_M \sigma = \frac{1}{2}\int_M \langle \mathbf{n}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f} \rangle. \tag{9.18}$$

The variation follows as

$$\begin{aligned}
\mathring{\text{Area}} &= \frac{1}{2}\int_M \langle \mathring{\mathbf{n}}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{f} \rangle + \int_M \langle \mathbf{n}, d\mathbf{f} \underset{\times}{\wedge} d\mathring{\mathbf{f}} \rangle \\
&= \int_M \langle \mathring{\mathbf{f}}, d\mathbf{n} \underset{\times}{\wedge} d\mathbf{f} \rangle - \int_M d\langle \mathbf{n}, d\mathbf{f} \times \mathring{\mathbf{f}} \rangle \\
&= \int_M \langle \mathring{\mathbf{f}}, 2H\mathbf{n}\sigma \rangle - \oint_{\partial M} \langle \mathring{\mathbf{f}}, \mathbf{n} \times d\mathbf{f} \rangle.
\end{aligned}$$

The first equality results again from the product rule. Here the first term vanishes because $\mathring{\mathbf{n}}$ lies in the tangent plane (is linearly dependent with two $d\mathbf{f}$'s), while the second term is transformed through integration by parts $d\langle \mathbf{n}, d\mathbf{f} \times \mathring{\mathbf{f}} \rangle = \langle d\mathbf{n} \underset{\times}{\wedge} d\mathbf{f}, \mathring{\mathbf{f}} \rangle - \langle \mathbf{n}, d\mathbf{f} \underset{\times}{\wedge} d\mathring{\mathbf{f}} \rangle$. The third equality uses (9.10).

Theorem 9.2 The functional gradient of the area is the twice the mean curvature normal $2H\mathbf{n}$ per unit area in the interior, and the clockwise 90° rotation of the boundary curve tangent, $-\mathbf{n} \times {}^{d\mathbf{f}}/_{ds}$, per unit length on the boundary:

$$\mathring{\text{Area}} = \int_M \langle \mathring{\mathbf{f}}, 2H\mathbf{n}\sigma \rangle - \oint_{\partial M} \langle \mathring{\mathbf{f}}, \mathbf{n} \times d\mathbf{f} \rangle. \tag{9.19}$$

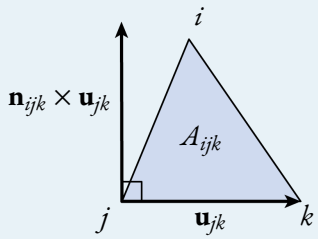
In fact, the negative (functional) gradient of the area is the surface tension per unit area (in the surface) or per unit length (on the boundary). At the boundary, the surface tension is pulling normal to the curve and tangential to the surface. In the surface, the net force is pointing in the normal direction proportional the mean curvature of the surface.

Definition 9.1 — Minimal surface. A surface $\mathbf{f}\colon M \to \mathbb{R}^3$ is called a **minimal surface** if the mean curvature vanishes $H = 0$ everywhere.

Exercise 9.3 — Discrete area gradient. Let $\mathbf{f}\colon \{p_0, \ldots, p_{n-1}\} \to \mathbb{R}^3$ be the point (vertex) positions of a triangulated surface. Let A be the total area of the discrete surface. Denote A_{ijk}, $\mathbf{n}_{ijk}$ the area and normal of the triangle t_{ijk}.

(a) For $p_i \prec t_{ijk}$, let e_{jk} be the edge in t_{ijk} across from p_i. Let $\mathbf{u}_{jk} = \mathbf{f}_k - \mathbf{f}_j \in \mathbb{R}^3$ be the edge vector of e_{jk}. Show that the gradient of the area with respect to $\mathbf{f}_i$ is

$$\frac{\partial A_{ijk}}{\partial \mathbf{f}_i} = \frac{1}{2}\mathbf{n}_{ijk} \times \mathbf{u}_{jk}.$$



Hint *Which direction should you "push" $\mathbf{f}_i$ in order to increase A_{ijk} most efficiently?*

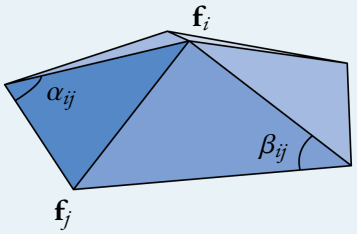
(b) The variation of total area with respect to $\mathbf{f}_i$ is

$$\frac{\partial A}{\partial \mathbf{f}_i} = \sum_{t_{ijk} > p_i} \frac{\partial A_{ijk}}{\partial \mathbf{f}_i}.$$

Show that it can be expressed as

$$\frac{\partial A}{\partial \mathbf{f}_i} = \frac{1}{2} \sum_{e_{ij} < p_i} \left(\cot \alpha_{ij} + \cot \beta_{ij}\right) \left(\mathbf{f}_i - \mathbf{f}_j\right).$$

where α_{ij} and β_{ij} are the angles across from e_{ij} in the two triangles incident to e_{ij}.



Hint *Express the vectors of interest using a coordinate basis containing the edge $(\mathbf{f}_i - \mathbf{f}_j)$.*

(c) Directly from (a), observe that

$$\frac{\partial A}{\partial \mathbf{f}_i} = \tfrac{1}{2} \sum_{t_{ijk} > p_i} \mathbf{n}_{ijk} \times \mathbf{u}_{jk}.$$

Using the result $\int_R H\mathbf{n}\sigma = \frac{1}{2} \oint_{\partial R} \mathbf{n} \times d\mathbf{f}$ from Exercise 9.1, argue that $\frac{\partial A}{\partial \mathbf{f}_i}$ is the integral of mean curvature normal $H\mathbf{n}$ over the one-ring neighborhood of p_i.

■

9.1.5 Variation of Mean Width (Total Mean Curvature)

We have seen that the area normal emerges from the variation of the volume. Extracting from this area normal, we integrate the area form to get total area. The mean curvature emerges from varying this total area. Following the pattern, we integrate the mean curvature, and investigate its variation.

Definition 9.2 — Total mean curvature and mean width. Let the **total mean curva-**

ture be

$$H_{\text{total}} := \int_M H\sigma = \frac{1}{2}\int_M \langle \mathbf{n}, d\mathbf{f} \underset{\times}{\wedge} d\mathbf{n}\rangle. \tag{9.20}$$

Up to a factor it is also known as the **mean width**

$$\text{MeanWidth} := \frac{1}{2\pi} H_{\text{total}}. \tag{9.21}$$

For example, the unit sphere has a mean width of 2 unit length.

The variation of total mean curvature is given by

$$\begin{aligned}
\mathring{H}_{\text{total}} &= \frac{1}{2}\int_M \langle \mathbf{n}, d\mathring{\mathbf{f}} \underset{\times}{\wedge} d\mathbf{n}\rangle + \frac{1}{2}\int_M \langle \mathbf{n}, d\mathbf{f} \underset{\times}{\wedge} d\mathring{\mathbf{n}}\rangle \\
&= \frac{1}{2}\int_M \langle \mathring{\mathbf{f}}, d\mathbf{n} \underset{\times}{\wedge} d\mathbf{n}\rangle - \frac{1}{2}\oint_{\partial M} \Big(\langle \mathbf{n}, d\mathbf{n} \times \mathring{\mathbf{f}}\rangle + \langle \mathbf{n}, d\mathbf{f} \times \mathring{\mathbf{n}}\rangle\Big) \\
&= \int_M \langle \mathring{f}, K\mathbf{n}\sigma\rangle - \frac{1}{2}\oint_{\partial M} \Big(\langle \mathring{f}, \mathbf{n} \times d\mathbf{n}\rangle + \langle \mathring{\mathbf{n}}, \mathbf{n} \times d\mathbf{f}\rangle\Big).
\end{aligned}$$

Theorem 9.3 The functional gradient of the total mean curvature is the Gaussian curvature normal $K\mathbf{n}$ per unit area in the interior of the surface:

$$\mathring{H}_{\text{total}} = \int_M \langle \mathring{f}, K\mathbf{n}\sigma\rangle - \frac{1}{2}\oint_{\partial M} \Big(\langle \mathring{f}, \mathbf{n} \times d\mathbf{n}\rangle + \langle \mathring{\mathbf{n}}, \mathbf{n} \times d\mathbf{f}\rangle\Big). \tag{9.22}$$

Do you have a good intuition of these boundary terms?

9.1.6 Variation of the Total Gaussian Curvature

Finally, for the variation of the total Gaussian curvature

$$K_{\text{total}} := \int_M K\sigma = \frac{1}{2}\int_M \langle \mathbf{n}, d\mathbf{n} \underset{\times}{\wedge} d\mathbf{n}\rangle \tag{9.23}$$

we find

$$\begin{aligned}
\mathring{K}_{\text{total}} &= \int_M \langle \mathbf{n}, d\mathbf{n} \underset{\times}{\wedge} d\mathring{\mathbf{n}}\rangle \\
&= -\oint_{\partial M} \langle \mathring{\mathbf{n}}, \mathbf{n} \times d\mathbf{n}\rangle.
\end{aligned}$$

The first equality follows from the product rule followed by the observation that $\mathring{\mathbf{n}}$ is orthogonal to $d\mathbf{n} \underset{\times}{\wedge} d\mathbf{n}$ and the fact that $d\mathbf{n} \underset{\times}{\wedge} d\mathring{\mathbf{n}} = d\mathring{\mathbf{n}} \underset{\times}{\wedge} d\mathbf{n}$.

In particular this result shows that the total Gaussian curvature is constant for surfaces without boundary, a version of the Gauss–Bonnet theorem. Recall that for a closed surface

$$K_{\text{total}} = \int_M K\sigma = 2\pi\chi(M) \tag{9.24}$$

where $\chi(M)$ is the Euler characteristic of the surface M.

9.2 Quermass

We used the $\mathbb{R}^3$ cross product extensively to derive the above formula for concrete calculation. In general for an $(n-1)$-dimensional manifold embedded in $\mathbb{R}^n$ dimensional space, we need a more general way to compute the variation. Here we use the machinery of Section 6.7 to calculate the variations.

Let us start with the n-dimensional enclosed volume. Let $\mathbf{f}$ be extended to U, an n-dimensional manifold with $\partial U = M$. Then the enclosed volume is given by

$$\mathrm{Vol} = \int_U \mathbf{f}^* \det_{\mathbb{R}^n} . \tag{9.25}$$

Now take variation in $\mathring{\mathbf{f}}$. We know that a tangential component of $\mathring{\mathbf{f}}$ corresponds to a perturbation in reparametrization, which yields no variation in the geometric functional such as Vol. Therefore we may just focus on taking normal variations

$$\mathring{\mathbf{f}} = \mathring{f}\mathbf{n}. \tag{9.26}$$

This gives

$$\begin{aligned}\mathring{\mathrm{Vol}} &= \int_U \mathbf{f}^* \mathscr{L}_{\mathring{\mathbf{f}}} \det_{\mathbb{R}^n} \\ &\overset{\text{Section 6.7.4}}{=} \int_U \mathbf{f}^*(d i_{\mathring{\mathbf{f}}} \det_{\mathbb{R}^n}) = \oint_{\mathbf{f}(M)} i_{\mathring{\mathbf{f}}} \det_{\mathbb{R}^n} = \oint_{\mathbf{f}(M)} \mathring{f}\, i_{\mathbf{n}} \det_{\mathbb{R}^n} .\end{aligned}$$

Therefore the area form

$$\sigma = i_{\mathbf{n}} \det_{\mathbb{R}^n} \in \Omega^{n-1}(M) \tag{9.27}$$

is the functional gradient of the total volume.

Now let us take the variation of total area. We will ignore the boundary term by assuming M is closed. Let us extend the area form σ to a neighborhood of $\mathbf{f}(M)$ by first extending $\mathbf{n}$ in a neighborhood of the surface (extending along the normal) and then define $\sigma = i_{\mathbf{n}} \det_{\mathbb{R}^n}$.

$$\left(\oint_M \mathbf{f}^*\sigma\right)^\circ = \oint_M \mathbf{f}^* \mathscr{L}_{\mathring{\mathbf{f}}} \sigma + \oint_M \mathbf{f}^* \mathring{\sigma} = \oint_{\mathbf{f}(M)} i_{\mathring{\mathbf{f}}} d\sigma + \underbrace{\oint_{\mathbf{f}(M)} d i_{\mathring{\mathbf{f}}} \sigma}_{\substack{=0 \text{ by Stokes,} \\ \text{or } i_{\mathring{\mathbf{f}}} i_{\mathbf{n}} \det = 0}} + \underbrace{\oint_{\mathbf{f}(M)} i_{\mathring{\mathbf{n}}} \det_{\mathbb{R}^n}}_{\substack{=0, \text{ since } \mathring{\mathbf{n}} \text{ is} \\ \text{tangent to } \mathbf{f}(M)}}$$

$$= \oint_{\mathbf{f}(M)} \mathring{f}\, i_{\mathbf{n}} d i_{\mathbf{n}} \det_{\mathbb{R}^n} .$$

Therefore the functional gradient of the total area is the $(n-1)$-form capturing the mean curvature:

$$\sigma_2 \coloneqq i_{\mathbf{n}} d i_{\mathbf{n}} \det_{\mathbb{R}^n} . \tag{9.28}$$

By the same calculation, one finds that the variation of the total mean curvature

$$\left(\oint_M \mathbf{f}^*\sigma_2\right)^\circ = \oint_{\mathbf{f}(M)} \mathring{f}\, \underbrace{i_{\mathbf{n}} d i_{\mathbf{n}} d i_{\mathbf{n}} \det_{\mathbb{R}^n}}_{=:\sigma_3} .$$

Following this pattern we obtain the j-th curvature:

$$\sigma_1 := \sigma = i_{\mathbf{n}} \det_{\mathbb{R}^n} \tag{9.29}$$

$$\sigma_j := \underbrace{i_{\mathbf{n}} d i_{\mathbf{n}} d \cdots i_{\mathbf{n}} d}_{(j-1) \text{ many } i_{\mathbf{n}} d\text{'s}} i_{\mathbf{n}} \det_{\mathbb{R}^n} . \tag{9.30}$$

We call σ_j the **quermass differentials**[2] and

$$Q_j := \int_{\mathbf{f}(M)} \sigma_j \tag{9.31}$$

the j-th **quermassintegral**.

In particular, for $n = 3$,

$$\sigma = \mathbf{f}^*(i_{\mathbf{n}} \det), \quad 2H\sigma = \mathbf{f}^*(i_{\mathbf{n}} d i_{\mathbf{n}} \det), \quad 2K\sigma = \mathbf{f}^*(i_{\mathbf{n}} d i_{\mathbf{n}} d i_{\mathbf{n}} \det). \tag{9.32}$$

9.3 Steiner's Formula

Recall Section 4.7. In the discrete setting, consider a closed polyhedron extruded in the normal direction uniformly. To first order, the rate of change of the enclosed volume will be proportional to the total area; to second order it is proportional to the total mean curvature contributed as small cylinder along each hinged edge; to third order it is proportional to the total Gaussian curvature given by the small cones bloomed from each vertex.

The smooth Steiner's formula also comprises the 4 geometric measurements Vol, Area, H_{total}, and K_{total} of a closed surface in 3D.

Theorem 9.4 — Steiner's formula in 3D. Let $\text{Vol} = \text{Vol}(\mathbf{f})$ denote the enclosed volume of the closed surface $\mathbf{f} : M \to \mathbb{R}^3$. Let $\mathbf{f} + \varepsilon\mathbf{n}$ denote the surface off-set by a uniform distance ε in the normal direction. Then

$$\text{Vol}(\mathbf{f} + \varepsilon\mathbf{n}) = \text{Vol} + \varepsilon \,\text{Area} + \frac{\varepsilon^2}{2} 2H_{\text{total}} + \frac{\varepsilon^3}{3} K_{\text{total}}. \tag{9.33}$$

In general the coefficients of this Taylor polynomial in ε are the quermassintegrals.

Theorem 9.5 — Steiner's formula in nD. For $(n-1)$-dimensional manifolds embedded in $\mathbb{R}^n$,

$$\begin{aligned} \text{Vol}(\mathbf{f} + \varepsilon\mathbf{n}) &= \text{Vol} + \varepsilon \oint_{\mathbf{f}(M)} \sigma_1 + \frac{\varepsilon^2}{2} \oint_{\mathbf{f}(M)} \sigma_2 + \cdots + \frac{\varepsilon^n}{n!} \oint_{\mathbf{f}(M)} \sigma_n \\ &= \text{Vol} + Q_1 \varepsilon + Q_2 \frac{\varepsilon^2}{2} + \cdots + Q_n \frac{\varepsilon^n}{n!} \end{aligned} \tag{9.34}$$

[2] "Quermass" comes from the German "Quermaß" where "quer-" means to measure in cross-section or in projection, and "mass" means mass.

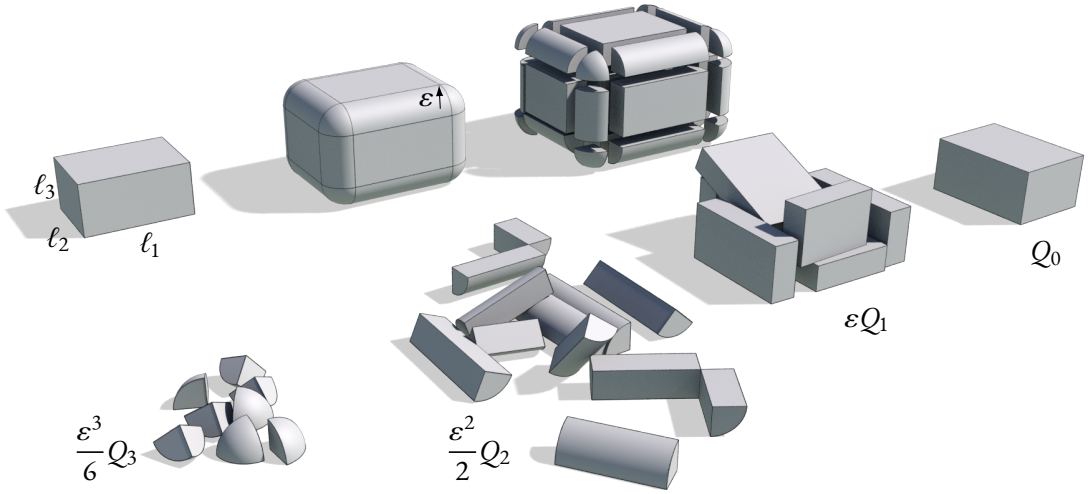


Figure 9.1 The volume of a solid body (here a box) off-set by an ε distance is a polynomial in ε called the **Steiner polynomial**. The quermassintegrals Q_0, Q_1, Q_2, Q_3 are the coefficients (up to the $k!$ factor) of the Steiner polynomial.

where

$$\sigma_j := \underbrace{i_{\mathbf{n}} d i_{\mathbf{n}} d \cdots i_{\mathbf{n}} d}_{(j-1) \text{ many } i_{\mathbf{n}} d\text{'s}} i_{\mathbf{n}} \det_{\mathbb{R}^n}.$$

How is the Steiner's formula related to the sequence of variational curvature normals? Observe that keeping the first order term (9.34) we find that the total area $\oint_{\mathbf{f}(M)} \sigma_1$ is the first variation of volume. Then observe that taking the derivative of (9.34) while keeping nonzero ε yields

$$\text{Area}(\mathbf{f} + \varepsilon \mathbf{n}) = \oint_{\mathbf{f}(M)} \sigma_1 + \varepsilon \oint_{\mathbf{f}(M)} \sigma_2 + \cdots + \frac{\varepsilon^{n-1}}{(n-1)!} \oint_{\mathbf{f}(M)} \sigma_n \tag{9.35}$$

suggesting that σ_2 is indeed the first order perturbation of the total area. Inductively each coefficient is the first-order perturbation term of the previous quermassintegral, which is exactly how we derive these quermassintegrals via computing their functional gradients consecutively.

▪ **Example 9.1 — Quermasses of a box in 3D.** As shown in Figure 9.1, let U be a box in 3D with side lengths ℓ_1, ℓ_2, ℓ_3. Then its volume is

$$Q_0 := \text{Vol} = \ell_1 \ell_2 \ell_3. \tag{9.36}$$

When the box is off-set by ε the extra volume it gains are contributed by the slabs over the faces, the quarter cylinders over the edges and the eighth balls over the corners:

$$\begin{aligned}
\varepsilon Q_1 &= (\varepsilon \ell_1 \ell_2 + \varepsilon \ell_1 \ell_3 + \varepsilon \ell_2 \ell_3) \cdot 2, \\
\frac{\varepsilon^2}{2!} Q_2 &= \left(\pi \varepsilon^2 \cdot \frac{1}{4} \cdot \ell_1 + \pi \varepsilon^2 \cdot \frac{1}{4} \cdot \ell_2 + \pi \varepsilon^2 \cdot \frac{1}{4} \cdot \ell_3 \right) \cdot 4 \\
\frac{\varepsilon^2}{3!} Q_3 &= \frac{4\pi^3}{3} \cdot \frac{1}{8} \cdot 8.
\end{aligned}$$

Therefore,

$$Q_0 = \ell_1\ell_2\ell_3 \tag{9.37}$$
$$Q_1 = 2(\ell_1\ell_2 + \ell_1\ell_3 + \ell_2\ell_3) \tag{9.38}$$
$$Q_2 = 2!\pi(\ell_1 + \ell_2 + \ell_3) \tag{9.39}$$
$$Q_3 = 3!\frac{4\pi}{3}. \tag{9.40}$$

▪

▪ **Example 9.2 — Quermasses of a box in nD.** Let U be a box in $\mathbb{R}^n$ with side lengths $\ell_1, \ldots, \ell_n$. Then

$$Q_0 = \ell_1 \cdots \ell_n, \quad Q_1 = |\mathbb{B}^1| \sum_{1\le i_1\le\cdots\le i_{n-1}\le n} \ell_{i_1}\cdots\ell_{i_{n-1}}, \tag{9.41}$$
$$Q_2 = 2!|\mathbb{B}^2| \sum_{1\le i_1\le\cdots\le i_{n-2}\le n} \ell_{i_1}\cdots\ell_{i_{n-2}}, \tag{9.42}$$
$$Q_k = k!|\mathbb{B}^k| \sum_{1\le i_1\le\cdots\le i_{n-k}\le n} \ell_{i_1}\cdots\ell_{i_{n-k}}, \tag{9.43}$$
$$Q_n = n!|\mathbb{B}^n|, \tag{9.44}$$

where

$$|\mathbb{B}^k| = \frac{\pi^{k/2}}{(k/2-1)!} \tag{9.45}$$

is the volume of the k-dimensional unit ball.[3] ▪

9.4 Hadwiger's Theorem

Earlier we focused on the surface, since those quermassintegrals involve the curvatures of the surface. But we may as well view the quermassintegrals $Q_0 := \mathrm{Vol}, Q_1, \ldots, Q_n$ as measurements of the a *solid body* $U \subset \mathbb{R}^n$ whose boundary is the aforementioned surface.

Each of the quermassintegrals satisfies the following properties:

- Additivity. For any two solid bodies $U, V \subset \mathbb{R}^n$,

$$Q_j(U \cup V) = Q_j(U) + Q_j(V) - Q_j(U \cap V), \quad \text{and } Q(\varnothing) = 0. \tag{9.46}$$

- Invariance under rigid motions. For any translation and rotation $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ in $\mathbb{R}^n$,

$$Q_j(\varphi(U)) = Q_j(U). \tag{9.47}$$

It is easy to see (9.46) for the volume Q_0 and the surface area Q_1. The additivity property for the rest of Q_k can either be seen as the consequence that Q_k comes from the variation of Q_{k-1}, or be shown using Steiner's polynomial formula for polytopes and then take the continuous limit.

[3] The factorial of non-integer is evaluated via the Gamma function. In this case it is enough to note that $(-1/2)! = \sqrt{\pi}$. The factorial of a half integer, say $5/2$, is given by $(5/2)! = 5/2 \cdot 3/2 \cdot 1/2 \cdot (-1/2)! = 5/2 \cdot 3/2 \cdot 1/2\sqrt{\pi}$.

Now we can see any solid body U as the union of many (translated and rotated) small boxes of various sizes like in Example 9.1 and Example 9.2. The two properties (9.46), (9.47) allow us to recover the quermasses Q_j of the whole body knowing only their evaluation on boxes.[4]

Next, we ask, how many kinds of measurements of solid bodies are there that satisfy (9.46) and (9.47) other than the quermasses? This question can be simplified to asking only for boxes. It turns out that $Q_0, \dots, Q_n$ and their linear combinations are the only measurements that satisfy these properties. This is because modulo rotation the only parameter of a box is its side lengths $\ell_1, \dots, \ell_n$. Due to rotational symmetry, the formula for any measurements must be symmetric in $\ell_1, \dots, \ell_n$. Symmetric polynomials in $\ell_1, \dots, \ell_n$ have all be listed by $Q_0, \dots, Q_n$ as shown in Example 9.1 and Example 9.2.

That is, $Q_0, \dots, Q_n$ exhaust all possible additive Euclidean invariant measurements of solid bodies.

Theorem 9.6 — Hadwiger Theorem. Any continuous[a] additive measurements W of solid bodies $U \subset \mathbb{R}^n$ that is invariant under rigid motions can be represented as

$$W(U) = \sum_{j=0}^{n} c_j Q_j(U) \quad \text{for some constants } c_j \in \mathbb{R}. \tag{9.48}$$

If the measurement has a homogeneous scaling rule $W(\lambda \cdot U) = \lambda^k W(U)$, where $\lambda \cdot U$ is U uniformly scaled in size by λ, then

$$W(U) = cQ_{n-k}(U) \quad \text{for some constant } c \in \mathbb{R}. \tag{9.49}$$

[a] The continuity condition here is to assert that these measurements can be taken as the limit of discrete unions of boxes. The technical term is that the measurement is continuous with respect to the Hausdorff metric.

9.5 Cauchy–Crofton Formulae

Crofton formulae relate the quermassintegrals with the expectation value of the number of intersections between the solid body with random points, random rays, random planes, *etc.*

9.5.1 Surface Area

Let us take a planar piece of shape Σ in 3D. Let us sample a random line in the space by the following way. The orientation of the line should be sampled uniformly in the space of all 3D rotations SO(3). Then the translational part of the line should be sampled uniformly and normalized in such a way that the expectation of the line hitting the unit ball is exactly its surface area 4π.[5] Now, the expectation of the random line hitting the planar piece is proportional to its area. Due to rotational invariance of the sampling, when we take the union of many planar pieces that stitch together into a unit sphere, we should obtain the expectation of the total intersection count $2 \cdot 4\pi$ (there twice as many intersections with the surface than just counting

[4] While taking the union of two rotated boxes, how to remove $Q_j(U \cap V)$ for the overlapping regions between rotated boxes might still be unclear here. Later we will see a practical Cauchy–Crofton formula that establishes this additivity rule.

[5] One may take the translation part of the random line be a bounded amount sufficiently large to make the final intersection count invariant under translation within a certain region.

whether the line hit the ball). Therefore,

$$\int_{E_1 \in \text{Lines}} \#\{\Sigma \cap E_1\} = 2\,\text{Area}(\Sigma) \tag{9.50}$$

or more generally for a closed surface $S \subset \mathbb{R}^3$

$$\int_{E_1 \in \text{Lines}} \#\{S \cap E_1\} = 2\,\text{Area}(S). \tag{9.51}$$

Again, the measure on the space of lines is uniform under rigid motions and is normalized so that $\int_{E_1 \in \text{Lines}} \#\{\mathbb{S}^2 \cap E_1\} = 8\pi$. These formulae by counting intersections and integrating all lines are the **Cauchy–Crofton formulae** or **kinematic formulae**.

To make this expression a measurement on a solid body U, we may as well count the number of times the line intersect the solid body. This count is captured by the Euler characteristic of the intersection.

$$\int_{E_1 \in \text{Lines}} \chi(U \cap E_1) = Q_1(U). \tag{9.52}$$

When U is a convex body, such as a ball, the set $U \cap E_1$ is always a single line segment and thus $\chi(U \cap E_1)$ (number of vertices minus the number of edges) is 1 if the line hit the convex body else 0.

Another important property of the Euler characteristics is that it is additive:

$$\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B). \tag{9.53}$$

This allows us to construct the Cauchy–Crofton formulae for measurements on unions of convex body that is automatically additive. These measurements will be invariant under rigid motion by construction (since we take average over all rotations and translations).

9.5.2 General Kinematic Formulae for Quermassintegrals

Let us take a box I in $\mathbb{R}^n$ and intersect it with random k-planes. Then

$$W(I) = \int_{E_k \in \{k\text{-planes}\}} \chi(I \cap E_k) \tag{9.54}$$

is a rigid-motion-invariant measurement on I that has homogeneous scaling law $W(\lambda \cdot I) = \lambda^{n-k} W(I)$. Therefore, $W(I)$ is proportional to $Q_k(I)$. Here we only use Hadwiger's Theorem for boxes, which only invokes exhausting all symmetric polynomials of the side lengths of the box (Example 9.2). Note that since I is convex, $\chi(I \cap E_k)$ is always either 0 (when $I \cap E_k = \varnothing$) or 1 (when $I \cap E_k$ is a k-dimensional topological disk). We use Euler characteristic to make this formula additive.

Using additivity, we immediately obtain that for any non-convex solid body U, the average count $\int_{E_k \in \{k\text{-planes}\}} \chi(U \cap E_k)$ is proportional to $Q_k(U)$. Let us calibrate the measure on $\{k$-planes$\}$ so that for the unit ball $\mathbb{B}^n \subset \mathbb{R}^n$

$$\int_{E_k \in \{k\text{-planes}\}} \chi(\mathbb{B}^n \cap E_k) = Q_k(\mathbb{B}^n). \tag{9.55}$$

Then we have the Cauchy–Crofton kinematic formula for general solid body and general quermassintegrals.

Theorem 9.7 With a normalized measure on the space of k-planes such that (9.55) holds, for each solid body $U \subset \mathbb{R}^n$ we have

$$Q_k(U) = \int_{E_k \in \{k\text{-planes}\}} \chi(U \cap E_k). \tag{9.56}$$

This formula also establishes Hadwiger's Theorem for general solid bodies.

Corollary 9.8 In 3D, the expectation of a uniformly sampled random point lying in a solid body U is proportional to the volume of U.

Corollary 9.9 In 3D, the expectation of the number of segments in which a uniformly sampled random line intersects a solid body U is proportional to the surface area of U.

Corollary 9.10 — Original Cauchy–Crofton formula. In 2D, the expectation of the number of times a uniformly sampled line intersect a curve is proportional to the total length of the curve.

Corollary 9.11 In 3D, the expectation value of the Euler characteristic of a uniformly-sampled planar cross-section of a solid body U is proportional to the total mean curvature of ∂U.

III Hodge Decomposition

10. Hodge Decomposition

In exterior calculus we have seen that the property $d \circ d = 0$ unifies identities such as "curl $\circ$ grad $= 0$", "div $\circ$ curl $= 0$" in vector calculus. In this chapter we further let this property lead us to a useful tool—the **Hodge decomposition**. A special case of the Hodge decomposition is the **Helmholtz decomposition** widely applied in classical physics such as in fluid dynamics and electromagnetism: Any $\mathbb{R}^3$ vector field $\mathbf{v}$ on a simply-connected domain can be decomposed into

$$\mathbf{v} = \operatorname{grad} u + \operatorname{curl} \mathbf{A}, \tag{10.1}$$

where u and $\mathbf{A}$ are known as the scalar and vector potentials for $\mathbf{v}$. Hodge decomposition is the exterior calculus version of the Helmholtz decomposition which applies to differential forms of arbitrary degree on a manifold of arbitrary topology and dimension. At a glance, Hodge decomposition states that any k-form $\omega \in \Omega^k(\mathcal{M})$ can be decomposed into

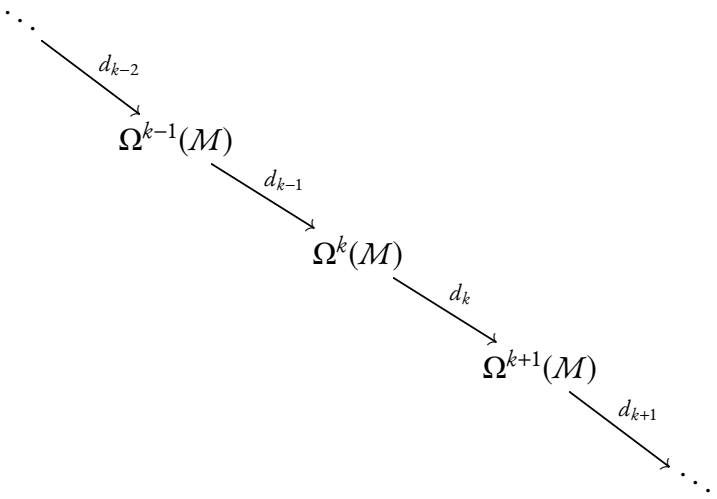
$$\omega = d\alpha + \delta\beta + h \tag{10.2}$$

for some $\alpha \in \Omega^{k-1}(\mathcal{M})$, $\beta \in \Omega^{k+1}(\mathcal{M})$, and a **harmonic form** $h \in \Omega^k(\mathcal{M})$ which satisfies $dh = 0$ and $\delta h = 0$. This harmonic part h shows up when $\mathcal{M}$ has nontrivial topology. For example, on a domain with "holes," there are vector fields which cannot be written as the derives of scalar potentials and vector potentials. The space of harmonic forms on a domain $\mathcal{M}$ is the **de Rham cohomology**. It is in fact directly related to the "number of holes on $\mathcal{M}$" which is algebraically measured by the **homology groups** of $\mathcal{M}$.

We will develop the theory first on closed manifolds, *i.e.*, compact manifolds without boundary, and later extend it to manifolds with boundaries.

10.1 Hodge Decomposition on Closed Manifolds

Suppose M is a closed manifold. Viewing each space of differential forms $\Omega^k(M)$ as a linear space, the exterior derivative d are the following sequence of linear maps:

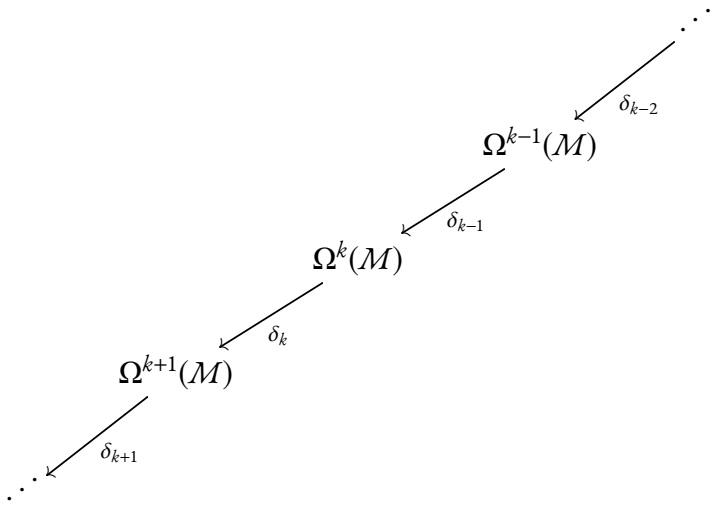


(10.3)

Such a sequence of linear maps from one vector space to the next is called a **chain complex** if $d_k \circ d_{k-1} = 0$, which is the case here. Specifically, the chain complex (10.3) consisting of differential forms and exterior derivatives is called the **de Rham complex**.

We strategically draw arrows with downward motion to emphasize that the differential form is loosing differentiability every time one takes its derivative.

Likewise,



(10.4)

is also a chain complex.

Definition 10.1 — (Co)closed and (co)exact. A k-form ω is said to be

- **closed** if $d\omega = 0$.
- **exact** if $\omega = d\alpha$ for some $(k-1)$-form α.
- **coclosed** if $\delta\omega = 0$.
- **coexact** if $\omega = \delta\beta$ for some $(k+1)$-form β.
- a **harmonic form** if ω is closed and coclosed.

Let

$$\mathcal{H}^k(M) := \ker(d_k) \cap \ker(\delta_k) \subset \Omega^k(M) \tag{10.5}$$

denote the subspace of all k-harmonic forms.

Note that

- $\operatorname{im}(d_{k-1}) \subset \Omega^k(M)$ is the subspace of exact k-forms.
- $\ker(d_k) \subset \Omega^k(M)$ is the subspace of closed k-forms.
- $\operatorname{im}(\delta_{k+1}) \subset \Omega^k(M)$ is the subspace of coexact k-forms.
- $\ker(\delta_k) \subset \Omega^k(M)$ is the subspace of coclosed k-forms.

10.1.1 Orthogonal Decomposition

Let us recall an important fact from linear algebra:

Theorem 10.1 — Four fundamental subspaces. Suppose V and W are linear spaces with inner product $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_W$ respectively. Let $A\colon V \to W$ be a linear map and $A^\star$ be its adjoint. That is, $A^\star$ is defined so that $\langle A\mathbf{v}, \mathbf{w}\rangle_W = \langle \mathbf{v}, A^\star\mathbf{w}\rangle_V$ for all $\mathbf{v} \in V$ and $\mathbf{w} \in W$. Then

$$(\operatorname{im}(A^\star))^\perp = \ker(A), \quad \text{and} \quad (\operatorname{im}(A))^\perp = \ker(A^\star). \tag{10.6}$$

Back to the chain complex. One thing we can infer from the chain complex is that

$$\operatorname{im}(d_{k-1}) \subset \ker(d_k); \tag{10.7}$$

that is, *exact forms are closed*. Also note that d_k and δ_{k+1} are adjoint operators with respect to the L^2 inner product $\langle\!\langle\cdot,\cdot\rangle\!\rangle = \int(\cdot)\wedge\star(\cdot)$:

$$\langle\!\langle d\alpha, \beta\rangle\!\rangle = \langle\!\langle \alpha, \delta\beta\rangle\!\rangle \quad \text{for all } \alpha \in \Omega^k(M) \text{ and } \beta \in \Omega^{k+1}(M). \tag{10.8}$$

Therefore, by Theorem 10.1 the space $\Omega^k(M)$ splits into orthogonal complements:

$$\Omega^k(M) = \ker(d_k) \oplus \operatorname{im}(\delta_{k+1}). \tag{10.9}$$

Since $\operatorname{im}(d_{k-1}) \subset \ker(d_k)$ is a subspace, we can further decompose $\ker(d_k)$ into orthogonal complements within $\ker(d_k)$:

$$\begin{aligned}
\ker(d_k) &= \operatorname{im}(d_{k-1}) \oplus \left((\operatorname{im} d_{k-1})^\perp \cap \ker(d_k)\right) && (10.10)\\
&= \operatorname{im}(d_{k-1}) \oplus (\ker(\delta_k) \cap \ker(d_k)) && (10.11)\\
&= \operatorname{im}(d_{k-1}) \oplus \mathcal{H}^k(M). && (10.12)
\end{aligned}$$

Putting the above together we have

Theorem 10.2 — Hodge decomposition on closed manifolds.

$$\Omega^k(M) = \operatorname{im}(d_{k-1}) \oplus \operatorname{im}(\delta_{k+1}) \oplus \mathcal{H}^k(M). \tag{10.13}$$

That is, for each $\omega \in \Omega^k(M)$, there is a unique $\alpha \in \Omega^{k-1}(M)$ up to a closed form, a unique

$\beta \in \Omega^{k+1}$ up to a coclosed form, and a unique $h \in \mathcal{H}^k(M)$ such that

$$\omega = d\alpha + \delta\beta + h. \tag{10.14}$$

Exercise 10.1 Let $h \in \Omega^k(M)$ be a k-form on a closed manifold M. Show that if $\Delta h = 0$, where $\Delta = -d\delta - \delta d$, then $h \in \mathcal{H}^k(M)$. Give a counterexample when M has a boundary.
Hint *Consider* $\langle\langle \Delta h, h \rangle\rangle$. ▪

10.1.2 Solve the Exact, Coexact and Harmonic Components

Let us first demonstrate how to extract the exact component $d\alpha$ of (10.14).

Take δ on both sides of (10.14) to arrive at

$$\delta d\alpha = \delta\omega. \tag{10.15}$$

The solution α to (10.15) exists due to the postulate (10.14) is true. However, this solution α is not unique: adding α by any closed form ($\in \ker(d_{k-1})$) does not affect the value of $d\alpha$. To pick a *canonical* α in the affine space $\alpha + \ker(d_{k-1})$ we choose the one with the least L^2-norm.

This canonical choice of minimal solution α to (10.15) is called **Coulomb gauge**.

Note that α itself has the Hodge decomposition

$$\alpha = \underbrace{\chi + d\phi}_{\in \ker(d)} + \delta\psi \tag{10.16}$$

where $\chi \in \mathcal{H}^{k-1}(M)$. Minimizing $\|\alpha\|^2 = \|\chi\|^2 + \|d\phi\|^2 + \|\delta\psi\|^2$ (we used the orthogonality of the decomposition) by having the freedom to choose the $\ker(d)$ part, we conclude that α under the Coulomb gauge must have $\chi = 0$ and $d\phi = 0$.

Therefore, to make (10.15) an equation that admits unique solution we pose

$$\begin{cases} \delta d\alpha = \delta\omega \\ \delta\alpha = 0 \\ \langle\langle \alpha, \chi \rangle\rangle = 0 \quad \text{for all } \chi \in \mathcal{H}^{k-1}(M). \end{cases} \tag{10.17}$$

A practical way of solving (10.17) is to rewrite it into a **Poisson problem**:

$$\begin{cases} -\Delta\alpha = \delta\omega \\ \langle\langle \alpha, \chi \rangle\rangle = 0 \quad \text{for all } \chi \in \mathcal{H}^{k-1}(M) \end{cases} \tag{10.18}$$

where $\Delta = -\delta d - d\delta$ is the Laplacian. Now one must show that (10.17) is equivalent to (10.18). It is obvious that the solution to (10.17) satisfies (10.18). Conversely, suppose α is a solution to (10.18). Then

$$\delta d\alpha + d\delta\alpha = \delta\omega.$$

Since the right-hand side is in $\mathrm{im}(\delta)$, by Hodge decomposition we conclude that $d\delta\alpha = 0$. This implies $0 = \langle\langle d\delta\alpha, \alpha \rangle\rangle = \|\delta\alpha\|^2$ and $\delta d\alpha = \delta\omega$. Hence α satisfies (10.17).

Exercise 10.2 Show that β in (10.14) under the Coulomb gauge (minimal $\|\beta\|^2$ solution) is recovered by solving

$$\begin{cases} d\delta\beta = d\omega \\ d\beta = 0 \\ \langle\!\langle \beta, \chi \rangle\!\rangle = 0 \quad \text{for all } \chi \in \mathcal{H}^{k+1}(M). \end{cases} \tag{10.19}$$

Show that this problem is equivalent to the Poisson equation

$$\begin{cases} -\Delta\beta = d\omega \\ \langle\!\langle \beta, \chi \rangle\!\rangle = 0 \quad \text{for all } \chi \in \mathcal{H}^{k+1}(M). \end{cases} \tag{10.20}$$

■

R Suppose $\omega = d\alpha \in \Omega^k(M)$ is given and known to be exact, and we want to find the Coulomb gauge α. Since we know the Coulomb gauge α is always coexact $\alpha = \delta\psi$, instead of solving (10.18), we can solve

$$\begin{cases} -\Delta\psi = \omega \\ \langle\!\langle \psi, \chi \rangle\!\rangle = 0 \quad \text{for all } \chi \in \mathcal{H}^{k}(M) \end{cases} \tag{10.21}$$

and set $\alpha = \delta\psi$. The practical difference between (10.18) and (10.21) is that one Poisson equation is for $(k-1)$-form and the other is for k-form. One of them might have fewer numbers of unknown and is cheaper to compute.

After finding $d\alpha$ and $\delta\beta$ components of $\omega = d\alpha + \delta\beta + h$, simply use

$$h = \omega - d\alpha - \delta\beta \tag{10.22}$$

to extract the harmonic component.

Alternatively, one may first find a basis for $\mathcal{H}^k(M)$ (it is a finite dimensional space, and a method "**tree-cotree algorithm**" for finding its basis will be discussed later in this chapter). One may apply the Gram–Schmidt method to further orthonormalize the basis. Then orthogonally project ω onto $\mathcal{H}^k(M)$ to extract the harmonic part.

Algorithm 10.1 Hodge decomposition on closed manifolds

Input: $\omega \in \Omega^k(M)$ ▷ $\partial M = \varnothing$

1: $\alpha \leftarrow$ Solve (the least-norm solution of) $-\Delta\alpha = \delta\omega$.
2: $\beta \leftarrow$ Solve (the least-norm solution of) $-\Delta\beta = d\omega$.
3: $h = \omega - d\alpha - \delta\beta$.

Output: $d\alpha, \delta\beta, h$.

10.1.3 Problems of Finding Potentials

Hodge decomposition is a powerful tool for asserting existence and uniqueness of certain PDE problems. For example, given a force field $f \in \Omega^1(M; \mathbb{R}\mathrm{J})$ can we find a scalar potential $u \in \Omega^0(M; \mathbb{R}\mathrm{J})$ such that $-du = f$? (Is f a conservative force?) If there exists such u, then how many?

▪ **Example 10.1 — Well-posedness of $d\alpha = \omega$.** Given $\omega \in \Omega^k(M)$ on a closed manifold M,

$$d\alpha = \omega \tag{10.23}$$

has a solution for α if and only if ω satisfies the conditions

$$\begin{cases} d\omega = 0 \\ \langle\!\langle \omega, h \rangle\!\rangle = 0 \quad \text{for all } h \in \mathcal{H}^k(M). \end{cases} \tag{10.24}$$

The solution is unique up to a closed $(k-1)$-form. That is, if α_0 is a solution to $d\alpha = \omega$, then α is also a solution if and only if $\alpha = \alpha_0 + d\phi + \chi$ for $\phi \in \Omega^{k-2}(M)$ and $\chi \in \mathcal{H}^{k-1}(M)$. The Coulomb gauge solution for α is the one that is coexact. ▪

In the case of finding a scalar potential, we are interested in the non-uniqueness of scalar potentials, which is given by $\mathcal{H}^0(M)$ consisting of 0-forms with vanishing d.

▪ **Example 10.2** If M is closed. Then $\mathcal{H}^0(M) = \mathbb{R}^c$, where c is the number of connected components of M. ▪

When studying potential forces we often say "potential energy is unique up to a constant (per connected component)." Now we see that this is a special case of the Hodge and de Rham theory.

10.2 Pseudoinverse of d

Here we introduce a concise expression to encapsulate operations in solving the Hodge decomposition using notions from least-squares problems.

In linear algebra, the **Moore–Penrose pseudoinverse** is a generalization of inverses that describes the least-squares solution to a linear problem.

Definition 10.2 — Moore–Penrose pseudoinverse. Suppose V and W are linear spaces with inner product $\langle \cdot, \cdot \rangle_V$ and $\langle \cdot, \cdot \rangle_W$ respectively. Let $A : V \to W$ be any linear map that is not necessarily invertible. The **pseudoinverse** $A^+ : W \to V$ is defined by the following. Given $\mathbf{w} \in W$, the vector $A^+\mathbf{w} \in V$ is the least-squares solution for $A\mathbf{v} = \mathbf{w}$. That is, $\mathbf{v}$ is found so that it minimizes $|A\mathbf{v} - \mathbf{w}|_W^2$, and among all $\mathbf{v}$ that yields the same minimum value of $|A\mathbf{v} - \mathbf{w}|_W^2$, we take the one with the least $|\mathbf{v}|_V^2$.

In other words, $A^+\mathbf{w}$ first orthogonally projects $\mathbf{w}$ to $\mathrm{im}(A)$, giving $\hat{\mathbf{w}} \coloneqq P_{\mathrm{im}(A)}\mathbf{w}$, and then takes the preimage $A^{-1}\{\hat{\mathbf{w}}\} = \{\mathbf{v} \in V \mid A\mathbf{v} = \hat{\mathbf{w}}\} \subset V$, which is an affine space $\mathbf{v}_0 + \ker(A)$ in V; $A^+\mathbf{w}$ is given by the orthogonal projection of $\mathbf{0}$ to this affine space $A^{-1}\{\hat{\mathbf{w}}\}$.

Consider $d_{k-1} : \Omega^{k-1}(M) \to \Omega^k(M)$. Use the L^2 products $\langle\!\langle \cdot, \cdot \rangle\!\rangle = \int_M (\cdot) \wedge \star(\cdot)$ as the inner product structure for both $\Omega^{k-1}(M)$ and $\Omega^k(M)$. What is $d_{k-1}^+ : \Omega^k(M) \to \Omega^{k-1}(M)$?

Given $\omega \in \Omega^k(M)$, $d_{k-1}^+\omega$ first orthogonally projects ω to $\mathrm{im}(d_{k-1})$, *i.e.* extracts the exact component $d\alpha$ in the Hodge decomposition $\omega = d\alpha + \delta\beta + h$. Then $d_{k-1}^+\omega$ gives the Coulomb gauge α.

In practice, $d_{k-1}^+\omega$ is computed by (10.18), or by (10.21) if ω is already exact.

Note that since d^+ first projects to the exact component, we in particular have $\ker(d^+) = (\mathrm{im}(d))^\perp = \ker(\delta)$. Also note that d^+ always produce a coexact form, we have $\mathrm{im}(d^+) \subset \mathrm{im}(\delta)$. On the other hand, any $\alpha \in \mathrm{im}(\delta)$ can be written as $\alpha = d^+(d\alpha)$. Therefore $\mathrm{im}(\delta) \subset \mathrm{im}(d^+)$. That is, $\mathrm{im}(d^+) = \mathrm{im}(\delta)$.

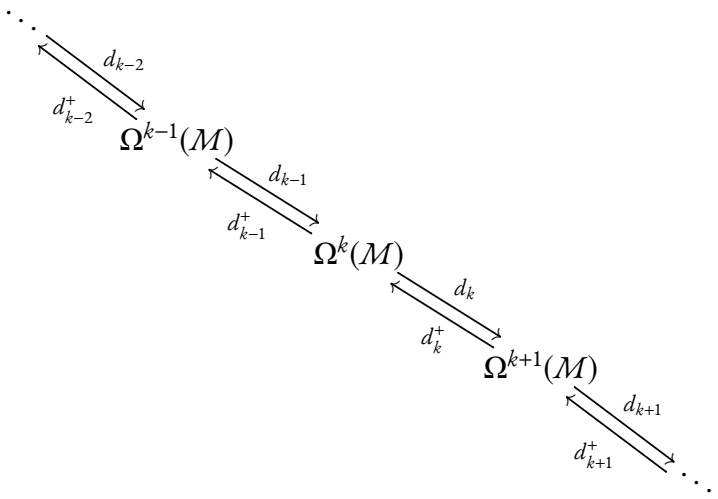
Theorem 10.3 The pseudoinverse d^+ of d forms a chain complex that gives the same notion of coexactness and coclosedness as δ:

$$d^+ \circ d^+ = 0, \tag{10.25}$$

$$\operatorname{im}(d^+) = \operatorname{im}(\delta), \quad \ker(d^+) = \ker(\delta). \tag{10.26}$$

Therefore, we can replace the chain complex of codifferentials (10.4) by the chain complex of d^+:

(10.27)



Note that while d or δ takes differentiation, d^+ effectively applies an integration. The differentiability of a form ascends when applied by d^+.

Theorem 10.4 — Hodge decomposition in terms of d^+. The Hodge decomposition components can be expressed as follows.

- (dd^+) extracts the exact component.
- (d^+d) extracts the coexact component.
- $(1 - dd^+)$ extracts the coclosed component.
- $(1 - d^+d)$ extracts the closed component.
- $(1 - dd^+ - d^+d)$ extracts the harmonic component.

▪ **Example 10.3 — Pressure projection.** In incompressible fluid computations, a common routine is given a vector field $\widetilde{\mathbf{v}} \in \Gamma(TM)$ find a pressure $p\colon M \to \mathbb{R}$ so that $\mathbf{v} = \widetilde{\mathbf{v}} - \operatorname{grad} p$ is divergence-free. In terms of exterior calculus $\widetilde{\eta} = \widetilde{\mathbf{v}}^\flat$, $\eta = \mathbf{v}^\flat$, the problem is that given a 1-form $\widetilde{\eta} \in \Omega^1(M)$, find $\eta = \widetilde{\eta} - dp$ for some $p \in \Omega^0(M)$ so that η is coclosed. This is achieved by $\eta = (1 - dd^+)\widetilde{\eta}$. That is, $\eta = \widetilde{\eta} - dp$ where p is the solution to the Poisson problem $-\Delta p = \delta\eta$.
▪

▪ **Example 10.4 — Magnetic field generated by an electric current.** In $\mathbb{R}^3$, the magnetic field generated by an electric current is given by the Biot–Savart formula (Eq. (3.34)). The formula however does not work on a general domain. In that case we solve the following problem. Let $J \in \Omega^{n-1}(M; \mathbb{R}\,\mathrm{C/s})$ be the electric current flux. Then $H = d^+J \in \Omega^{n-2}(M; \mathbb{R}\,\mathrm{C/s})$ and the magnetic field $B = \star_2^{-1}H \in \Omega^2(M; \mathbb{R}\,\mathrm{Js/C})$ where $\star_2$ is the magnetic permeability. Note that the physical law for magnetic 2-form B is that it is not only closed but also exact (*i.e.* $H = \star_2 B$ is coexact). ▪

▪ **Example 10.5 — Parametrization by Poisson reconstruction.** In surface quadrangulation (quad-meshing) a popular method is to first design some guiding field $\alpha, \beta \in \Omega^1(M)$ over a surface M and then find the functions $u = d^+\alpha, v = d^+\beta$ which has d best approximating the given guiding fields. Then $\mathbf{r} = (u, v) \colon M \to \mathbb{R}^2$ is the texture coordinate that pulls back a square grid as a quad mesh. ▪

10.3 Homology and Cohomology

The harmonic forms in the Hodge decomposition are the closed forms that are not exact. In general, in a chain complex, the space of those non-exact closed elements are the **(co)homology space** or **(co)homology group**.

We will be looking at two types of chain complexes. One of them is the chain complex on k-chains (k-dimensional pieces of geometry) $C_k(M) \xrightarrow{\partial} C_{k+1}(M)$ involving the boundary operators ∂. There the homological question concerns "how many closed chain (cycles) are not exact (boundaries of higher dimensional cell)?" This is a geometric/topological question about the domain. The other chain complex, which we have looked at earlier, is the de Rham complex $\Omega^k(M) \xrightarrow{d} \Omega^{k+1}(M)$ involving the differential operator d. The corresponding homological question is to determine the solvability of PDEs (*e.g.* Can we solve the potential for a given closed differential form?).

These two seemingly different type of problems (PDE vs. topology) are in fact isomorphic.

In this section we will once again assume $\partial M = \varnothing$.

10.3.1 De Rham Cohomology

Let us begin with the de Rham chain complex (10.3). We always have $\operatorname{im}(d_{k-1}) \subset \ker(d_k)$. This means that exactness implies closedness. However, there might be closed forms that are not exact. We capture this "gap" between $\operatorname{im}(d_{k-1})$ and $\ker(d_k)$ by the quotient space ${}^{\ker(d_k)}\!/_{\operatorname{im}(d_{k-1})}$.[1]

Definition 10.3 — De Rham cohomology. The **k-th de Rham cohomology** of a manifold M is the quotient space

$$H^k_{\text{deRham}}(M) \coloneqq \frac{\ker(d_k)}{\operatorname{im}(d_{k-1})} \tag{10.28}$$

Likewise we denote

$$H^{\text{deRham}}_k(M) \coloneqq \frac{\ker(\delta_k)}{\operatorname{im}(\delta_{k+1})} = \frac{\ker(d^+_{k-1})}{\operatorname{im}(d^+_k)} \tag{10.29}$$

for the δ- or d^+-homology.

[1] Let $W \subset V$ be a vector subspace. The notion of quotient space ${}^V\!/_W$ is similar to the notion of orthogonal complement ${}^V\!/_W \cong W^\perp$ but it is not prompted by any inner product structure on V. It is defined by the following. First define an equivalence relation $\sim_W$ in V that two vectors $\mathbf{u}, \mathbf{v} \in V$ is said to be equivalent $\mathbf{u} \sim_W \mathbf{v}$ if their difference is parallel to W: $\mathbf{u} - \mathbf{v} \in W$. The quotient space ${}^V\!/_W$ is the equivalence class defined by this equivalence relation. Concretely, each element $[\mathbf{u}]_W \in {}^V\!/_W$ is an affine subspace $[\mathbf{u}]_W = \{\mathbf{u}\} + W \subset V$ parallel to W. If there is an inner product on V, then we can enumerate this space of affine planes parallel to W by the orthogonal complement $W^\perp$. Therefore ${}^V\!/_W \cong W^\perp$.

Recall that the harmonic part $\mathcal{H}^k(M)$ in the Hodge decomposition is the orthogonal complement of $\mathrm{im}(d_{k-1})$ within $\ker(d_k)$. This means that $\mathcal{H}^k(M) \cong {}^{\ker(d_k)}/_{\mathrm{im}(d_{k-1})}$

Theorem 10.5 — Hodge isomorphism. On a closed manifold

$$\mathcal{H}^k(M) \cong H^k_{\mathrm{deRham}}(M) \cong H_k^{\mathrm{deRham}}(M). \tag{10.30}$$

10.3.2 Homology of a Discrete Mesh

Now, we have another set of chain complexes in the discrete setting. Recall that on a cell decomposition of M we have the chain complex

$$\cdots \xrightarrow{\partial_{k+2}} C_{k+1}(M) \xrightarrow{\partial_{k+1}} C_k(M) \xrightarrow{\partial_k} C_{k-1}(M) \xrightarrow{\partial_{k-1}} \cdots \tag{10.31}$$

and the cochain complex (which is the same as the discrete de Rham complex by discrete exterior calculus)

$$\cdots \xrightarrow[=d_{k-2}]{\partial^\intercal_{k-1}} C^{k-1}(M) \xrightarrow[=d_{k-1}]{\partial^\intercal_k} C^k(M) \xrightarrow[=d_k]{\partial^\intercal_{k+1}} C^{k+1}(M) \xrightarrow[=d_{k+1}]{\partial^\intercal_{k+2}} \cdots . \tag{10.32}$$

Note that $\partial_k \circ \partial_{k+1} = 0$.

Definition 10.4 — Discrete homology. A k-chain $\Gamma \in C_k(M)$ is

- **closed** if $\partial\Gamma = 0$, *i.e.* it has no boundary.
- **exact** if $\Gamma = \partial\Sigma$ for some $\Sigma \in C_{k+1}(M)$.

The *k*-**th homology** is defined as the quotient space

$$H_k(M) \coloneqq \frac{\ker(\partial_k)}{\mathrm{im}(\partial_{k+1})}. \tag{10.33}$$

An element of $H_1(M)$ is represented by 1-chains that are closed but not the boundaries of any 2-chain on a surface M. They exist on surfaces with genus such as a torus. Two such cycles Γ_1, Γ_2 represent the same element in $H_k(M)$ if Γ_1 can be "deformed to" (or more accurately "coborder with") Γ_2 by adding a boundary of a 2-chain, *i.e.* $\Gamma_1 - \Gamma_2 = \partial\Sigma$ for some 2-chain Σ.

Intuitively, for a oriented connected closed surface M has a finite dimensional $H_1(M)$. Each genus (handle) of the surface gives 2 dimensions in the 1st homology.

In general $H_k(M)$ is a finite dimensional space depending only on the global topology of M and is independent of the resolution of the discretization.

Definition 10.5 — Betti number. The k-th **Betti number** is defined by

$$\beta_k \coloneqq \dim\left(H_k(M)\right). \tag{10.34}$$

That is, $H_k(M) \cong \mathbb{R}^{\beta_k}$.

Similarly we have the discrete de Rham cohomology

$$H^k(M) \coloneqq \frac{\ker(\partial^\intercal_{k+1})}{\mathrm{im}(\partial^\intercal_k)} \tag{10.35}$$

as well as the homologies on the dual mesh

$$H_k(M^\star), \quad H^k(M^\star). \tag{10.36}$$

Fortunately, they are all isomorphic.

$$\begin{aligned}\mathbb{R}^{\beta_k} &\cong H_k(M) \cong H^k(M) \cong H_k(M^\star) \cong H^k(M^\star)\\ &\cong H^k_{\text{deRham}}(M) \cong H^{\text{deRham}}_k(M) \cong \mathcal{H}^k(M)\\ &\cong H_{n-k}(M) \cong H^{n-k}(M) \cong H_{n-k}(M^\star) \cong H^{n-k}(M^\star)\\ &\cong H^{n-k}_{\text{deRham}}(M) \cong H^{\text{deRham}}_{n-k}(M) \cong \mathcal{H}^{n-k}(M) \cong \mathbb{R}^{\beta_{n-k}}\end{aligned} \tag{10.37}$$

10.3.3 Isomorphisms between Homologies

Theorem 10.6 — Universal coefficient theorem.

$$H^k(M) \cong H_k(M) \cong \mathbb{R}^{\beta_k}. \tag{10.38}$$

Proof. This can be directly shown by exploiting the 4 fundamental subspaces of matrices (Theorem 10.1). First of all $H_k(M) = {}^{\ker(\partial_k)}/_{\text{im}(\partial_{k+1})}$, so $\ker(\partial_k) \cong \text{im}(\partial_{k+1}) \oplus H_k(M)$. We also have $C_k(M) \cong \ker(\partial_k) \oplus \text{im}(\partial_k^\intercal)$. Thus,

$$C_k(M) \cong \text{im}(\partial_{k+1}) \oplus \text{im}(\partial_k^\intercal) \oplus H_k(M).$$

On the other hand we have $C_k(M) \cong \text{im}(\partial_{k+1}) \oplus \ker(\partial_{k+1}^\intercal)$. Hence,

$$\ker \partial_{k+1}^\intercal \cong \text{im}(\partial_k^\intercal) \oplus H_k(M).$$

Therefore,

$$H^k(M) = \frac{\ker(\partial_{k+1}^\intercal)}{\text{im}(\partial_k^\intercal)} \cong H_k(M).$$

□

Theorem 10.7 — De Rham Theorem.

$$H^k(M) \cong H^k_{\text{deRham}}(M). \tag{10.39}$$

Proof. The isomorphism is induced by the sampling map, that takes a continuous form to a discrete form $\int : \Omega^k(M) \to C^k(M)$. That is it takes a k-form ω and returns a cochain $\int_{(\cdot)} \omega$. By Stokes theorem we have $\int \circ d = \partial^\intercal \circ \int$. That is the following diagram commute

$$\begin{array}{ccc}
\vdots & & \vdots \\
\downarrow d & & \downarrow \partial^\intercal \\
\Omega^k(M) & \xrightarrow{\int} & C^k(M) \\
\downarrow d & & \downarrow \partial^\intercal \\
\Omega^{k+1}(M) & \xrightarrow{\int} & C^{k+1}(M) \\
\downarrow d & & \downarrow \partial^\intercal \\
\vdots & & \vdots
\end{array} \tag{10.40}$$

This induces a map on the cohomology space $\int : H^k_{\text{deRham}}(M) \to H^k(M)$. Moreover $\int$ is surjective: any discrete chain $C^k(M)$ can be extended to a smooth form $\Omega^k(M)$ (*e.g.* via some local spline basis such as the **Witney element**). One can check that the induced $\int$ on the cohomologies is also surjective. Furthermore $\int$ is injective. To see this, one shows that $\ker(\int) \cap \ker(d) \subset \Omega^k(M)$ is always exact. That is, the higher resolution information of a closed form in $\Omega^k(M)$ that cannot be resolved in the discrete sampling $\int$ will consist only of exact forms. This last claim is a bit technical. It can be accomplished by integrating ω into τ so that $d\tau = \omega$ over a domain that grows one cell at a time until it pads over the entire manifold. □

Theorem 10.8 — Poincaré Duality.

$$H_k(M) \cong H^{n-k}(M^\star), \quad \mathcal{H}^k(M) = \mathcal{H}^{n-k}(M). \tag{10.41}$$

Proof. The discrete case $H_k(M) \cong H^{n-k}(M^\star)$ is a direct consequence of how $\partial, \partial^\intercal$ on the dual $(n-k)$-(co)chain is related to $\partial, \partial^\intercal$ on the primal k-chain. See (7.16) and (7.17). For the smooth harmonic forms the equality $\mathcal{H}^k(M) = \mathcal{H}^{n-k}(M)$ is obtained by taking the Hodge star. □

10.4 Poincaré Duality

Here we describe a procedure that turns a homological chain $\Gamma \in C_{n-k}(M^\star)$ into a harmonic field $h \in \mathcal{H}^k(M)$. This will realize the Poincaré duality which maps a homology class of $H_{n-k}(M^\star)$ to a homology class of $H^k(M)$.

Suppose we are given a closed chain $\Gamma \in C_{n-k}(M^\star)$ on the dual mesh,

$$\Gamma = \sum_{e^\star \in M^\star_{n-k}} a_e\, e^\star, \quad \partial^\star_{n-k}\Gamma = 0.$$

Typically $a_e \in \{-1, 0, 1\}$ but it could also be larger integers, or real numbers, as a chain. Then, by placing the same coefficients on the corresponding basis components, we construct the primal cochain (a discrete k-form) over the primal mesh

$$\mathbb{1}_\Gamma := \sum_{e \in M_k} a_e\, \mathbb{1}_e \in C^k(M).$$

Note that

$$\partial^\star_{n-k}\Gamma = 0 \quad \Longrightarrow \quad \partial^\intercal_{k+1}\mathbb{1}_\Gamma = d_k\mathbb{1}_\Gamma = 0.$$

R On that note, in general the Poincaré duality map $\mathbb{1}_{(\cdot)} : C_{n-k}(M^\star) \to C^k(M)$ preserves the boundary/coboundary operators:

$$d_k\mathbb{1}_\Gamma = (-1)^{k+1}\mathbb{1}_{\partial^\star_{n-k}\Gamma}. \tag{10.42}$$

The k-form $\mathbb{1}_\Gamma$ is now a closed discrete k-form that is an impulse at Γ and zero elsewhere. Namely, for each primal k-chain $\Sigma \in C_k(M)$, the evaluation $\langle \mathbb{1}_\Gamma | \Sigma \rangle = \int_\Sigma \mathbb{1}_\Gamma$ counts the number of times (counted with orientation) Γ and Σ intersect.

Finally, to turn the impulse $\mathbb{1}_\Gamma$ into a discrete harmonic form in the same cohomology class, we apply (10.4):

$$h_\Gamma := (1 - dd^+ - d^+d)\mathbb{1}_\Gamma \overset{d\mathbb{1}_\Gamma = 0}{=} \mathbb{1}_\Gamma - dd^+\mathbb{1}_\Gamma. \tag{10.43}$$

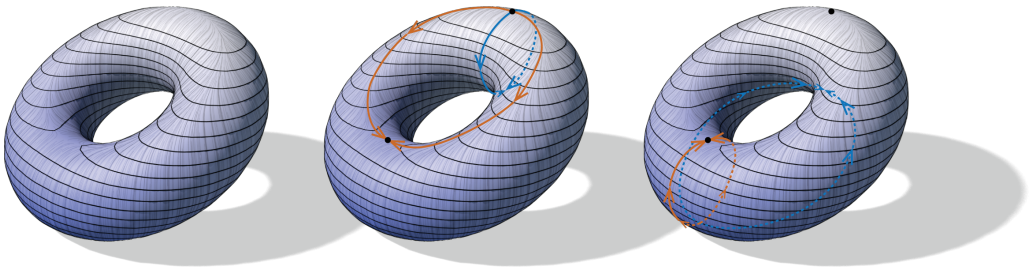


Figure 10.1 A generic smooth function defined over a torus (here the height function) will have at least one minimum, one maximum, and two saddle points. Consider a function with only one maximum and one minimum. Backward tracing the downward gradient flow from the saddles gives a basis (a.k.a. a set of generators) for $H_1(M)$. One can also backward trace the upward gradient flow from those saddles, which will be another set of basis for $H_1(M)$. These two sets of generators are related by the Poincaré duality.

10.5 Morse Theory

Here we talk about an algorithm to find the generators (basis) for $H_k(M)$ called the **tree-cotree algorithm**. The underlying theory is the **Morse theory** where we look at yet another chain complex.

The idea is to find an extremely coarse cell decomposition of M so that the skeleton of that discretization automatically reveals the homology generators. A cell is a region that can be retracted via deformation to a point. Let us represent such a motion of retraction by a flow that concentrates to a point. Such a flow must be a gradient flow. (Can you see using Hodge decomposition what could happen if the flow is not a gradient flow?)

We choose an arbitrary function $f : M \to \mathbb{R}$ and consider its gradient flow. On most region the flow will drain down to the minimum of the function, except for the "drainage divides." These drainage divides are lower dimensional manifolds connecting saddle points. Within each of the divides the flow retracts most region to some critical point, divided by even lower dimensional drainage divides. This gives rise to a coarse cell decomposition.

Figure 10.1 and Figure 10.2 are examples on surfaces. The function $f : M \to \mathbb{R}$ is chosen to be the height function of the given embedding in $\mathbb{R}^3$. The drainage divides are found by starting with the saddle point and tracing the flow backwards to the maximum. As shown in both figures, when f has only one local minimum and one local maximum, the "territory" that flows to the minimum is one large cell (as the only element in the 2-skeleton) bordered by the drainage divide loops (as the 1-skeleton) anchored at the maximum point (as the only vertex for the 0-skeleton).

In this case the drainage divides must be the homology generators. They are closed loops as shown in the figures. They cannot be the boundary of any 2-chain, and any two of them cannot coborder any 2-chain, otherwise the flow separated by the these divides would have more than one minimum.

10.5.1 Morse Homology

The above intuition about the drainage divide is enough for the reader to proceed to the tree-cotree algorithm. Here we explain some mathematical terminologies in the Morse theory.

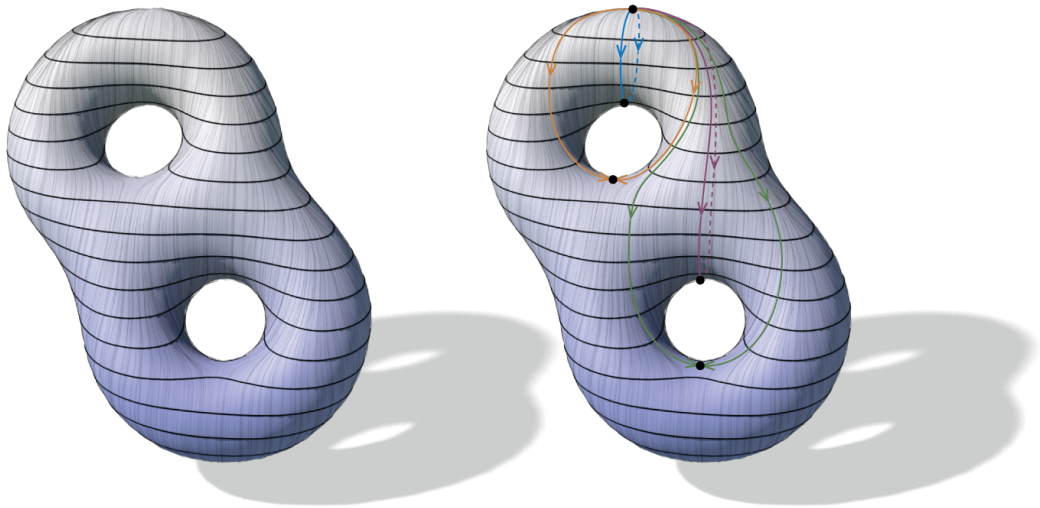


Figure 10.2 Consider a function over the genus-2 surface M so that there is only one maximum and one minimum. Tracing the gradient flow from the saddle points back to the maximum gives a set of generators for $H_1(M)$

Consider a generic smooth function $f\colon M \to \mathbb{R}$. A typical example of f on a surface is the height function of an embedding of the surface M in $\mathbb{R}^3$. Let us also assume that M has a Riemannian metric, so we may talk about the gradient descend flow on M.

A point $p \in M$ is called a **critical point** of f if $(df)|_p = 0$, or equivalently $(\operatorname{grad} f)|_p = 0$. Up to perturbation, we may assume that the critical points of f are finitely many isolated points. Moreover f has non-degenerate second derivative (Hessian) at each critical point. A function with these assumptions are called a **Morse function**.

Definition 10.6 Each critical point p of f has a **Morse index**. The Morse index at p is the number of directions (counted in dimension) in which the gradient descent flow $-\operatorname{grad} g$ departs from p.

This definition can be made precise by the linear stability analysis for flows around critical points. The Morse index $\operatorname{index}(p)$ of a critical point p of a function f is the number of negative eigenvalues of the Hessian matrix[a]

$$\operatorname{Hess}_{\mathbf{r}} g = \begin{bmatrix} \frac{\partial^2 f}{\partial r_1^2} & \frac{\partial^2 f}{\partial r_1 \partial r_2} & \cdots & \frac{\partial^2 f}{\partial r_1 \partial r_n} \\ \frac{\partial^2 f}{\partial r_1 \partial r_2} & \frac{\partial^2 f}{\partial r_2^2} & \cdots & \frac{\partial^2 f}{\partial r_2 \partial r_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial r_1 \partial r_n} & \frac{\partial^2 f}{\partial r_2 \partial r_n} & \cdots & \frac{\partial^2 f}{\partial r_n^2} \end{bmatrix} \tag{10.44}$$

using any coordinate $\mathbf{r}\colon U \ni p \to \mathbb{R}^n$ around p.

[a] The actual coordinate-independent Hessian operator is given by $\operatorname{Hess} f = \nabla(\operatorname{grad} f)$, the covariant derivative of $\operatorname{grad} f$. If it is just the number of negative eigenvalues is concerned, it is fine to just take the coordinate-dependent Hessian matrix.

If the Morse index of a critical point is 0 (*i.e.* all eigenvalues of the Hessian is positive), then this critical point is a **local minimum**. If the Morse index of a critical point is n (*i.e.* all eigenval-

ues of the Hessian is negative), then this critical point is a **local maximum**. If the Morse index is any other number in between, the critical point is called a **saddle**.

Definition 10.7 — Stable and unstable manifold. Let $\phi_t : M \to M$ be the flow map generated by the gradient descent flow. That is $\frac{d}{dt}\phi_t = -\operatorname{grad} f|_{\phi_t}$ for $-\infty < t < \infty$ and $\phi_{t=0} = \mathrm{id}_M$. Let p be a critical point of f. We call the collection of points that converge to p under the flow the **stable manifold** of p:

$$\mathrm{Stable}(p) := \left\{ x \in M \,\middle|\, \lim_{t\to+\infty} \phi_t(x) = p \right\}. \tag{10.45}$$

We call the collection of points that depart from p (*i.e.* converge to p under the backward flow) the **unstable manifold** of p

$$\mathrm{Unstable}(p) := \left\{ x \in M \,\middle|\, \lim_{t\to-\infty} \phi_t(x) = p \right\}. \tag{10.46}$$

In fact,

$$\dim\big(\mathrm{Unstable}(p)\big) = \mathrm{index}(p). \tag{10.47}$$

Let p, q be two critical points, $\mathrm{index}(p) > \mathrm{index}(q)$. The **flow line** (or flow surface) from p to q is given by

$$\mathrm{Flow}(p,q) := \mathrm{Ustable}(p) \cap \mathrm{Stable}(q). \tag{10.48}$$

Note that

$$\dim\big(\mathrm{Flow}(p,q)\big) = \mathrm{index}(p) - \mathrm{index}(q). \tag{10.49}$$

Let $\mathrm{Crit}_k := \{p_1^{(k)}, \ldots, p_{m_k}^{(k)}\}$ denote all the critical points with Morse index k. Define the **Morse k-chain** as the formal linear space with each $p_i^{(k)}$ being a basis element

$$C_k^{\mathrm{Morse}}(f) := \mathrm{span}\left\{p_1^{(k)}, \ldots, p_{m_k}^{(k)}\right\} = \left\{ \sum_{i=1}^{m_k} c_i\, p_i^{(k)} \,\middle|\, c_i \in \mathbb{R} \right\} \cong \mathbb{R}^{m_k}. \tag{10.50}$$

There is a boundary operator $\partial_k^{\mathrm{Morse}} : C_k^{\mathrm{Morse}}(f) \to C_{k-1}^{\mathrm{Morse}}(f)$, a linear operator, which sends each basis element $p_i^{(k)}$ to a linear combination of index-$(k-1)$-critical points weighted by the number of flow lines connecting with $p_i^{(k)}$:

$$\partial_k^{\mathrm{Morse}} p_i^{(k)} = \sum_{q \in \mathrm{Crit}_{k-1}} \#\,\mathrm{Flow}(p_i^{(k)}, q)\, q. \tag{10.51}$$

The count $\#\,\mathrm{Flow}(p_i^{(k)}, q)$ of flow lines are signed integers. A flow line $\gamma \subset \mathrm{Flow}(p,q)$, $\gamma(-\infty) = p$, $\gamma(\infty) = q$, is counted as $+1$ (resp. -1) if concatenating the direction γ' with a positively-oriented basis for the unstable subspace at q (tangent plane $\mathrm{Unstable}(q)$ at q) gives a positively (resp. negatively) oriented basis for the surrounding $\mathrm{Unstable}(p)$.

In both examples Figure 10.1 and Figure 10.2, each pair of flow lines (in the same color) are counted with opposite signs, hence $\partial_2^{\mathrm{Morse}} = 0$.

In general, one can show that

$$\partial_{k-1}^{\text{Morse}} \circ \partial_k^{\text{Morse}} = 0. \tag{10.52}$$

In other words,

$$\cdots \xrightarrow{\partial_{k+2}^{\text{Morse}}} C_{k+1}^{\text{Morse}}(f) \xrightarrow{\partial_{k+1}^{\text{Morse}}} C_k^{\text{Morse}}(f) \xrightarrow{\partial_k^{\text{Morse}}} C_{k-1}^{\text{Morse}}(f) \xrightarrow{\partial_{k-1}^{\text{Morse}}} \cdots \tag{10.53}$$

is a chain complex.

Theorem 10.9 The Morse homology

$$H_k^{\text{Morse}}(f) := \frac{\ker(\partial_k^{\text{Morse}})}{\text{im}(\partial_{k+1}^{\text{Morse}})} \tag{10.54}$$

is isomorphic to the de Rham and the cellular homology

$$H_k^{\text{Morse}}(f) \cong H_k(M) \tag{10.55}$$

on finite dimensional M.

Now we see that the de Rham cohomology, homology of a discrete mesh, and the Morse homology are all isomorphic. To find the generators for the homology, we work with the one where the chain complex is most manageable. The (smooth) de Rham complex consists of the space of differential forms $\Omega^k(M)$ which are infinite dimensional spaces. The chains $C_k(M)$ and cochains $C^k(M)$ on a discrete mesh are finite dimensional spaces but typically the dimension is large depending on the mesh resolution. In contrast, the Morse chain complex is a tiny space; it is just spanned by the critical points of a smooth function f. Therefore we search the generators for the homology by working with the Morse chain complex.

10.5.2 Tree-Cotree Algorithm

The tree-cotree algorithm finds a set of generators for $H_1(M^\star)$ and/or $H_1(M)$ for a discrete surface M. Here we assume M is closed. The algorithm is based on the discrete Morse homology theory.

The idea is to effectively generate some gradient flow over M. We will let the flow to have only one local maximum and one local minimum, so that the drainage divides automatically form the homology generators.

We label an arbitrary face $\sigma \in M_2 = M_0^\star$ as the "summit", and an arbitrary vertex $p \in M_0 = M_2^\star$ as the "oceanic trench."

Take $\sigma \in M_0^\star$ as the root and build a spanning tree over the faces, *i.e.* a spanning tree in $M_1^\star$, by a breath-first search or a depth-first search algorithm. We call this tree a **cotree** $T^\star \subset M_1^\star$. It represents mountain ridges. It effectively gives a Morse function value f in a decreasing order from the root to the leaves.

Next, take the trench p as the root and build a spanning tree in $M_1 \setminus T^\star$. That is, build a spanning tree for the primal mesh using remaining edges that have not been crossed by the cotree. We call this resulting tree $T \subset M_1 \setminus T^\star$ a **primal tree**. It represents a river system. One may immagine Morse function values f are now given along the rivers in a ascending order from the root to the leaves. To make sure it is a proper Morse function without additional local

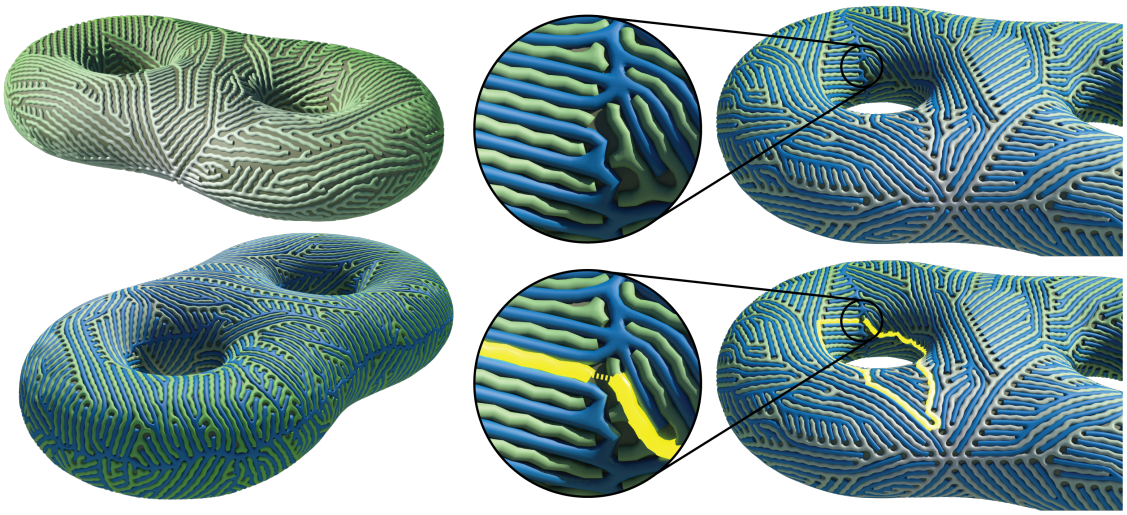


Figure 10.3 Tree-cotree algorithm. Generate a spanning tree over the dual mesh (*i.e.* the dual tree, or cotree) representing the mountain ridges (top left). Then generate a spanning tree over the remaining primal edges (the primal tree) that are not crossed by any mountain ridge; this represents the river system (bottom left). Then there will be exactly $\beta_1 = \dim(H_1(M))$ many edges that are neither crossed by a ridge nor a river; these are the saddles (top right). Connecting the dual edge (resp. primal edge) at each of the saddles gives a cycle in the mountain range (resp. river system). These cycles are a set of generators for $H_1(M^\star)$ (resp. $H_1(M)$).

maximum or minimum, each of the f value at a river (primal vertex) is strictly less than an f value at any mountain ridge (dual vertex).

Then, one finds a few edges $\{s_1, \ldots, s_{\beta_1}\} \in M_1 \setminus (T \cup T^\star)$ that are neither crossed by a river nor a mountain ridge. These edges are the **saddle edges**. There are exactly β_1 saddle edges, where $\beta_1 = \dim(H_1(M)) = \dim(H_1(M^\star))$ is the first Betti number. For a closed oriented surface $\beta_1 = 2g$ where g is the genus.

Finally, for each saddle edge s_j, connect the mountain ridge then we obtain a cycle $\gamma_j^\star$ in the cotree $T^\star \cup \{s_j^\star\}$. If we instead connect the river ends, we obtain a cycle γ_j in the tree $T \cup \{s_j\}$. From an earlier discussion about drainage divide, we know that $\gamma_1^\star, \ldots, \gamma_{\beta_1}^\star$ will be linearly independent in $H_1(M^\star)$, and similarly for $\gamma_1, \ldots, \gamma_{\beta_1}$ in $H_1(M)$. These cycles are the output of the algorithm.

Exercise 10.3 Using that the Euler characteristic $\chi(M)$ is given by $\chi(M) = |V| - |E| + |F| = 2 - 2g$, show that there are exactly $2g$ saddle edges.
Hint *Every node except for the root in a spanning has one edge pointing to a parent node. So, how many edges are there in the cotree and how many edges are there in the primal tree?* ∎

Algorithm 10.2 First homology generators for a closed surface

Input: Closed discrete surface $M = (V, E, F)$.

Let the cotree $T^\star \subset E^\star$ be a face spanning tree.
Let the tree $T \subset E \setminus (T^\star)^\star$ be a vertex spanning tree.
Let $S = E \setminus (T \cup (T^\star)^\star)$. ▷ $|S| = \beta_1$.
$k \leftarrow 0$. ▷ Counter
for $s \in S$ **do**
    $k \leftarrow k + 1$.
    Let $\gamma_k^\star$ be the unique cycle in $\{s^\star\} \cup T^\star$.
    Let γ_k be the unique cycle in $\{s\} \cup T$.
end for

Output: $(\gamma_1, \ldots, \gamma_{\beta_1})$ is a set of generators for $H_1(M)$, and $(\gamma_1^\star, \ldots, \gamma_{\beta_1}^\star)$ is a set of generators for $H_1(M^\star)$.

In fact, the (transversal) intersection between γ_i and $\gamma_j^\star$ will be 1 if $i = j$ and otherwise 0. This correspondance between the respective generators can be taken as the Poincaré duality between $H_1(M) \cong H^1(M)$ and $H_1(M^\star)$.

▪ **Example 10.6 — Basis for harmonic 1-forms.** To generate a basis for the harmonic forms $\mathcal{H}^1(M)$, take each generator $\gamma_j^\star \in C_1(M^\star)$, consider the impulse $\mathbb{1}_{\gamma_j^\star} \in C^1(M)$, and define $h_j := (1 - dd^+)\mathbb{1}_{\gamma_j^\star}$. See Section 10.4. Note that $(h_1, \ldots, h_{\beta_1})$ are not necessarily an orthonormal basis for $\mathcal{H}^1(M)$ with respect to the L^2-inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle = \int_M (\cdot) \wedge_1 \star(\cdot)$. Using this L^2-inner product (given by the discrete $\star_1$ as the inner product matrix) one may apply the Gram–Schmidt process, or an (economic) QR factorization functional call in a numerical linear algebra library, to produce an orthonormal basis $(\hat{h}_1, \ldots, \hat{h}_{\beta_1})$ for $\mathcal{H}^1(M)$.

An orthonormal basis is convenient for orthogonal projection:

$$P_{\mathcal{H}^1(M)}\alpha = \sum_{i=1}^{\beta_1} \langle\!\langle \hat{h}_i, \alpha \rangle\!\rangle \hat{h}_i.$$

▪

11. Hodge Decomposition with Boundary

In Chapter 10 we looked at the Hodge decomposition on manifolds without boundary. In this chapter we consider manifolds with boundary $\partial M \neq \varnothing$.

In the case of closed manifold, we have seen that the harmonic part of the Hodge decomposition is given by the homology and cohomology theory. There was not a real distinction between the different homologies (de Rham, cellular, Morse, *etc.*) for closed manifolds. In contrast, for manifolds with boundary there is another homology called the **relative homology** that shows up depending on the boundary condition of the problem.

At the boundary, we will need to talk about the tangential and normal components of a differential form. This would be the ingredient for setting up the right boundary condition for the Hodge decomposition.

11.1 Exterior Algebra at the Boundary

Let M be an n-dimensional Riemannian manifold with boundary ∂M. the boundary ∂M is regarded as an $(n-1)$-dimensional submanifold in M. We let the **unit outward normal** vector field of ∂M be denoted by $\mathbf{n}$.

Note that the normal vector $\mathbf{n}$ is still a *tangent vector of* M. It is defined only at ∂M and is normal to the tangent space $T(\partial M)$ of ∂M. For formula that involves $\mathbf{n}$ we automatically assume that the formula takes place only at the boundary.

11.1.1 Pullback to Boundary

Sometimes we will explicitly write the **inclusion map**

$$j_\partial : \partial M \to M \tag{11.1}$$

as the embedding of ∂M in M. It is a useful notation. For example, if $\mu = \star_0 1 \in \Omega^n(M)$ is the volume form of M, then "$\mu_\partial = i_{\mathbf{n}}\mu$" is the area form for ∂M. However technically, $i_{\mathbf{n}}\mu$ is

still an $(n-1)$-form of M (taking place at the boundary), instead of an $(n-1)$-form of ∂M as should μ_∂. To be precise

$$i_{\mathbf{n}}\mu \in \Omega^{n-1}(M)|_{\partial M}, \quad \mu_\partial \in \Omega^{n-1}(\partial M). \tag{11.2}$$

In this case, we will use the pullback operator by the inclusion map to obtain consistent types

$$\mu_\partial = j_\partial^*(i_{\mathbf{n}}\mu). \tag{11.3}$$

Once we get used to how to insert j_∂^* in various places to get precise formula such as (11.3), we will sometimes omit j_∂^*. That is, when we write

$$\mu_\partial = i_{\mathbf{n}}\mu \tag{11.4}$$

we mean (11.3). In fact we have been omitting the pullback operator in Section 9.2.

Now let us take a closer look at the pullback operator to the boundary

$$j_\partial^* \colon \Omega^k(M)|_{\partial M} \to \Omega^k(\partial M). \tag{11.5}$$

A k-form $\alpha \in \Omega^k(M)$ of M restricted at the boundary, $\alpha|_{\partial M} \in \Omega^k(M)|_{\partial M}$ is still a k-multilinear form of the entire tangent space $T_p M$ of M at any $p \in \partial M$. In contrast, $j_\partial^*\alpha = j_\partial^*(\alpha|_{\partial M})$ is a k-multilinear form of only the subspace $T_p(\partial M) \subset T_p M$. The form $j_\partial^*\alpha$ can only take vectors that are tangent to ∂M. Hence in a sense $j_\partial^*\alpha$ projects α to its tangent component.

Just like other pullback operators it distributes over wedges:

$$j_\partial^*(\omega \wedge \eta) = (j_\partial^*\omega) \wedge (j_\partial^*\eta). \tag{11.6}$$

The pullback j_∂^* commute with the exterior derivative in the following sense

$$d\Big(j_\partial^*\left(\omega|_{\partial M}\right)\Big) = j_\partial^*\Big((d\omega)|_{\partial M}\Big). \tag{11.7}$$

R Note that the restriction symbol $(\cdot)|_{\partial M}$ does not commute with the exterior derivative ($(d\omega)|_{\partial M} \neq d(\omega|_{\partial M})$); in fact it is not clear what $d(\omega|_{\partial M})$ means since we need the information of ω in a full neighborhood (not only $\omega|_{\partial M}$) to take d.

As a reminder, pullbacks do not commute with the Hodge stars $\star$.

11.1.2 Interior and Exterior Product with Normal

Recall the following facts about interior product $i_{\mathbf{n}}$

$$\begin{aligned} &i_{\mathbf{n}} \colon \Omega^k(M)|_{\partial M} \to \Omega^{k-1}(M)|_{\partial M} \\ &(i_{\mathbf{n}}\alpha)(\mathbf{v}_1, \dots, \mathbf{v}_{k-1}) = \alpha(\mathbf{n}, \mathbf{v}_1, \dots, \mathbf{v}_{k-1}). \end{aligned}$$

- $i_{\mathbf{n}} \circ i_{\mathbf{n}} = 0$.
- Leibniz rule: $i_{\mathbf{n}}(\alpha \wedge \beta) = (i_{\mathbf{n}}\alpha) \wedge \beta + (-1)^k \alpha \wedge (i_{\mathbf{n}}\beta)$ where α is a k-form.
- For each $\omega \in \Omega^k(M)|_{\partial M}$,

$$(-1)^k \star^{-1} i_{\mathbf{n}} \star \omega = \mathbf{n}^\flat \wedge \omega. \tag{11.8}$$

- Equivalent to the previous item, for $\alpha \in \Omega^k(M)|_{\partial M}, \beta \in \Omega^{k-1}(M)|_{\partial M}$,

$$\langle i_{\mathbf{n}}\alpha, \beta\rangle = \left\langle \alpha, \mathbf{n}^\flat \wedge \beta \right\rangle \tag{11.9}$$

where $\langle \omega, \eta\rangle = \star(\omega \wedge \star\eta)$.

Eq. (11.8) can be summarized as

$$\begin{array}{ccccccccc}
\Omega^0(M)|_{\partial M} & \xleftarrow{i_{\mathbf{n}}} & \Omega^1(M)|_{\partial M} & \xleftarrow{i_{\mathbf{n}}} & \Omega^2(M)|_{\partial M} & \longleftarrow\cdots & \Omega^{n-1}(M)|_{\partial M} & \xleftarrow{i_{\mathbf{n}}} & \Omega^n(M)|_{\partial M} \\
\downarrow{\scriptstyle \star_0} & \nearrow_{\mathbf{n}^\flat\wedge} & \downarrow{\scriptstyle \star_1} & \nearrow_{\mathbf{n}^\flat\wedge} & \downarrow{\scriptstyle \star_3} & & \downarrow{\scriptstyle \star_{n-1}} & \nearrow_{\mathbf{n}^\flat\wedge} & \downarrow{\scriptstyle \star_n} \\
\Omega^n(M)|_{\partial M} & \xrightarrow[i_{\mathbf{n}}]{} & \Omega^{n-1}(M)|_{\partial M} & \xrightarrow[-i_{\mathbf{n}}]{} & \Omega^{n-2}(M)|_{\partial M} & \longrightarrow\cdots & \Omega^1(M)|_{\partial M} & \xrightarrow[(-1)^{n-1}i_{\mathbf{n}}]{} & \Omega^0(M)|_{\partial M}
\end{array} \tag{11.10}$$

Both rows of (11.10) are indeed **chain complexes**, that is $i_{\mathbf{n}} \circ i_{\mathbf{n}} = 0$ and $(\mathbf{n}^\flat\wedge) \circ (\mathbf{n}^\flat\wedge) = 0$.

Unlike the chain complex for d and δ (de Rham complex), here there is in fact no gap between $\ker(i_{\mathbf{n}})$ and $\mathrm{im}(i_{\mathbf{n}})$. That is $i_{\mathbf{n}}$-closed implies $i_{\mathbf{n}}$-exact. Similarly, $(\mathbf{n}^\flat\wedge)$-closed implies $(\mathbf{n}^\flat\wedge)$-exact. In technical term, such a trivial (no-homology) chain complex is called an **exact sequence**. The following theorem asserts that this is true.

Theorem 11.1 Suppose $\omega \in \Omega^k(M)|_{\partial M}$.
- If $i_{\mathbf{n}}\omega = 0$, then $\omega = i_{\mathbf{n}}(\mathbf{n}^\flat \wedge \omega)$.
- If $\mathbf{n}^\flat \wedge \omega = 0$, then $\omega = \mathbf{n}^\flat \wedge (i_{\mathbf{n}}\omega)$.

In other words, $i_{\mathbf{n}}$-closed implies $i_{\mathbf{n}}$-exact, and $(\mathbf{n}^\flat\wedge)$-closed implies $(\mathbf{n}^\flat\wedge)$-exact.

Proof. By Leibniz rule $i_{\mathbf{n}}(\mathbf{n}^\flat \wedge \omega) = \underbrace{(i_{\mathbf{n}}\mathbf{n}^\flat)}_{=1}\wedge\omega - \mathbf{n}^\flat \wedge (i_{\mathbf{n}}\omega)$ we have

$$\omega = i_{\mathbf{n}}(\mathbf{n}^\flat \wedge \omega) + \mathbf{n}^\flat \wedge (i_{\mathbf{n}}\omega). \tag{11.11}$$

If $i_{\mathbf{n}}\omega = 0$, then the second term vanishes, yielding $\omega = i_{\mathbf{n}}(\mathbf{n}^\flat \wedge \omega)$. If $\mathbf{n}^\flat \wedge \omega = 0$ then the first term vanishes, and thus $\omega = \mathbf{n}^\flat \wedge (i_{\mathbf{n}}\omega)$. □

Corollary 11.2

$$\ker\left(i_{\mathbf{n}}|_{\Omega^k}\right) = \mathrm{im}(i_{\mathbf{n}}|_{\Omega^{k+1}}), \quad \ker((\mathbf{n}^\flat\wedge)_{\Omega^k}) = \mathrm{im}((\mathbf{n}^\flat\wedge)_{\Omega^{k-1}}) \tag{11.12}$$

11.1.3 Tangent and Normal Components

Definition 11.1 — Tangent/normal to the boundary. We say $\omega \in \Omega^k(M)|_{\partial M}$ is
- **tangent to the boundary** if $i_{\mathbf{n}}\omega = 0$; *i.e.* $\omega \in \ker(i_{\mathbf{n}}) = \mathrm{im}(i_{\mathbf{n}})$.
- **normal to the boundary** if $\mathbf{n}^\flat \wedge \omega = 0$; *i.e.* $\omega \in \ker(\mathbf{n}^\flat\wedge) = \mathrm{im}(\mathbf{n}^\flat\wedge)$.

Since $i_{\mathbf{n}}$ and $\mathbf{n}^\flat\wedge$ are adjoint operators in the sense of (11.9), the forms tangent to boundary and the forms normal to boundary are in fact pointwise orthogonal complements to each other

(Theorem 10.1). That is, each $\Omega^k(M)|_{\partial M}$ orthogonally splits into the tangent and normal components:

$$\Omega^\bullet(M)|_{\partial M} = \ker(i_{\mathbf{n}}) \oplus \operatorname{im}(\mathbf{n}^\flat \wedge) = \underbrace{\ker(i_{\mathbf{n}})}_{\text{tangent}} \oplus \underbrace{\ker(\mathbf{n}^\flat \wedge)}_{\text{normal}} \tag{11.13}$$

In fact (11.11) is the formula that talks about splitting a form ω into its tangent and normal components. We define the following projection operators to these components.

Definition 11.2 — Projections to the tangent and normal components. Define

$$\mathbb{t} := i_{\mathbf{n}} \circ (\mathbf{n}^\flat \wedge), \quad \mathbb{n} := (\mathbf{n}^\flat \wedge) \circ i_{\mathbf{n}}, \tag{11.14}$$

both of which are maps $\Omega^k(M)|_{\partial M} \to \Omega^k(M)|_{\partial M}$. The $\mathbb{t}$ operator extracts the **tangent component**, which we call the **Dirichlet data**, while $\mathbb{n}$ extracts the **normal component**, also known as the **Neumann data**. Sometimes $\mathbb{t}$ is also called the **projection** and $\mathbb{n}$ is called the **rejection**.

Corollary 11.3 $\mathbb{t}, \mathbb{n}$ are projection operators, *i.e.* $\mathbb{t}^2 = \mathbb{t}$ and $\mathbb{n}^2 = \mathbb{n}$.

Due to (11.11) we have

Corollary 11.4

$$\mathbb{t} + \mathbb{n} = \mathrm{id}\,. \tag{11.15}$$

Corollary 11.5 The space of k-forms that are tangent (resp. normal) to the boundary can be written as

$$\ker(i_{\mathbf{n}}) = \operatorname{im}(i_{\mathbf{n}}) = \operatorname{im}(\mathbb{t}) = \ker(\mathbb{n}) \tag{11.16}$$

$$\ker(\mathbf{n}^\flat \wedge) = \operatorname{im}(\mathbf{n}^\flat \wedge) = \operatorname{im}(\mathbb{n}) = \ker(\mathbb{t}). \tag{11.17}$$

Since forms that are normal to the boundary ($\ker(\mathbf{n}^\flat \wedge)$) are in fact in $\ker(\mathbb{t})$, these normal-to-boundary forms are also said to be **zero Dirichlet**. Likewise, we say forms tangent to the boundary are **zero Neumann**.

The following relation between $\mathbb{t}, \mathbb{n}$ and Hodge star $\star$ is useful later on.

Theorem 11.6 For each $\omega \in \Omega^k(M)|_{\partial M}$

$$\mathbb{t} \star \omega = \star\, \mathbb{n}\omega, \quad \mathbb{n} \star \omega = \star\, \mathbb{t}\omega. \tag{11.18}$$

Another characterization of $\mathbb{t}$ is the following. Let $\omega \in \Omega^k(M)|_{\partial M}$. Then $\mathbb{t}\omega \in \Omega^k(M)|_{\partial M}$ is given by

$$(\mathbb{t}\omega)(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \omega(P\mathbf{v}_1, \ldots, P\mathbf{v}_k) \tag{11.19}$$

where P is the orthogonal projection $T_x M \to T_x(\partial M)$ at each $x \in \partial M \subset M$. Note that $\mathbb{n}$ does not have a comparably simple formula as (11.19). The implication of (11.19) is that $\mathbb{t}$ projects ω

to a component that only sees the vectors tangent to the boundary, which is effectively what j_∂^* does. In other words, j_∂^* is an isomorphism between $\operatorname{im}(\mathbb{t}) = \ker(i_{\mathbf{n}}) = \ker(\mathbb{n})$ and $\Omega^\bullet(\partial M)$:

$$j_\partial^* : \underbrace{\operatorname{im}(\mathbb{t}) \subset \Omega^k(M)|_{\partial M}}_{\substack{\text{tangent} \\ \text{component}}} \xrightarrow{\cong} \Omega^k(\partial M). \tag{11.20}$$

This tells us when it is fine to omit j_∂^* in the discussion in Section 11.1.1. We may omit the pullback j_∂^* in (11.3) and write (11.4) since $i_{\mathbf{n}}\mu \in \operatorname{im}(i_{\mathbf{n}}) = \operatorname{im}(\mathbb{t})$. In general, one treats the tangent component $\operatorname{im}(\mathbb{t}) = \ker(i_{\mathbf{n}})$ the same as $\Omega^k(\partial M)$, and even treat $\mathbb{t}$ as the pullback operator j_∂^*:

$$j_\partial^* \omega = j_\partial^* \mathbb{t}\omega = \mathbb{t}\omega. \tag{11.21}$$

Corollary 11.7 $\mathbb{t}$ has similar properties as pullbacks such as

- it distributes over wedges:

$$\mathbb{t}(\omega \wedge \eta) = (\mathbb{t}\omega) \wedge (\mathbb{t}\eta). \tag{11.22}$$

- it commutes with exterior derivatives:

$$\mathbb{t} d\omega = d_\partial \mathbb{t}\omega. \tag{11.23}$$

Finally, suppose $\star_\partial$ is the Hodge star on ∂M. Then it has the following relation with the Hodge star $\star$ in M.

Theorem 11.8 For each tangential $\alpha \in \operatorname{im}(\mathbb{t}) \subset \Omega^k(M)|_{\partial M}$, we have

$$\star_\partial \alpha = \star(\mathbf{n}^\flat \wedge \alpha) = (-1)^k i_{\mathbf{n}}(\star\alpha). \tag{11.24}$$

Or more generally, for any $\alpha \in \Omega^k(M)|_{\partial M}$,

$$\star_\partial \mathbb{t}\alpha = \star(\mathbf{n}^\flat \wedge \alpha) = (-1)^k i_{\mathbf{n}}(\star\alpha). \tag{11.25}$$

Corollary 11.9 Suppose $\alpha \in \Omega^k(M)|_{\partial M}$ and $\beta \in \Omega^{k+1}(M)|_{\partial M}$. Then

$$(\mathbb{t}\alpha) \wedge \star(\mathbb{n}\beta) = (\mathbb{t}\alpha) \wedge \star_\partial(i_{\mathbf{n}}\beta). \tag{11.26}$$

Exercise 11.1 Use Theorem 11.8 to show Corollary 11.9.
Hint *$i_{\mathbf{n}}\beta$ is tangent to the boundary.* ∎

11.2 Green's Identity

In Theorem 6.9 we have that on manifolds without boundary $\langle\!\langle d\alpha, \beta \rangle\!\rangle = \langle\!\langle \alpha, \delta\beta \rangle\!\rangle$. This is no longer true if the domain has a boundary, unless we include a boundary term.

Theorem 11.10 — Integration by parts with boundary (Green's Identity). Let $\mathcal{M}$ be a compact Riemannian manifold with boundary. Then for each $\alpha \in \Omega^{k-1}(\mathcal{M})$ and $\beta \in \Omega^k(\mathcal{M})$,

$$\langle\!\langle d\alpha, \beta \rangle\!\rangle = \langle\!\langle \alpha, \delta\beta \rangle\!\rangle + \oint_{\partial\mathcal{M}} (\mathbb{t}\alpha) \wedge \star(\mathbb{n}\beta). \tag{11.27}$$

By Corollary 11.9 one may also write (11.27) as

$$\langle\!\langle d\alpha, \beta \rangle\!\rangle = \langle\!\langle \alpha, \delta\beta \rangle\!\rangle + \langle\!\langle \mathbb{t}\alpha, i_{\mathbf{n}}\beta \rangle\!\rangle_{\partial\mathcal{M}}. \tag{11.28}$$

Proof. By Leibniz rule $d(\alpha \wedge \star\beta) = (d\alpha) \wedge \star\beta - (-1)^k \alpha \wedge d(\star\beta) = (d\alpha) \wedge \star\beta - \alpha \wedge \star(\delta\beta)$. Integrating the expression over $\mathcal{M}$ and apply Stokes theorem we find

$$\oint_{\partial\mathcal{M}} \alpha \wedge \star\beta = \int_{\mathcal{M}} (d\alpha) \wedge \star\beta - \int_{\mathcal{M}} \alpha \wedge \star(\delta\beta) = \langle\!\langle d\alpha, \beta \rangle\!\rangle - \langle\!\langle \alpha, \delta\beta \rangle\!\rangle.$$

Then rewrite the left-hand side as follows

$$\begin{aligned}\oint_{\partial\mathcal{M}} \alpha \wedge \star\beta = \oint_{\partial\mathcal{M}} j_{\partial\mathcal{M}}^*(\alpha \wedge \star\beta) = \oint_{\partial\mathcal{M}} \mathbb{t}(\alpha \wedge \star\beta) \\ \overset{(11.22)}{=} \oint_{\partial\mathcal{M}} (\mathbb{t}\alpha) \wedge (\mathbb{t} \star \beta) \overset{\text{Theorem 11.6}}{=} \oint_{\partial\mathcal{M}} (\mathbb{t}\alpha) \wedge \star(\mathbb{n}\beta).\end{aligned}$$

□

Therefore, d and δ are not adjoint operators with respect to $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ unless the boundary term vanishes.

Corollary 11.11 $\langle\!\langle d\alpha, \beta \rangle\!\rangle = \langle\!\langle \alpha, \delta\beta \rangle\!\rangle$ if $\mathbb{t}\alpha = 0$ or $\mathbb{n}\beta = 0$.

11.3 Dirichlet and Neumann Boundary Conditions

In Corollary 11.11 we see that we recover the adjointness between d and δ if we restrict ourselves to $\mathbb{t}\alpha = 0$ or $\mathbb{n}\beta = 0$ on the boundary. To develop a Hodge decomposition theory we first consider the subspaces satisfying the $\mathbb{t}\alpha = 0$ or $\mathbb{n}\beta = 0$ boundary condition.

Definition 11.3 We define the following subspaces of $\Omega^k(\mathcal{M})$:

- **Dirichlet subspace**

$$\Omega_D^k(\mathcal{M}) := \left\{ \omega \in \Omega^k(\mathcal{M}) \,\middle|\, \mathbb{t}\omega|_{\partial\mathcal{M}} = 0 \right\}, \tag{11.29}$$

  i.e. the set of k-forms satisfying the zero-Dirichlet (normal-to-boundary) boundary condition.

- **Neumann subspace**

$$\Omega_N^k(\mathcal{M}) := \left\{ \omega \in \Omega^k(\mathcal{M}) \,\middle|\, \mathbb{n}\omega|_{\partial\mathcal{M}} = 0 \right\}, \tag{11.30}$$

i.e. the set of k-forms satisfying the zero-Neumann (tangent-to-boundary) boundary condition.

R In the theory of Hodge decomposition with boundary, the **Neumann boundary condition** is different from the condition what is usually referred to in PDE. The former is $\text{nt}\alpha = 0$, *i.e.* $j_\partial^* \star \alpha = 0$, while the latter is $\partial u/\partial \mathbf{n} = 0$ for 0-forms u, *i.e.* $\text{nt} du = 0$.

It turns out that these boundary conditions interplay nicely with the exterior derivatives and the codifferentials.

Theorem 11.12 Let $\alpha, \beta \in \Omega^k(M)$. Then

$$\text{tr}\alpha|_{\partial M} = 0 \quad \Longrightarrow \quad \text{tr}(d\alpha)|_{\partial M} = 0 \tag{11.31}$$

$$\text{nt}\beta|_{\partial M} = 0 \quad \Longrightarrow \quad \text{nt}(\delta\beta)|_{\partial M} = 0. \tag{11.32}$$

In other words, a Dirichlet form stays Dirichlet after being applied by d, and a Neumann form stays Neumann after being applied by δ.

Proof. The case of d and the Dirichlet condition follows from (11.23). For the Neumann case, note that $\text{nt}\delta\beta = 0$ is equivalent to $\star\text{nt}\delta\beta = \text{tr} \star \delta\beta = 0$, *i.e.* $\text{tr} d(\star\beta) = 0$. Similarly, $\text{nt}\beta = 0 \Leftrightarrow \text{tr}(\star\beta) = 0$. Hence we obtain the proof for the Neumann case by applying the result for the Dirichlet case on $\star\beta$. □

Therefore, the Dirichlet subspaces form a sub-chain complex in the de Rham complex for d:

$$\begin{array}{ccccccccc}
\ddots & & \ddots & & & & & & \\
 & {\scriptstyle d_{k-2}}\searrow & & {\scriptstyle d_{k-2}}\searrow & & & & & \\
 & & \Omega_D^{k-1}(M) & \subset & \Omega^{k-1}(M) & & & & \\
 & & & {\scriptstyle d_{k-1}}\searrow & & {\scriptstyle d_{k-1}}\searrow & & & \\
 & & & & \Omega_D^{k}(M) & \subset & \Omega^{k}(M) & & \\
 & & & & & {\scriptstyle d_{k}}\searrow & & {\scriptstyle d_{k}}\searrow & \\
 & & & & & & \Omega_D^{k+1}(M) & \subset & \Omega^{k+1}(M) \\
 & & & & & & & {\scriptstyle d_{k+1}}\searrow & & {\scriptstyle d_{k+1}}\searrow \\
 & & & & & & & & \ddots & & \ddots
\end{array} \tag{11.33}$$

Similarly the Neumann subspaces form a sub-chain complex in the de Rham complex for δ.

11.3.1 Examples in 2D and 3D

In a 3-dimensional volume, we interpret a 1-form η as the circulation of a vector field via musical isomorphism $\eta = \mathbf{v}^\flat$, and we interpet a 2-form ω as the flux of a vector field $\omega = i_{\mathbf{w}} \det = \star\mathbf{w}^\flat$ where det is the volume form. In a 2-dimensional domain, a 1-form can either be viewed as the flux $\star\mathbf{v}^\flat$ or the circulation $\mathbf{v}^\flat$.

0-forms

Let $u \in \Omega^0(M)$ be a function. Then $\mathbb{t}u = 0$ means that

$$u|_{\partial M} = 0.$$

Note that for 0-forms $\mathbb{n}u = \mathbf{n}^\flat \wedge \underbrace{i_{\mathbf{n}} u}_{=0}$ always vanishes.

▪ **Example 11.1 — Grounded boundary.** If $u \in \Omega^0(M; \mathbb{R}\,\mathrm{J/C})$ represents the electric potential, then $\mathbb{t}u = 0$ is the condition that ∂M is a grounded conductor. ▪

▪ **Example 11.2 — Zero temperature.** Suppose $u \in \Omega^0(M; \mathbb{R}\,^\circ\mathrm{C})$ represents temperature, then $\mathbb{t}u = 0$ is the condition that ∂M is made of ice which maintains a constant temperature $0\,^\circ\mathrm{C}$ at the boundary. ▪

Circulation Forms

Let $\eta = \mathbf{v}^\flat \in \Omega^1(M)$ be the circulation form of a vector field $\mathbf{v} \in \Gamma(TM)$. Then

$$\begin{aligned} \mathbb{t}\eta = 0 \quad &\Leftrightarrow \quad \mathbf{v} \text{ is normal to the boundary,} \\ \mathbb{n}\eta = 0 \quad &\Leftrightarrow \quad \mathbf{v} \text{ is tangent to the boundary.} \end{aligned}$$

The condition $\mathbb{n}\eta = 0$ is equivalent to $\mathbf{v} \cdot \mathbf{n} = 0$ at the boundary, which is also known as the **no-through boundary condition**.

▪ **Example 11.3 — Electrostatic field around conductors.** Let $E = \mathbf{E}^\flat \in \Omega^1(M; \mathbb{R}\,\mathrm{J/C})$ be a static electric field $E = -du$. Suppose ∂M is made of a conductor (not necessarily grounded). Then the electric potential must be constant on each connected component of ∂M; in particular, $\mathbf{E}$ has no tangential component on the boundary. The Dirichlet condition $\mathbb{t}E = 0$ describes it. One also sees that a grounded conductor condition ($\mathbb{t}u = 0$) implies conductor condition ($\mathbb{t}(du) = 0$) in reference to Theorem 11.12. ▪

▪ **Example 11.4 — Velocity around obstacles.** Let $\mathbf{v}$ represent the velocity field of a fluid flow. If ∂M represents the container wall or obstacles, then $\mathbb{n}\mathbf{v}^\flat = 0$. ▪

Flux Forms

Let $\sigma = \star\, \mathbf{v}^\flat \in \Omega^{n-1}(M)$ be the flux form of a vector field $\mathbf{v} \in \Gamma(TM)$. Then

$$\begin{aligned} \mathbb{t}\sigma = 0 \quad &\Leftrightarrow \quad \mathbf{v} \text{ is tangent to the boundary,} \\ \mathbb{n}\sigma = 0 \quad &\Leftrightarrow \quad \mathbf{v} \text{ is normal to the boundary.} \end{aligned}$$

Note that $\star$ swaps the tangency and normality of the vector field with the boundary. Now it is $\mathbb{t}\sigma = 0$ (equivalently $j_\partial^* \sigma = 0$) that represents the no-through boundary condition. It is intuitive to think of the no-through boundary condition as that the pullback $j_\partial^* \sigma$ of the flux σ to the boundary vanishes.

n-forms

Similar to how the 1-form and flux form are conjugate to each other, n-form and 0-form are related by $\star$ which swaps $\mathbb{n}$ and $\mathbb{t}$. Let $\rho = \star u$ be an n-form. Then $\mathbb{n}\rho = 0$ is equivalent to $u|_{\partial M} = 0$. Note that $\mathbb{t}\rho$ always vanishes. (Think of $\mathbb{t}$ as j_∂^*. The pullback of an n-form to an $(n-1)$-dimensional boundary must be zero.)

11.4 Hodge–Morrey–Friedrich Decomposition

The Hodge decomposition on a manifold with boundary involves boundary conditions. First, let us look at the L^2-orthogonal complement of various spaces.

Here we will let d_D denote the exterior derivative d restricted on the Dirichlet subspace $\Omega_D^{k-1}(M) \xrightarrow{d_D} \Omega_D^k(M)$. In particular, the exact derivative of form satisfying Dirichlet boundary condition is described as an element in $\mathrm{im}(d_D)$

$$d\Omega_D^{k-1}(M) = \mathrm{im}(d_D) \subset \Omega_D^k(M) \subset \Omega^k(M) \tag{11.34}$$

Similarly, let δ_N denote the codifferential δ restricted on the Neumann subspace $\Omega_N^{k+1}(M) \xrightarrow{\delta_N} \Omega_N^k(M)$.

Theorem 11.13 — Exact $\perp$ Neumann coclosed.

$$(\mathrm{im}(d))^\perp = \ker(\delta) \cap \ker(\mathbb{n}). \tag{11.35}$$

Proof. The space $\mathrm{im}(d)$ is given by $\{d\xi \,|\, \xi \in \Omega^{k-1}(M)\}$. Its orthogonal complement is the collection of all $\omega \in \Omega^k(M)$ such that $\langle\!\langle d\xi, \omega \rangle\!\rangle = 0$ for all $\xi \in \Omega^{k-1}(M)$. By Green's identity (11.28),

$$0 = \langle\!\langle d\xi, \omega \rangle\!\rangle = \langle\!\langle \xi, \delta\omega \rangle\!\rangle + \langle\!\langle \mathbb{t}\xi, i_{\mathbf{n}}\omega \rangle\!\rangle_{\partial M} \quad \text{for all } \xi \in \Omega^{k-1}(M). \tag{11.36}$$

This holds if and only if $\delta\omega = 0$ and $i_{\mathbf{n}}\omega|_{\partial M} = 0$. Note that $\ker(i_{\mathbf{n}}) = \ker(\mathbb{n})$. □

Theorem 11.14 — Dirichlet exact $\perp$ coclosed.

$$(\mathrm{im}(d_D))^\perp = \ker(\delta). \tag{11.37}$$

Proof. We look for the condition for an $\omega \in \Omega^k(M)$ being orthogonal to all elements in $\mathrm{im}(d_D) = \{d\xi \,|\, \xi \in \Omega_D^{k-1}(M)\}$. By Green's identity (11.28),

$$0 = \langle\!\langle d\xi, \omega \rangle\!\rangle = \langle\!\langle \xi, \delta\omega \rangle\!\rangle + \langle\!\langle \underbrace{\mathbb{t}\xi}_{=0}, i_{\mathbf{n}}\omega \rangle\!\rangle_{\partial M} = \langle\!\langle \xi, \delta\omega \rangle\!\rangle \quad \text{for all } \xi \in \Omega_D^{k-1}(M). \tag{11.38}$$

This holds for all $\xi \in \Omega_D^{k-1}(M)$ if and only if $\delta\omega = 0$. That is, $(\mathrm{im}(d_D))^\perp = \ker(\delta)$. □

Exercise 11.2 — Closed $\perp$ Neumann coexact.

$$(\mathrm{im}(\delta_N))^\perp = \ker(d). \tag{11.39}$$

▪

Exercise 11.3 — Dirichlet closed $\perp$ Coexact.

$$(\mathrm{im}(\delta))^\perp = \ker(d) \cap \ker(\mathbb{t}). \tag{11.40}$$

Now let us take $\Omega^k(M)$ and orthogonally split it into the Dirichlet exact and coclosed parts (Theorem 11.14):

$$\Omega^k(M) = \operatorname{im}(d_D) \oplus \ker(\delta). \tag{11.41}$$

Next, we observe that $\operatorname{im}(\delta_N)$ is a subspace of $\ker(\delta)$. We split $\ker(\delta)$ into $\operatorname{im}(\delta_N)$ and the orthogonal complement within $\ker(\delta)$: $\ker(\delta) = \operatorname{im}(\delta_N) \oplus (\operatorname{im}(\delta_N)^\perp \cap \ker(\delta))$. Combining with (11.41) and using Exercise 11.2 we obtain

$$\Omega^k(M) = \operatorname{im}(d_D) \oplus \operatorname{im}(\delta_N) \oplus \big(\ker(d) \cap \ker(\delta)\big). \tag{11.42}$$

Definition 11.4 — Harmonic forms. Let the space of

- **harmonic forms** be $\mathcal{H}^k(M) = \ker(d) \cap \ker(\delta) \subset \Omega^k(M)$.
- **Dirichlet harmonic forms** be $\mathcal{H}^k_D(M) = \mathcal{H}^k(M) \cap \Omega^k_D(M) = \mathcal{H}^k(M) \cap \ker(\mathbb{t})$.
- **Neumann harmonic forms** be $\mathcal{H}^k_N(M) = \mathcal{H}^k(M) \cap \Omega^k_N(M) = \mathcal{H}^k(M) \cap \ker(\mathbb{n})$.

In terms of the space of harmonic forms (11.42) is written as

$$\Omega^k(M) = \operatorname{im}(d_D) \oplus \operatorname{im}(\delta_N) \oplus \mathcal{H}^k(M). \tag{11.43}$$

This decomposition (11.43) is called the **Hodge–Morrey decomposition**. Note that unlike the Hodge decomposition for closed manifolds where $\mathcal{H}^k(M) \cong H^k(M)$ is a finite dimensional space, here $\mathcal{H}^k(M)$ is still an infinite dimensional space. The reason is that $\mathcal{H}^k(M)$ still contains exact forms and coexact forms, since (11.43) only separates out the Dirichlet exact and Neumann coexact instead of all exact and coexact forms. We shall further decompose $\mathcal{H}^k(M)$ into smaller components.

Taking $\operatorname{im}(d) \subset \mathcal{H}^k(M)$, and using $(\operatorname{im}(d))^\perp = \ker(\delta) \cap \ker(\mathbb{n})$ to find its orthogonal complement within $\mathcal{H}^k(M)$, we have

$$\begin{aligned}\mathcal{H}^k(M) = \ker(d) \cap \ker(\delta) &= \big(\operatorname{im}(d) \cap \ker(\delta)\big) \oplus \big(\ker(d) \cap \ker(\delta) \cap \ker(\mathbb{n})\big) \\ &= \underbrace{\Big(\operatorname{im}(d) \cap \mathcal{H}^k(M)\Big)}_{\text{exact harmonic}} \oplus \underbrace{\mathcal{H}^k_N(M)}_{\text{Neumann harmonic}}.\end{aligned} \tag{11.44}$$

Another possible route is to take $\operatorname{im}(\delta) \subset \mathcal{H}^k(M)$ and to orthogonally split $\mathcal{H}^k(M)$ into the coexact and the Dirichlet harmonic part (using Exercise 11.3):

$$\begin{aligned}\mathcal{H}^k(M) = \ker(d) \cap \ker(\delta) &= \big(\operatorname{im}(\delta) \cap \ker(d)\big) \oplus \big(\ker(d) \cap \ker(\delta) \cap \ker(\mathbb{t})\big) \\ &= \underbrace{\Big(\operatorname{im}(\delta) \cap \mathcal{H}^k(M)\Big)}_{\text{coexact harmonic}} \oplus \underbrace{\mathcal{H}^k_D(M)}_{\text{Dirichlet harmonic}}.\end{aligned} \tag{11.45}$$

In conclusion:

Corollary 11.15 — Hodge–Morrey–Friedrich Decomposition. The space $\Omega^k(M)$ of k-forms is orthogonally decomposed into

$$\Omega^k(M) = \underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \oplus \underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \oplus \underbrace{\operatorname{im}(d) \cap \mathcal{H}^k(M)}_{\text{exact harmonic}} \oplus \underbrace{\mathcal{H}^k_N(M)}_{\text{Neumann harmonic}} \tag{11.46}$$

or

$$\Omega^k(M) = \underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \oplus \underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \oplus \underbrace{\operatorname{im}(\delta) \cap \mathcal{H}^k(M)}_{\text{coexact harmonic}} \oplus \underbrace{\mathcal{H}^k_D(M)}_{\text{Dirichlet harmonic}} \tag{11.47}$$

Regrouping different components yields

$$\Omega^k(M) = \underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \oplus \underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \oplus \underbrace{\mathcal{H}^k(M)}_{\text{harmonic}} \tag{11.48}$$

$$\Omega^k(M) = \underbrace{\underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \oplus \underbrace{\operatorname{im}(d) \cap \mathcal{H}^k(M)}_{\text{exact harmonic}}}_{\operatorname{im}(d)\ \text{exact}} \oplus \underbrace{\underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \oplus \underbrace{\mathcal{H}^k_N(M)}_{\text{Neumann harmonic}}}_{\ker(\delta)\cap\ker(\mathbb{n})\ \text{Neumann coclosed}} \tag{11.49}$$

$$\Omega^k(M) = \underbrace{\underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \oplus \underbrace{\operatorname{im}(\delta) \cap \mathcal{H}^k(M)}_{\text{coexact harmonic}}}_{\operatorname{im}(\delta)\ \text{coexact}} \oplus \underbrace{\underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \oplus \underbrace{\mathcal{H}^k_D(M)}_{\text{Dirichlet harmonic}}}_{\ker(d)\cap\ker(\mathbb{t})\ \text{Dirichlet closed}} \tag{11.50}$$

$$\Omega^k(M) = \underbrace{\underbrace{\underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \oplus \underbrace{\operatorname{im}(d) \cap \mathcal{H}^k(M)}_{\text{exact harmonic}}}_{\operatorname{im}(d)\ \text{exact}} \oplus \underbrace{\mathcal{H}^k_N(M)}_{\text{Neumann harmonic}}}_{\ker(d)\ \text{closed}} \oplus \underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \tag{11.51}$$

$$\Omega^k(M) = \underbrace{\underbrace{\underbrace{\operatorname{im}(\delta_N)}_{\text{Neumann coexact}} \oplus \underbrace{\operatorname{im}(\delta) \cap \mathcal{H}^k(M)}_{\text{coexact harmonic}}}_{\operatorname{im}(\delta)\ \text{coexact}} \oplus \underbrace{\mathcal{H}^k_D(M)}_{\text{Dirichlet harmonic}}}_{\ker(\delta)\ \text{coclosed}} \oplus \underbrace{\operatorname{im}(d_D)}_{\text{Dirichlet exact}} \tag{11.52}$$

In Section 11.49 we see that $\mathcal{H}^k_N(M)$ is the space of forms that are Neumann coclosed but not Neumann coexact. The same space appears in Section 11.51 as the forms that are closed but not exact. Similarly, as shown in Section 11.50 and Section 11.52 $\mathcal{H}^k_D(M)$ is the space of forms that

are Dirichlet closed but not Dirichlet exact, or it is the coclosed form but not coexact.

Corollary 11.16 — Dirichlet and Neumann harmonic forms and cohomologies. The space of Neumann (resp. Dirichlet) harmonic forms is the cohomology

$$\mathcal{H}^k_N(M) \cong H^k_{\text{deRham}}(M) := \frac{\ker(d|_{\Omega^k(M)})}{\operatorname{im}(d|_{\Omega^{k-1}(M)})} \tag{11.53}$$

$$\mathcal{H}^k_D(M) \cong H^k_{\text{deRham}}(M, \partial M) := \frac{\ker(d|_{\Omega^k_D(M)})}{\operatorname{im}(d|_{\Omega^{k-1}_D(M)})} \tag{11.54}$$

The space $H^k(M)$ is called the **(absolute) cohomology**. The space $H^k(M, \partial M)$ is called the **relative cohomology** (relative to ∂M). Both are finite dimensional spaces.

Here we take a Glimpse of the so-called **Poincaré–Lefschetz duality**, the generalization of the Poincaré duality on manifolds with boundary. Note that $\star$ maps a Neumann k-form to a Dirichlet $(n-k)$-form and vice versa, and keeps the harmonicity for harmonic form. Therefore,

Theorem 11.17 — Poincaré–Lefschetz duality for the de Rham cohomologies. The Hodge star $\star$ gives the isomorphism

$$\mathcal{H}^k_D(M) \cong \mathcal{H}^{n-k}_N(M). \tag{11.55}$$

In particular,

$$H^k_{\text{deRham}}(M, \partial M) \cong H^{n-k}_{\text{deRham}}(M). \tag{11.56}$$

We will explore more about the relative cohomology in Section 11.6

The Hodge–Morrey–Friedrich decomposition allows us to determine the correct set of boundary conditions and Coulomb gauge conditions for the the potential problems such as "given ω find α such that $d\alpha = \omega$."

We will first talk about extracting each components in the Hodge–Morrey–Friedrich decomposition. These correspond to problems without specific boundary values. Then we will explore problems where a specific Dirichlet data or Neumann data is given.

11.5 Components of the Hodge–Morrey–Friedrich Decomposition

Let $\omega \in \Omega^k(M)$. By the Hodge–Morrey–Fiedrich decomposition there exists $\alpha \in \Omega^{k-1}_D(M)$, $\beta \in \Omega^{k+1}_N(M)$, $\phi \in \Omega^{k-1}(M)$ with $\delta d\phi = 0$, $\psi \in \Omega^{k+1}(M)$ with $d\delta\psi = 0$, $h_N \in \mathcal{H}^k_N(M)$ and $h_D \in \mathcal{H}^k_D(M)$ such that

$$\omega = d\alpha + \delta\beta + d\phi + h_N, \tag{11.57}$$

$$\omega = d\alpha + \delta\beta + \delta\psi + h_D. \tag{11.58}$$

11.5.1 Dirichlet Potential (Pseudoinverse of d_D)

Here we find the Dirichlet potential α for (11.57) or (11.58). Taking δ on either equation yields the equation for α:

$$\begin{cases} \delta d\alpha = \delta\omega & \text{in } M \\ \mathbb{t}\alpha = 0 & \text{on } \partial M \end{cases} \tag{11.59}$$

The solution α is non-unique: Addition by any Dirichlet closed forms to α yields an equally valid $d\alpha \in \mathrm{im}(d_D)$ component for the Hodge–Morrey–Friedrich decomposition of ω. To make (11.59) into a problem that has a unique solution, we choose the α that has the least L^2-norm. This amounts to imposing the additional conditions $\alpha \perp (\ker(d) \cap \ker(\mathbb{t}))$ (which we call the **Coulomb gauge** conditions)

$$\begin{cases} \delta d\alpha = \delta\omega \\ \alpha \perp (\ker(d) \cap \ker(\mathbb{t})) \\ \mathbb{t}\alpha|_{\partial M} = 0 \end{cases} \overset{(11.50)}{\Leftrightarrow} \begin{cases} \delta d\alpha = \delta\omega \\ \alpha \in \mathrm{im}(\delta) \\ \mathbb{t}\alpha|_{\partial M} = 0 \end{cases} \overset{(11.52)}{\Leftrightarrow} \begin{cases} \delta d\alpha = \delta\omega \\ \delta\alpha = 0 \\ \alpha \perp \mathcal{H}_N^{k-1}(M) \\ \mathbb{t}\alpha|_{\partial M} = 0. \end{cases} \tag{11.60}$$

This last equation is equivalent to the following **Poisson problem** with zero Dirichlet boundary condition:

Theorem 11.18 — Poisson problem for the Dirichlet exact part. The potential α under the Coulomb gauge for the Dirichlet exact part $d\alpha$ of ω in (11.57) or (11.57) is found by solving the Poisson problem

$$\begin{cases} -\Delta\alpha = \delta\omega & \text{in } M \\ \mathbb{t}\alpha = 0 & \text{on } \partial M \\ \mathbb{t}\delta\alpha = 0 & \text{on } \partial M \\ \langle\!\langle \alpha, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}_N^{k-1}(M) \end{cases} \tag{11.61}$$

where $-\Delta = \delta d + d\delta$.

Proof. The implication from (11.60) to (11.61) is clear. Conversely, start with $\delta d\alpha + d\delta\alpha = \delta\omega$ and observe that both the right-hand side and the term $\delta d\alpha$ are coexact. Hence $d\delta\alpha$ must be coexact. Using $\mathbb{t}\delta\alpha = 0$ we find $d\delta\alpha$ is Dirichlet closed, which is orthogonal to coexact forms. Therefore $d\delta\alpha = 0$. Again using $\mathbb{t}\delta\alpha = 0$ and Green's identity we have $0 = \langle\!\langle d\delta\alpha, \alpha \rangle\!\rangle = \langle\!\langle \delta\alpha, \delta\alpha \rangle\!\rangle$ so $\delta\alpha = 0$. □

Definition 11.5 — Pseudoinverse of d_D. Let

$$d_D^+ \colon \Omega^k(M) \to \Omega_D^{k-1}(M) \tag{11.62}$$

defined by that $d_D^+\omega := \alpha$ where α is the solution to (11.61). That is, $d_D^+\omega$ is the least-squares solution in $\Omega_D^{k-1}(M)$ to the linear problem $d_D\alpha = \omega$.

The image and kernel of the pseudoinverse d_D^+ are

$$\mathrm{im}(d_D^+) = \mathrm{im}(\delta) \cap \ker(\mathbb{t}), \quad \ker(d_D^+) = \ker(\delta). \tag{11.63}$$

R Note that the space $\mathrm{im}(d_D^+) = \mathrm{im}(\delta) \cap \ker(\mathbb{t}) = \mathrm{im}(\delta) \cap \Omega_D^{k-1}(M)$ is not a *closed subspace* in $\Omega^{k-1}(M)$ under the L^2 inner product structure. What we mean is that when we take twice orthogonal complement

$$\begin{aligned}\left(\mathrm{im}(d_D^+)\right)^\perp &= \ker(d_D) = \ker(d) \cap \ker(\mathbb{t}),\\ \overline{\mathrm{im}(d_D^+)} &= \left(\mathrm{im}(d_D^+)\right)^{\perp\perp} = (\ker(d) \cap \ker(\mathbb{t}))^\perp \overset{(11.50)}{=} \mathrm{im}(\delta)\end{aligned}$$

we obtain a space $\mathrm{im}(\delta)$ that is slightly bigger than what we started with $(\mathrm{im}(\delta) \cap \ker(\mathbb{t}))$. In fact, the implication of this is that we can use $\mathrm{im}(d_D^+)$ to make arbitrarily close approximation (in L^2 distance) to any given element in $\mathrm{im}(\delta)$, even the ones that does not satisfy the Dirichlet boundary condition $(\ker(\mathbb{t}))$. They are however generated by $d_D^+\omega$ where ω has a strong impulse near the boundary. For regular ω, we find $d_D^+\omega$ always satisfy the Dirichlet boundary condition as explicitly described in (11.61). But if we are talking about cohomology, we do observe ${}^{\ker(\delta)}/_{\mathrm{im}(\delta)} = {}^{\ker(d_D^+)}/_{\mathrm{im}(d_D^+)}$.

In the special case where ω is already Dirichlet exact $\omega \in \mathrm{im}(d_D)$ we usually take an alternative Poisson problem. We know a Coulomb gauge α for $d\alpha = 0$ must be coexact. Therefore we take the representation $\alpha = \delta\sigma$ where σ must satisfy $d\delta\sigma = \omega$.

$$\begin{cases} d\delta\sigma = \omega \\ \sigma \in (\ker(\delta))^\perp = \mathrm{im}(d_D) \\ \mathbb{t}\delta\sigma|_{\partial M} = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} d\delta\sigma = \omega \\ d\sigma = 0 \\ \mathbb{t}\sigma|_{\partial M} = 0 \\ \mathbb{t}\delta\sigma|_{\partial M} = 0 \\ \sigma \perp \mathcal{H}_D^k(M) \end{cases} \quad \Leftrightarrow \quad \begin{cases} -\Delta\sigma = \omega \\ \mathbb{t}\sigma|_{\partial M} = 0 \\ \mathbb{t}\delta\sigma|_{\partial M} = 0 \\ \sigma \perp \mathcal{H}_D^k(M). \end{cases} \tag{11.64}$$

To see the last equation implies the second-last equation, using the assumption $\omega \in \mathrm{im}(d_D)$ and the condition $\mathbb{t}\delta\sigma$ observe that $d\delta\sigma + \delta d\sigma = \omega$ implies that $\delta d\sigma \in \mathrm{im}(d_D)$. But $\delta d\sigma$ is also coclosed. Thus by (11.52) we have $\delta d\sigma = 0$. By Green's identity and $\sigma \in \ker(\mathbb{t})$ we have $0 = \langle\!\langle \sigma, \delta d\sigma \rangle\!\rangle = \langle\!\langle d\sigma, d\sigma \rangle\!\rangle$ and thus $d\sigma = 0$.

Theorem 11.19 — Another Poisson problem for the Dirichlet exact part. Suppose $\omega \in \mathrm{im}(d_D) \subset \Omega^k(M)$ is Dirichlet exact.[a] Then the potential $\alpha \in \Omega_D^{k-1}$ for $d\alpha = \omega$ under the Coulomb gauge is given by $\alpha = \delta\sigma$ where σ is the solution to the Poisson equation

$$\begin{cases} -\Delta\sigma = \omega & \text{in } M \\ \mathbb{t}\sigma = 0 & \text{on } \partial M \\ \mathbb{t}\delta\sigma = 0 & \text{on } \partial M \\ \langle\!\langle \sigma, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}_D^k(M). \end{cases} \tag{11.65}$$

[a] That is $d\omega = 0$, $\mathbb{t}\omega|_{\partial M} = 0$ and $\omega \perp \mathcal{H}_D^k$

11.5.2 Neumann Potential (Pseudoinverse of δ_N)

Finding $\beta \in \Omega_N^{k+1}(M)$ in (11.57) or (11.58) is similar to how we find α. We look for β that satisfies $d\delta\beta = d\omega$ and $\mathbb{n}\beta = 0$, together with an additional Coulomb gauge condition $\beta \in (\ker(\delta) \cap \ker(\mathbb{n}))^\perp = \mathrm{im}(d)$. Such least norm solution can be found by solving the following Poisson problem.

Theorem 11.20 — Poisson problem for the Neumann coexact part. The potential β under the Coulomb gauge for the Neumann coexact part $\delta\beta$ of ω in (11.57) or (11.57) is found by solving the Poisson problem

$$\begin{cases} -\Delta\beta = d\omega & \text{in } M \\ \text{tr}\beta = 0 & \text{on } \partial M \\ \text{tr}d\beta = 0 & \text{on } \partial M \\ \langle\!\langle \beta, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}_D^{k+1}(M). \end{cases} \tag{11.66}$$

Definition 11.6 — Pseudoinverse of δ_N. We call the solution β to the Poisson problem (11.66) the pseudoinverse δ_N^+ of δ_N applied on $\omega \in \Omega^k(M)$.

Note that

$$\text{im}(\delta_N^+) = \text{im}(d) \cap \ker(\text{tr}), \quad \ker(\delta_N^+) = \ker(d). \tag{11.67}$$

Theorem 11.21 — Another Poisson problem for the Neumann coexact part. Suppose $\omega \in \text{im}(\delta_N) \subset \Omega^k(M)$ is Neumann coexact.[a] Then the potential $\beta \in \Omega_D^{k+1}$ for $\delta\beta = \omega$ under the Coulomb gauge is given by $\beta = d\sigma$ where σ is the solution to the Poisson equation

$$\begin{cases} -\Delta\sigma = \omega & \text{in } M \\ \text{tr}\sigma = 0 & \text{on } \partial M \\ \text{tr}d\sigma = 0 & \text{on } \partial M \\ \langle\!\langle \sigma, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}_N^k(M). \end{cases} \tag{11.68}$$

[a] That is $\delta\omega = 0$, $\text{tr}\omega|_{\partial M} = 0$ and $\omega \perp \mathcal{H}_N^k$

11.5.3 Exact Component (Pseudoinverse of d)

Here we find the potential $\eta = \alpha + \phi$ for the exact component $d(\alpha + \phi)$ in (11.57). Taking δ on the equation gives

$$\delta d\eta = \delta\omega \tag{11.69}$$

and this time there is no Dirichlet boundary condition anymore. The Coulomb gauge condition for this problem is that $\eta \in (\ker(d))^\perp = \text{im}(\delta_N)$. By (11.49) the Neumann coexact forms are Neumann coclosed and orthogonal to Neumann harmonic. Therefore, the well-posed problem for η is

$$\begin{cases} \delta d\eta = \delta\omega \\ \delta\eta = 0 \\ \eta \perp \mathcal{H}_N^{k-1}(M) \\ \text{tr}\eta|_{\partial M} = 0 \end{cases} \tag{11.70}$$

which is equivalent to the Poisson equation with Neumann boundary conditions:

Theorem 11.22 — Poisson problem for the exact component. The potential $\eta = \alpha + \phi$ under the Coulomb gauge for the exact part of (11.57) is given by the solution of the Poisson equation

$$\begin{cases} -\Delta\eta = \delta\omega & \text{in } M \\ \text{ⓝ}\eta = 0 & \text{on } \partial M \\ \text{ⓝ}d\eta = \text{ⓝ}\omega & \text{on } \partial M \\ \langle\!\langle \eta, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}_N^{k-1}(M) \end{cases} \tag{11.71}$$

Proof. To see how (11.71) implies (11.70) observe that $\delta d\eta + d\delta\eta = \delta\omega$ implies that $d\delta\eta = \delta(d\eta - \omega)$ is Neumann coexact (since $\text{ⓝ}d\eta = \text{ⓝ}\omega$). But $d\delta\eta$ is also closed. Therefore by the decomposition (11.51) we must have $d\delta\eta = 0$. Now using $\eta \in \ker(\text{ⓝ})$ and Green's identity we have $0 = \langle\!\langle d\delta\eta, \eta \rangle\!\rangle = \langle\!\langle \delta\eta, \delta\eta \rangle\!\rangle$, so $\delta\eta = 0$. □

Definition 11.7 — Pseudoinverse of d. Let

$$d^+ : \Omega^k(M) \to \Omega^{k-1}(M) \tag{11.72}$$

be the operator so that $d^+\omega = \eta$ is the solution to (11.71). That is, $d^+\omega$ is the least-squares solution in $\Omega^{k-1}(M)$ to the linear problem $d\eta = \omega$.

Note that

$$\operatorname{im}(d^+) = \operatorname{im}(\delta_N), \quad \ker(d^+) = \ker(\delta) \cap \ker(\text{ⓝ}). \tag{11.73}$$

In the special case where $\omega = d\eta$ is known to be exact, we may recover the Coulomb gauge $\eta = d^+\omega$ by another Poisson equation. Since η must be Neumann coexact, we let $\eta = \delta\sigma$ for some $\sigma \in \Omega_N^k(M)$. The condition for σ is $d\delta\sigma = \omega$. Imposing the Coulomb gauge for σ we have $\sigma \in (\ker(\delta) \cap \ker(\text{ⓝ}))^\perp = \operatorname{im}(d)$:

$$\begin{cases} d\delta\sigma = \omega \\ d\sigma = 0 \\ \text{ⓝ}\sigma|_{\partial M} = 0 \\ \sigma \perp \mathcal{H}_N^k(M) \end{cases} \tag{11.74}$$

This is equivalent to the following Poisson problem.

Theorem 11.23 — Another Poisson problem for the exact part. Suppose $\omega \in \Omega^k(M)$ is known to be exact.[a] Then the potential $\eta \in \Omega^{k-1}(M)$ for $d\eta = \omega$ under the Coulomb gauge is given by $\eta = \delta\sigma$, where σ is the solution to the Poisson problem

$$\begin{cases} -\Delta\sigma = \omega & \text{in } M \\ \text{ⓝ}\sigma = 0 & \text{on } \partial M \\ \text{ⓝ}d\sigma = 0 & \text{on } \partial M \\ \langle\!\langle \sigma, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}_N^k(M) \end{cases} \tag{11.75}$$

[a] That is, $d\omega = 0$ and $\omega \perp \mathcal{H}^k_N(M)$.

Proof. The implication from (11.74) to (11.75) is clear. Conversely, suppose σ satisfies (11.74). Then $d\delta\sigma + \delta d\sigma = \omega$ together with the assumption that ω is exact implies that $\delta d\sigma$ is exact; in particular it is closed. However, using $\mathbb{n} d\sigma = 0$, we know $\delta d\sigma$ Neumann coexact. By the orthogonality (11.51) between Neumann coexactness and closedness we have $\delta d\sigma = 0$. Therefore $0 = \langle\!\langle \sigma, \delta d\sigma \rangle\!\rangle = \langle\!\langle d\sigma, d\sigma \rangle\!\rangle$ (using $\mathbb{n} d\sigma = 0$) and hence $d\sigma = 0$. □

11.5.4 Coexact Component (Pseudoinverse of δ)

In analogy to the exact component, the coexact component $\zeta = \beta + \psi$ for the coexact component $\delta(\beta + \psi)$ in (11.58) is the solution to the following Poisson problem.

Theorem 11.24 — Poisson problem for the coexact component. The potential $\zeta = \beta + \psi$ under the Coulomb gauge for the coexact part of (11.58) is given by the solution of the Poisson equation

$$\begin{cases} -\Delta\zeta = d\zeta & \text{in } M \\ \mathbb{t}\zeta = 0 & \text{on } \partial M \\ \mathbb{t}\delta\zeta = \mathbb{t}\omega & \text{on } \partial M \\ \langle\!\langle \zeta, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}^{k-1}_D(M) \end{cases} \tag{11.76}$$

Definition 11.8 — Pseudoinverse of δ. Let

$$\delta^+ \colon \Omega^k(M) \to \Omega^{k+1}(M) \tag{11.77}$$

be the operator so that $\delta^+\omega = \zeta$ is the solution to (11.76). That is, $\delta^+\omega$ is the least-squares solution in $\Omega^{k+1}(M)$ to the linear problem $\delta\zeta = \omega$.

Note that

$$\operatorname{im}(\delta^+) = \operatorname{im}(d_D), \quad \ker(\delta^+) = \ker(d) \cap \ker(\mathbb{t}). \tag{11.78}$$

Theorem 11.25 — Another Poisson problem for the coexact part. Suppose $\omega \in \Omega^k(M)$ is coexact.[a] Then the potential $\zeta = \delta^+\omega \in \Omega^{k-1}(M)$ for $\delta\zeta = \omega$ under the Coulomb gauge is given by $\zeta = d\sigma$ where σ is the solution to the Poisson problem

$$\begin{cases} -\Delta\sigma = \omega & \text{in } M \\ \mathbb{t}\sigma = 0 & \text{on } \partial M \\ \mathbb{t}\delta\sigma = 0 & \text{on } \partial M \\ \langle\!\langle \sigma, \chi \rangle\!\rangle = 0 & \text{for all } \chi \in \mathcal{H}^k_D(M) \end{cases} \tag{11.79}$$

[a] That is, $\delta\omega = 0$ and $\omega \perp \mathcal{H}^k_D(M)$.

11.5.5 Hodge–Morrey–Friedrich Decomposition in terms of the Pseudoinverses of the Differentials

We summarize the section by expressing each of the components of the Hodge–Morrey–Fiedrich decomposition in terms of d^+, d_D^+, δ^+ and δ_N^+.

Recall (11.57) that

$$\omega = \underbrace{d\alpha}_{\substack{\text{Dirichlet}\\ \text{exact}}} + \underbrace{\delta\beta}_{\substack{\text{Neumann}\\ \text{coexact}}} + \underbrace{d\phi}_{\substack{\text{exact}\\ \text{harmonic}}} + \underbrace{h_N}_{\substack{\text{Neumann}\\ \text{harmonic}}} \tag{11.80}$$

and (11.58) that

$$\omega = \underbrace{d\alpha}_{\substack{\text{Dirichlet}\\ \text{exact}}} + \underbrace{\delta\beta}_{\substack{\text{Neumann}\\ \text{coexact}}} + \underbrace{d\psi}_{\substack{\text{coexact}\\ \text{harmonic}}} + \underbrace{h_D}_{\substack{\text{Dirichlet}\\ \text{harmonic}}} \tag{11.81}$$

Each potential under the Coulomb gauge is given by

$$\alpha = d_D^+\omega, \quad \beta = \delta_N^+\omega \tag{11.82}$$
$$\eta = \alpha + \phi = d^+\omega, \quad \phi = \eta - \alpha = (d^+ - d_D^+)\omega \tag{11.83}$$
$$\zeta = \beta + \psi = \delta^+\omega, \quad \psi = \zeta - \beta = (\delta^+ - \delta_N^+)\omega \tag{11.84}$$
$$h_N = (1 - dd^+ - \delta\delta_N^+)\omega, \quad h_D = (1 - dd_D^+ - \delta\delta^+)\omega \tag{11.85}$$

A component such as $\delta\beta = \delta\delta_N^+\omega$ can also be expreseed as $\delta\beta = d^+d\omega$ since d^+ gives Neumann coexact forms. In fact each component can be expressed using only d, d^+, d_D^+.

Theorem 11.26 The Hodge–Morrey–Friedrich decomposition components can be expressed as follows.

- $dd_D^+ = \delta^+\delta$ extracts the Dirichlet exact component.
- $dd^+ = \delta_N^+\delta$ extracts the exact component.
- $d^+d = \delta\delta_N^+$ extracts the Neumann coexact component.
- $d_D^+d = \delta\delta^+$ extracts the coexact component.
- $(dd^+ - dd_D^+) = (\delta_N^+\delta - \delta^+\delta)$ extracts the exact harmonic component.
- $(d_D^+d - d^+d) = (\delta\delta^+ - \delta\delta_N^+)$ extracts the coexact harmonic component.
- $(1 - d_D^+d) = (1 - \delta\delta^+)$ extracts the Dirichlet closed component.
- $(1 - d^+d) = (1 - \delta\delta_N^+)$ extracts the closed component.
- $(1 - dd^+) = (1 - \delta_N^+\delta)$ extracts the Neumann coclosed component.
- $(1 - dd_D^+) = (1 - \delta^+\delta)$ extracts the coclosed component.
- $(1 - dd^+ - d^+d) = (1 - \delta_N^+\delta - \delta\delta_N^+)$ extracts the Neumann harmonic component.
- $(1 - dd_D^+ - d_D^+d) = (1 - \delta^+\delta - \delta\delta^+)$ extracts the Dirichlet harmonic component.

▪ **Example 11.5 — Pressure projection.** See also Example 10.3 for the boundariless case. Let $\mathcal{M}$ be a domain with boundary. For an incompressible flow the velocity $\mathbf{v}$ must satisfy the divergence-free condition $\operatorname{div}\mathbf{v} = 0$ and the no-through boundary condition $\mathbf{v}\cdot\mathbf{n}|_{\partial\mathcal{M}} = 0$. In terms of 1-form $\eta = \mathbf{v}^\flat$, the conditions are $\delta\eta = 0$ and $\text{tr}\,\eta|_{\partial\mathcal{M}} = 0$. That is, η is *Neumann coclosed*. Let $\widetilde{\mathbf{v}} \in \Gamma(T\mathcal{M})$ ($\widetilde{\eta} := \widetilde{\mathbf{v}}^\flat$) be an arbitrarily given vector field that might not satisfy the above conditions. The Hodge–Morrey–Friedrich decomposition asserts that the orthogonal complement of Neumann coclosed space is the exact space, and thus there exists a unique exact

form dp such that $\widetilde{\eta} - dp = \eta$ is Neumann coclosed. This is accomplished by $\eta := (1 - dd^+)\widetilde{\eta}$. In practice $p = d^+\widetilde{\eta} \in \Omega^0(M)$ is given by solving the Poisson equation (from (11.71) and note that $\text{tr}\, p$ always vanishes for 0-forms)

$$\begin{cases} -\Delta p = \delta\widetilde{\eta} \\ \text{tr}\, dp|_{\partial M} = \text{tr}\,\widetilde{\eta} \\ p \perp \mathcal{H}^0_N(M) \end{cases} \qquad \eta = \widetilde{\eta} - dp. \tag{11.86}$$

Note that $\mathcal{H}^0_N(M)$ consists of constant functions (constant per connected components). Let us assume M is connected, so $\mathcal{H}^0_N(M)$ is spanned by just one constant function. In classical PDE terms

$$\begin{cases} -\Delta p = \nabla \cdot \widetilde{\mathbf{v}} & \text{in } M \\ \partial p/\partial \mathbf{n} = \widetilde{\mathbf{v}} \cdot \mathbf{n} & \text{on } \partial M \\ \int_M p = 0. \end{cases} \qquad \mathbf{v} = \widetilde{\mathbf{v}} - \nabla p. \tag{11.87}$$

∎

11.6 Relative Homology and Cohomology

Now we move on to the discrete setting. At the same time we will talk about the **relative homology**.

In the case of closed manifolds, the k-th homology is the space of all closed k-dimensional chains that are not the boundary of any $(k-1)$-dimensional chain. This notion is still the same for manifolds with boundary.

What is new is that when there is a boundary, we have another notion of closedness alongside with the classical one: **relative closedness**. We say a chain Γ is relatively closed (relative to ∂M) if $\partial\Gamma \subset \partial M$. That is, Γ may have a boundary, but this boundary must be contained in ∂M. The k-th **relative homology**, denoted by $H_k(M, \partial M)$, is the space of all relatively closed k chains that do not coincide with the boundary of any $(k-1)$ chain in the region $M \setminus \partial M$.

It turns out that the relative (co)homology is isomorphic to the Dirichlet harmonic forms:

$$H_k(M, \partial M) \cong H^k(M, \partial M) \cong H^k_{\text{deRham}}(M, \partial M) \cong \mathcal{H}^k_D(M). \tag{11.88}$$

The classical version of closedness (without taking relativity to ∂M) is the absolute (co)homology, and it is isomorphic to the Neumann harmonic forms:

$$H_k(M) \cong H^k(M) \cong H^k_{\text{deRham}}(M) \cong \mathcal{H}^k_N(M). \tag{11.89}$$

By Theorem 11.17 the absolute and relative (co)homologies are related by the Poincaré–Lefschetz duality:

$$H_k(M, \partial M) \cong H_{n-k}(M). \tag{11.90}$$

11.6.1 Relative Chain and Cochain

Let us talk about the discrete analog of the de Rham complex with boundary condition (11.33). Let $M = (M_0, \dots, M_n)$ be a discrete mesh. The boundary of mesh is first labeled in the vertices $(\partial M)_0 \subset M_0$, *i.e.* a set of boundary vertices. Then we mark each edge e a boundary edge if both

incident vertices of the edge are boundary vertices. That is, a boundary edge lies completely in the boundary of the domain. Similarly, a face is marked as a boundary face if all incident edges are boundary edges, and so on. In terms of chains, define $C_0(\partial M) \coloneqq \operatorname{span}((\partial M)_0)$; and then for $k = 1, \dots, n$

$$(\partial M)_k \coloneqq \{e \in M_k \mid \partial e \subset C_{k-1}(\partial M)\} \quad \text{(set of boundary } k\text{-cells)}, \tag{11.91}$$

$$C_k(\partial M) \coloneqq \operatorname{span}((\partial M)_k) \quad \text{(boundary } k\text{-chains are linear combinations} \tag{11.92}$$

$$\text{of boundary } k\text{-cells)}. \tag{11.93}$$

Hence we have the chain complex for the boundary mesh:

$$\cdots \xrightarrow{\partial_\partial} C_{k+1}(\partial M) \xrightarrow{\partial_\partial} C_k(\partial M) \xrightarrow{\partial_\partial} C_{k-1}(\partial M) \xrightarrow{\partial_\partial} \cdots \tag{11.94}$$

Taking their dual spaces we also have the cochain (discrete differential form) complex connected by the coboundary operator (discrete exterior derivative)

$$C^k(\partial M) = (C_k(\partial M))^\star \cong C_k(\partial M) \cong \mathbb{R}^{|(\partial M)_k|}, \tag{11.95}$$

$$\cdots \xrightarrow{d_\partial = \partial_\partial^\mathsf{T}} C^{k-1}(\partial M) \xrightarrow{d_\partial = \partial_\partial^\mathsf{T}} C^k(\partial M) \xrightarrow{d_\partial = \partial_\partial^\mathsf{T}} C^{k-1}(\partial M) \xrightarrow{d_\partial = \partial_\partial^\mathsf{T}} \cdots \tag{11.96}$$

In practice, suppose ∂ (resp. d) is the (co)boundary operator on $C_k(M)$ (resp. $C^k(M)$) given as a matrix, we obtain ∂_∂ (resp. d_∂) by slicing the submatrix corresponding to the boundary cells:

$$\partial_\partial = \partial_{\mathrm{bdy}_{k-1}, \mathrm{bdy}_k}, \quad d_\partial = d_{\mathrm{bdy}_{k+1}, \mathrm{bdy}_k} \tag{11.97}$$

where $\mathrm{bdy}_k \subset \{0, \dots, |M_k| - 1\}$ is the index set for the boundary k-cells.

So now we have the original chain complex on M and a chain complex on ∂M. Now consider the remaining cells that are not in the boundary

$$(M^\circ)_k \coloneqq M_k \setminus (\partial M)_k \quad \text{(set of interior } k\text{-cells)} \tag{11.98}$$

which spans the so-called **relative k-chains** (or **k-chains relative to boundary**)

$$C_k(M, \partial M) \coloneqq \operatorname{span}((M^\circ)_k). \tag{11.99}$$

Equivalently,

$$C_k(M, \partial M) \cong \frac{C_k(M)}{C_k(\partial M)} \cong \mathbb{R}^{|M_k| - |(\partial M)_k|} = \mathbb{R}^{|(M^\circ)_k|}. \tag{11.100}$$

It consists of the chains in $C_k(M)$ with the condition that the coefficients in front of any boundary cell is zero.

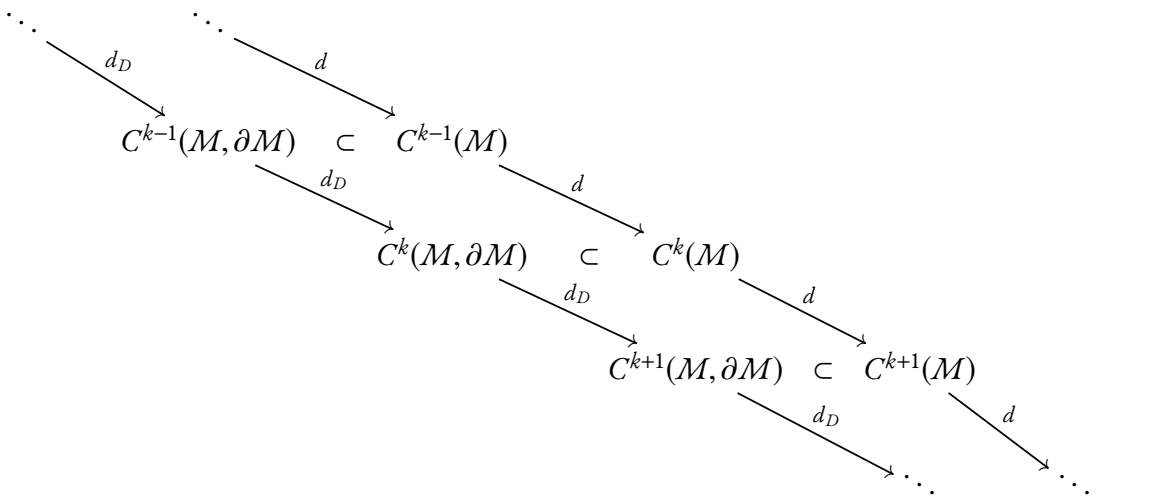
Similarly the space of **relative k-cochains** is

$$C^k(M, \partial M) \cong \frac{C^k(M)}{C^k(\partial M)} \cong \mathbb{R}^{|(M^\circ)_k|}, \tag{11.101}$$

which is the collection of discrete differential forms whose evaluation at any boundary cell is zero. That is, it is the Dirichlet subspace in the space of k-forms:

$$C^k(M, \partial M) \text{ is the discrete version of } \Omega^k_D(M). \tag{11.102}$$

As the discrete analog of (11.33) we have relative cochain complex



(11.103)

where d_D denotes the d operator on the Dirichlet subspace.

In practice, suppose d is the discrete exterior derivative on $C^k(M)$ given as a matrix, we get d_D by slicing the submatrix corresponding to the interior cells:

$$\partial_{\text{rel}} = \partial_{\text{int}_{k-1},\text{int}_k}, \quad d_D = d_{\text{int}_{k+1},\text{int}_k}, \quad d_{D,k} = \partial^{\mathsf{T}}_{\text{rel},k+1}, \tag{11.104}$$

where $\text{int}_k = (\{0, \dots, |M_k| - 1\} \setminus \text{bdy}_k) \subset \{0, \dots, |M_k| - 1\}$ is the index set for the interior k-cells.

If we had labeled the cells in such a way that the interior and boundary cells were grouped together, then d would take the block matrix form

$$d = \left[\begin{array}{c|c} d_D & d_{\text{int}_{k+1},\text{bdy}_k} \\ \hline 0 & d_\partial \end{array}\right]. \tag{11.105}$$

11.6.2 Dual Mesh

For a mesh $M = (M_0, \dots, M_n)$ with boundary, the dual mesh

$$(M^\star)_k = \{e^\star \mid e \in M_{n-k}\} \tag{11.106}$$

contains a few cells $e^\star$ associated with the primal boundary cells $e \in (\partial M)_{n-k}$. These cells do not have a properly defined boundary $\partial^\star e^\star$, because some boundary cells of $e^\star$ are not contained in the mesh $M^\star$. Instead, we understand $\partial^\star e^\star$ as $\partial^\star_{\text{rel}} e^\star$ to accept the fact that there are missing boundary cells. That is, we imagine that $M^\star$ is a bigger mesh that also include some dummy boundary cells "$\partial M^\star$," with an additional constraint that all chains and cochains vanish on the dummy $\partial M^\star$.

Another way to take the dual mesh is the following. Take the dual mesh consisting of only the dual cells associated with the interior mesh:

$$(M^{\circ\star})_k = \{e^\star \mid e \in (M^\circ)_{n-k}\}. \tag{11.107}$$

All boundary pieces of every cells in $(M^{\circ\star})_k$ are contained in the mesh. So the boundary operator $\partial^\star$ is understood in the traditional sense.

The chains and cochains on the dual meshes $M^\star, M^{\circ\star}$ are respectively

$$C_k(M^\star, \partial M^\star) = \left\{ \sum_{e \in M_{n-k}} c_e\, e^\star \,\middle|\, c_e \in \mathbb{R} \right\} \cong C_{n-k}(M) \tag{11.108}$$

$$C^k(M^\star, \partial M^\star) = \left\{ \sum_{e \in M_{n-k}} \alpha_e\, \mathbb{1}_{e^\star} \,\middle|\, \alpha_e \in \mathbb{R} \right\} \cong C^{n-k}(M) \tag{11.109}$$

$$C_k(M^{\circ\star}) = \left\{ \sum_{e \in (M^\circ)_{n-k}} c_e\, e^\star \,\middle|\, c_e \in \mathbb{R} \right\} \cong C_{n-k}(M, \partial M) \tag{11.110}$$

$$C^k(M^{\circ\star}) = \left\{ \sum_{e \in (M^\circ)_{n-k}} \alpha_e\, \mathbb{1}_{e^\star} \,\middle|\, \alpha_e \in \mathbb{R} \right\} \cong C^{n-k}(M, \partial M). \tag{11.111}$$

The incidence matrix (boundary operator) on the dual mesh is related to the relative boundary operator:

$$\partial^\star_{\mathrm{rel},k} \colon C_k(M^\star, \partial M^\star) \to C_{k-1}(M^\star, \partial M^\star), \qquad \partial^\star_{\mathrm{rel},k} = (-1)^{n-k+1} \partial^\intercal_{\mathrm{rel},n-k+1} \tag{11.112}$$

$$\partial^{\circ\star}_k \colon C_k(M^{\circ\star}) \to C_{k-1}(M^{\circ\star}), \qquad \partial^{\circ\star}_k = (-1)^{n-k+1} \partial^\intercal_{\mathrm{rel},n-k+1} \tag{11.113}$$

and similarly for the d operator

$$d^\star_{D,k} \colon C^k(M^\star, \partial M^\star) \to C^{k+1}(M^\star, \partial M^\star), \qquad d^\star_{D,k} = (-1)^{n-k-1} d^\intercal_{n-k} \tag{11.114}$$

$$d^{\circ\star}_k \colon C^k(M^{\circ\star}) \to C^{k+1}(M^{\circ\star}), \qquad d^{\circ\star}_k = (-1)^{n-k-1} d^\intercal_{D,n-k}. \tag{11.115}$$

Just like in the discrete exterior calculus for closed surfaces, one provides a discrete Hodge star (*e.g.* a diagonal Hodge star)

$$\star_k \colon C^k(M) \to C^{n-k}(M^\star, \partial M^\star), \tag{11.116}$$

$$\star_k \colon C^k(M, \partial M) \to C^{n-k}(M^{\circ\star}). \tag{11.117}$$

11.6.3 Discrete Exterior Calculus with Boundary

Now, we have all the ingredient to translate the smooth Hodge–Morrey–Friedrich theory to the discrete setting. The operator $\mathbb{t}$ is taken as $j^*_{\partial M}$, which is to take the part in the cochain $C^k(M)$ associated to the cells in the boundary:

$$\mathbb{t} \colon C^k(M) \to C^k(\partial M), \quad (\mathbb{t}\alpha)(e) = \alpha(e) \text{ for each } e \in (\partial M)_k. \tag{11.118}$$

The operator $\mathbb{n}$ on the dual mesh is defined as the dual of $\mathbb{t}$

$$\mathbb{n} \colon C^k(M^\star, \partial M^\star) \to C^k\big((\partial M)^\star\big), \quad \mathbb{n} = \star_{n-k} \mathbb{t}\, \star^{-1}_{n-k}. \tag{11.119}$$

The operator $\sqcap$ on the primal mesh would require the virtual boundary cells $\partial M^\star$ (do not confuse it with the dual mesh $(\partial M)^\star$ of ∂M)

$$\sqcap\colon C^k(M) \to C^k(\partial M^\star) \tag{11.120}$$

▪ **Example 11.6 — Poisson equation with prescribed normal derivative.** Suppose we want to solve for $u \in \Omega^0(M)$

$$\begin{cases} -\Delta u = f & \text{in } M \\ \sqcap du = g & \text{on } \partial M \end{cases} \tag{11.121}$$

We let u be a primal 0-form. ▪

IV Appendix

A. Quaternions

Quaternions form a number system that represents 3D rotations. It is similar to how complex numbers $\mathbb{C}$ represent 2D rotations. Before diving in quaternion algebra, let us remind ourselves how complex numbers represent 2D rotations.

A.1 Complex Numbers

The rotation in the 2D vector space goes as follows. The counterclockwise rotation R^θ by an angle θ about the origin applied on an $\mathbb{R}^2$ vector $(x, y) \in \mathbb{R}^2$ is given by

$$R^\theta(x, y) = \big(\cos(\theta)x - \sin(\theta)y, \sin(\theta)x + \cos(\theta)y\big). \tag{A.1}$$

Complex numbers $\mathbb{C}$ are $\mathbb{R}^2$ vectors equipped with a multiplication

$$\begin{aligned} &(\cdot)\cdot(\cdot)\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \\ &(a, b)\cdot(x, y) := (xa - yb, xb + ya). \end{aligned} \tag{A.2}$$

This product of two $\mathbb{R}^2$ vectors encapsulates the 2D rotation (A.1) as

$$R^\theta(x, y) = (\cos\theta, \sin\theta)\cdot(x, y).$$

A vector space equipped with a multiplication is called an **algebra**. A multiplication is any bilinear form, that is, it satisfies $\mathbf{u}\cdot(c_1\mathbf{v}_1 + c_2\mathbf{v}_2) = c_1\mathbf{u}\cdot\mathbf{v}_1 + c_2\mathbf{u}\cdot\mathbf{v}_2$ and $(c_1\mathbf{u}_1 + c_2\mathbf{u}_2)\cdot\mathbf{v} = c_1\mathbf{u}_1\cdot\mathbf{v} + c_2\mathbf{u}_2\cdot\mathbf{v}$. Indeed, the multiplication (A.2) is a linear expression of (a, b) when (x, y) is fixed; and is linear in x and y when a and b are fixed. The multiplication (A.2) makes $\mathbb{R}^2$ an algebra.

Due to linearity, we can describe the multiplication just for the basis vectors. Call $(1, 0) = \mathbb{1}$ and $(0, 1) = \mathbb{i}$:

$$\mathbb{1}\cdot\mathbb{1} = \mathbb{1}, \quad \mathbb{1}\cdot\mathbb{i} = \mathbb{i}\cdot\mathbb{1} = \mathbb{i}, \quad \mathbb{i}\cdot\mathbb{i} = \mathbb{i}^2 = -\mathbb{1}. \tag{A.3}$$

Figure A.1 William Rowan Hamilton and his wife Helen Marie Bayly carving the formula for his quaternions into Brougham Bridge (now Broom Bridge) in Cabra, Dublin, Ireland, on October 16th, 1843. Left: Daniel Doyle's sand sculpture. Right: plaque on Broom Bridge.

In the end, one writes $\mathbb{C} := \{x + \mathbb{i}y \,|\, x, y \in \mathbb{R}\}$. The basis vector $\mathbb{1}$ is suppressed because it behaves exactly like the real number 1.

Then, one checks that the complex algebra

- has **division**—for any $z_1, z_2 \in \mathbb{C}$, $z_2 \neq 0$, there exists exactly one element $u \in \mathbb{C}$ so that $z_1 = u \cdot z_2$ and exactly one element $v \in \mathbb{C}$ such that $z_1 = z_2 \cdot v$;
- is **associative**—$z_1 \cdot (z_2 \cdot z_3) = (z_1 \cdot z_2) \cdot z_3$;
- is **commutative**—$z_1 z_2 = z_2 z_1$.

An algebra that has division is called a **division algebra**. An associative division algebra is also called a **division ring** or a **skew field**. A commutative division ring is a **field**.

Besides these familiar properties, the complex algebra

- has a **norm**—there is a quadratic form $|\cdot|^2 \colon \mathbb{C} \to \mathbb{R}_{\geq 0}$ so that $|z_1 \cdot z_2|^2 = |z_1|^2 |z_2|^2$.

Using the norm and division, we have **complex conjugate** $\bar{z} := z^{-1}|z|^2$.

A.1.1 Can you Multiply Triples?

Though complex numbers have been used since the 16th century, it was in the late 18th century that mathematicians such as Leonhard Euler discovered the link between complex numbers and the geometry 2D rotations. The above algebraic approach of assigning a multiplication (A.2) between the then called *couplets* was due to William Rowan Hamilton in 1833. That was also the time when there was a great excitement in complex analysis. The natural question Hamilton asked was, can a similar algebraic structure (a multiplication) be equipped by $\mathbb{R}^3$ so that it represents 3D rotations.

A famous story goes as follows. Hamilton worked on this "Theory of Triplets" unsuccessfully for 10 years. According to Hamilton, every morning, both his 9- and 10-year-old sons asked him "Well, Papa, can you multiply triples?" "No, I can only add and subtract them" William Rowan Hamilton always replied with a sad shake of the head. On October 16th, 1843, Hamilton and his wife took a walk along the Royal Canal in Dublin. While they walked across Brougham Bridge (now Broom Bridge), a solution suddenly occurred to him. While he could not multiply triplets, he saw a way to do so for quadruples. He then carved the basic rules for multiplication into the bridge (Figure A.1):

$$\mathbb{i}^2 = \mathbb{j}^2 = \mathbb{k}^2 = \mathbb{i}\mathbb{j}\mathbb{k} = -\mathbb{1}. \tag{A.4}$$

Hamilton called a quadruple with these rules of multiplication a *quaternion*.

A.2 Quaternion Algebra

Quaternions $\mathbb{H}$ are $\mathbb{R}^4$ vectors equipped with a multiplication. We write $\mathbb{R}^4$ as the direct sum of a scalar and a 3D vector $\mathbb{R}^4 = \mathbb{R} \oplus \mathbb{R}^3$. That is,

$$(\alpha, v_1, v_2, v_3) \in \mathbb{R}^4 \quad \text{is identified as} \quad \alpha + (v_1, v_2, v_3) = \alpha + \mathbf{v} \in \mathbb{R} \oplus \mathbb{R}^3.$$

We call the scalar part the real part of a quaternion, and the 3D vector part the imaginary part.

The multiplication always boils down to

$$(\alpha + \mathbf{v})(\beta + \mathbf{w}) = \alpha\beta + \alpha\mathbf{w} + \beta\mathbf{v} + \mathbf{v}\mathbf{w}$$

where the only nontrivial part is the multiplication between two $\mathbb{R}^3$ vector. Define

$$\mathbf{v}\mathbf{w} := -\langle \mathbf{v}, \mathbf{w} \rangle + \mathbf{v} \times \mathbf{w} \quad \text{for } \mathbf{v}, \mathbf{w} \in \mathbb{R}^3 \tag{A.5}$$

using the Euclidean scalar product $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + v_2 w_2 + v_3 w_3$ and the vector cross product $\mathbf{v} \times \mathbf{w} = (v_2 w_3 - v_3 w_2, v_3 w_1 - v_1 w_3, v_1 w_2 - v_2 w_1)$. In the end, one writes

$$\mathbb{H} = \left\{ \alpha + v_1 \mathbb{i} + v_2 \mathbb{j} + v_3 \mathbb{k} \,\middle|\, \alpha, v_1, v_2, v_3 \in \mathbb{R} \right\}$$

with the multiplication rules on the basis vectors:

$$\begin{aligned}
&\mathbb{i}^2 = -\mathbb{1}, \quad \mathbb{j}^2 = -\mathbb{1}, \quad \mathbb{k}^2 = -\mathbb{1} \\
&\mathbb{i}\mathbb{j} = -\mathbb{j}\mathbb{i} = \mathbb{k} \\
&\mathbb{j}\mathbb{k} = -\mathbb{k}\mathbb{j} = \mathbb{i} \\
&\mathbb{k}\mathbb{i} = -\mathbb{i}\mathbb{k} = \mathbb{j}.
\end{aligned}$$

Note particularly that the product does not commute, *i.e.* for $q_1, q_2 \in \mathbb{H}$, in general $q_1 q_2 \neq q_2 q_1$. Define for $q = \alpha + \mathbf{v} \in \mathbb{H}$,

$$\mathrm{Re}(q) := \alpha \in \mathbb{R}, \quad \mathrm{Im}(q) := \mathbf{v} \in \mathbb{R}^3.$$

A.2.1 Algebraic Structure

Here we show that the quaternion algebra is a normed associative division algebra, *i.e.* a normed skew field.

Theorem A.1 — Normed algebra. For $p, q \in \mathbb{H}$, one has $|pq|^2 = |p|^2 |q|^2$. Here $|\alpha + \mathbf{v}|^2$ is defined by $\alpha^2 + |\mathbf{v}|^2$.

Proof. Let $p = \alpha + \mathbf{v}$ and $q = \beta + \mathbf{w}$. We have $pq = (\alpha + \mathbf{v})(\beta + \mathbf{w}) = \alpha\beta - \langle \mathbf{v}, \mathbf{w} \rangle + \alpha\mathbf{w} + \beta\mathbf{v} + \mathbf{v} \times \mathbf{w}$. Thus

$$\begin{aligned}
|pq|^2 &= \alpha^2\beta^2 - 2\alpha\beta\langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle^2 \\
&\quad + \alpha^2 |\mathbf{w}|^2 + \beta^2 |\mathbf{v}|^2 + |\mathbf{v}|^2 |\mathbf{w}|^2 - \langle \mathbf{v}, \mathbf{w} \rangle^2 + 2\alpha\beta\langle \mathbf{w}, \mathbf{v} \rangle \\
&= (\alpha^2 + |\mathbf{v}|^2)(\beta^2 + |\mathbf{w}|^2) = |p|^2 |q|^2.
\end{aligned}$$

□

R Historically, $|pq|^2 = |p|^2|q|^2$ of quaternions was already written in a letter by Euler sent to Goldbach, May 4, 1748. It is called the **Euler's four-square identity**

$$\begin{aligned}(a_1^2+a_2^2+a_3^2+a_4^2)(b_1^2+b_2^2+b_3^2+b_4^2) = \quad & (a_1b_1-a_2b_2-a_3b_3-a_4b_4)^2 \\ & + (a_1b_2+a_2b_1+a_3b_4-a_4b_3)^2 \\ & + (a_1b_3-a_2b_4+a_3b_1+a_4b_2)^2 \\ & + (a_1b_4+a_2b_3-a_3b_2+a_4b_1)^2.\end{aligned}$$

R Adolf Hurwitz (1859–1919) showed that formula $(\sum_{i=1}^n a_i^2)(\sum_{j=1}^n b_j^2) = \sum_{k=1}^n c_k$ with $\mathbf{c}$ bilinear in $\mathbf{a}$ and $\mathbf{b}$ (similar to Euler's four-square identity for n tuples of real numbers) can hold only for $n = 1, 2, 4, 8$. The **Hurwitz Theorem** states that the only normed division algebra for $\mathbb{R}^n$ are the *real* ($n = 1$), *complex* ($n = 2$), *quaternion* ($n = 4$), and the *octonion* ($n = 8$) number systems.

Theorem A.2 — Associativity. The quaternionic multiplication is associative. That is, for $p, q, r \in \mathbb{H}$, one has $(pq)r = p(qr)$.

Proof. Define 2×2 complex matrices

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad X = \begin{bmatrix} \mathbb{i} & 0 \\ 0 & -\mathbb{i} \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & -\mathbb{i} \\ -\mathbb{i} & 0 \end{bmatrix}.$$

In physics, these matrices are usually written as $X = \mathbb{i}\sigma_z, Y = -\mathbb{i}\sigma_y, Z = -\mathbb{i}\sigma_x$ where σ's are the **Pauli matrices**. Now, one checks that

$$\begin{aligned} &I^2 = I, \quad IX = XI = X, \quad IY = YI = Y, \quad IZ = ZI = Z \\ &XY = -YX = Z, \quad YZ = -ZY = X, \quad ZX = -XZ = Y. \end{aligned}$$

This says that $\operatorname{span}\{I, X, Y, Z\}$ is a closed subalgebra of $\mathbb{C}^{2\times 2}$, and

$$\begin{aligned} \mathbb{H} &\to \operatorname{span}\{I, X, Y, Z\} \subset \mathbb{C}^{2\times 2} \\ \alpha + v_1\mathbb{i} + v_2\mathbb{j} + v_3\mathbb{k} &\mapsto \alpha I + v_1 X + v_2 Y + v_3 Z \end{aligned}$$

is an algebra isomorphism. Since matrix multiplication is associative, we have that quaternionic multiplication is associative. □

Theorem A.3 — Division algebra. The quaternion algebra is a division algebra. Since it is an associative algebra, being a division algebra means that for each $q \in \mathbb{H}$, $q \neq 0$, there exists $q^{-1} \in \mathbb{H}$ such that $qq^{-1} = q^{-1}q = -\mathbb{1}$.

Proof. For $q = \alpha + \mathbf{v}$, let $q^{-1} = {}^{(\alpha-\mathbf{v})}/_{|q|^2}$. It is easy to verify that $(\alpha+\mathbf{v})(\alpha-\mathbf{v}) = (\alpha-\mathbf{v})(\alpha+\mathbf{v}) = \alpha^2 + |\mathbf{v}|^2 = |q|^2$. Therefore $qq^{-1} = q^{-1}q = -\mathbb{1}$. □

Definition A.1 — Quaternion conjugate. Define for each quaternion $q = \alpha + \mathbf{v}$ the quaternion conjugate $\bar{q} := \alpha - \mathbf{v}$. That is, $\bar{q} = q^{-1}|q|^2$ if $q \neq 0$, and $\bar{0} = 0$.

Theorem A.4 For $p, q \in \mathbb{H}$, one has $\overline{pq} = \bar{q}\bar{p}$.

Proof. Use $\bar{q} = q^{-1}|q|^2$ and the general fact that inverses reverse the multiplication order $(pq)^{-1} = q^{-1}p^{-1}$. □

A.2.2 Quaternion Exponentials

Theorem A.5 Let $\mathbf{v} \in \mathbb{R}^3$ with $|\mathbf{v}| = 1$. Let $t \in \mathbb{R}$. Then

$$e^{t\mathbf{v}} = \cos(t) + \sin(t)\mathbf{v}. \tag{A.6}$$

Proof. Here $e^{t\mathbf{v}}$ takes the definition through power series $e^{t\mathbf{v}} = \sum_{n=0}^{\infty} \frac{1}{n!}(t\mathbf{v})^n$. (It is an absolutely convergent power series.) The terms with even powers contribute the real part

$$\sum_{m=0}^{\infty} \frac{1}{(2m)!}(t\mathbf{v})^{2m} = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!}t^{2m} = \cos(t),$$

whereas the odd-powered terms are collected into the imaginary part

$$\sum_{m=0}^{\infty} \frac{1}{(2m+1)!}(t\mathbf{v})^{2m+1} = \sum_{m=0}^{\infty} \frac{(-1)^m\mathbf{v}}{(2m+1)!}t^{2m+1} = \sin(t)\mathbf{v}.$$

□

R In general $e^{t\mathbf{v}}e^{s\mathbf{w}} \neq e^{t\mathbf{v}+s\mathbf{w}}$. The formula $e^{t\mathbf{v}}e^{s\mathbf{w}} = e^{t\mathbf{v}+s\mathbf{w}}$ holds only when the two exponents commute $\mathbf{vw} = \mathbf{wv}$; this happens only if $\mathbf{v}, \mathbf{w}$ are parallel.

Theorem A.6 Let $\mathbf{v} \in \mathbb{R}^3$ with $|\mathbf{v}| = 1$. Let $t \in \mathbb{R}$, $\alpha \in \mathbb{R}$. Then

$$e^{\alpha+t\mathbf{v}} = e^{\alpha}(\cos(t) + \sin(t)\mathbf{v}). \tag{A.7}$$

Exercise A.1 Show Theorem A.6 using the power series. ∎

A.3 Rotations in 3D

Here we explain how quaternions represent 3D linear rotations. A general 3D rotation takes the form of a rotation about an axis through the origin. For $\mathbf{v} \in \mathbb{R}^3$, $|\mathbf{v}| = 1$, let

$$R^{\theta}_{\mathbf{v}} \colon \mathbb{R}^3 \to \mathbb{R}^3 \quad \text{denote the rotation by angle } \theta \text{ about } \mathbf{v}. \tag{A.8}$$

Explicitly, for each $\mathbf{y} \in \mathbb{R}^3$, $R^{\theta}_{\mathbf{v}}\mathbf{y}$ is given by the following **Rodrigues formula**. Decompose $\mathbf{y}$ into the $\mathbf{v}$ component and its orthogonal complement $\mathbf{y} = \langle \mathbf{y}, \mathbf{v}\rangle\mathbf{v} + \underbrace{(\mathbf{y} - \langle \mathbf{y}, \mathbf{v}\rangle\mathbf{v})}_{=:\mathbf{w}}$. Then apply

the rotation on this decomposition, which leaves the $\mathbf{v}$ component invariant, and rotates $\mathbf{w}$ like a 2D rotation in $\mathbf{v}^\perp$:

$$R^\theta_{\mathbf{v}}\mathbf{y} = \langle \mathbf{y}, \mathbf{v}\rangle \mathbf{v} + \cos(\theta)\mathbf{w} + \sin(\theta)(\mathbf{v} \times \mathbf{w}). \tag{A.9}$$

Theorem A.7 Suppose $\mathbf{v} \in \mathbb{R}^3$, $|\mathbf{v}| = 1$ and $\theta \in \mathbb{R}$. Consider

$$q = e^{(\theta/2)\mathbf{v}}, \quad \text{that is,} \quad q = \cos\left(\tfrac{\theta}{2}\right) + \sin\left(\tfrac{\theta}{2}\right)\mathbf{v}.$$

Then for each imaginary quaternion $\mathbf{y} \in \mathbb{R}^3$, the quaternion $q\mathbf{y}\bar{q}$ is always imaginary

$$q\mathbf{y}\bar{q} \in \mathbb{R}^3.$$

The map $\mathbf{y} \mapsto q\mathbf{y}\bar{q}$, as a linear map from $\mathbb{R}^3$ to $\mathbb{R}^3$, is a rotation around $\mathbf{v}$ by the angle θ. That is,

$$R^\theta_{\mathbf{v}}\mathbf{y} = q\mathbf{y}\bar{q} = e^{(\theta/2)\mathbf{v}}\mathbf{y}e^{-(\theta/2)\mathbf{v}}. \tag{A.10}$$

Proof. In this proof, we first show that $q\mathbf{y}\bar{q} \in \mathbb{R}^3$, and then we show that $\mathbf{y} \mapsto q\mathbf{y}\bar{q}$ is a rotation. Finally we show that it is the rotation around $\mathbf{v}$ with angle θ.

Observe that $p \in \mathbb{R}^3$ if and only if $\bar{p} = -p$. So, to show that $q\mathbf{y}\bar{q} \in \mathbb{R}^3$ we have to check that $\overline{q\mathbf{y}\bar{q}} = -q\mathbf{y}\bar{q}$:

$$\overline{q\mathbf{y}\bar{q}} = \bar{\bar{q}}\bar{\mathbf{y}}\bar{q} = q\bar{\mathbf{y}}\bar{q} = -q\mathbf{y}\bar{q}.$$

Hence $q\mathbf{y}\bar{q} \in \mathbb{R}^3$.

The map $F\colon \mathbb{R}^3 \to \mathbb{R}^3$, $F(\mathbf{y}) := q\mathbf{y}\bar{q}$, is linear in $\mathbf{y}$. And it always preserve the norm

$$|F(\mathbf{y})| = |q\mathbf{y}\bar{q}| = |q||\mathbf{y}||\bar{q}| = |\mathbf{y}|.$$

So $F\colon \mathbb{R}^3 \to \mathbb{R}^3$ is an orthogonal linear transformation, which can only be rotation (the case $\det(F) = 1$) or mirrored-rotation (the case $\det(F) = -1$). All we have to check is $\det(F) = 1$. Observe that

$$\det\left(\mathbf{y} \mapsto \left(\cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\mathbf{v}\right)\mathbf{y}\overline{\left(\cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\mathbf{v}\right)}\right) \in \{+1, -1\}$$

varies continuously with θ, and equals to 1 for $\theta = 0$. This ensures $\det(F) = 1$ for all θ.

Finally we show that F is a rotation about $\mathbf{v}$ with angle θ. By multiplying out, one can check that $F(\mathbf{v}) = \mathbf{v}$, which says that it is $\mathbf{v}$ is the rotation axis. All that is left to show is that F is a rotates with angle θ on the plane $\mathbf{v}^\perp$. Now, for $\mathbf{w} \in \mathbf{v}^\perp$, *i.e.* $\langle \mathbf{w}, \mathbf{v}\rangle = 0$, we have $\mathbf{wv} = \mathbf{w} \times \mathbf{v}$, and therefore $\mathbf{wv} = -\mathbf{vw}$. So for $\mathbf{w} \in \mathbf{v}^\perp$ we have

$$\begin{aligned} q\mathbf{w}\bar{q} &= \left(\cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\mathbf{v}\right)\mathbf{w}\left(\cos\tfrac{\theta}{2} - \sin\tfrac{\theta}{2}\mathbf{v}\right) \\ &= \left(\cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\mathbf{v}\right)\left(\cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\mathbf{v}\right)\mathbf{w} \\ &= (\cos\theta + \sin\theta\mathbf{v})\,\mathbf{w} \\ &= \cos(\theta)\mathbf{w} + \sin(\theta)(\mathbf{v} \times \mathbf{w}). \end{aligned}$$

Viewed in the orthogonal basis $(\mathbf{w}, \mathbf{v} \times \mathbf{w}, \mathbf{v})$, this vector $q\mathbf{w}\bar{q}$ is indeed rotated with angle θ about $\mathbf{v}$. □

Corollary A.8 For every $q \in \mathbb{H}, q \neq 0$, the map

$$\begin{aligned} \mathbb{R}^3 &\to \mathbb{R}^3 \\ \mathbf{y} &\mapsto q\mathbf{y}q^{-1} \end{aligned}$$

is a rotation.

Proof. Recall that $q^{-1} = \bar{q}/|q|^2$. Hence $q\mathbf{y}q^{-1} = \frac{q}{|q|}\mathbf{y}(\overline{\frac{q}{|q|}})$, and the normalized $\frac{q}{|q|} = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{v}$ for some $\theta \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}, |\mathbf{v}| = 1$. □

Notation A.1 Let $\mathbb{S}^3$ denote the 3-dimensional hypersphere (the **3-sphere**). It is the collection of quaternions with unit norm

$$\mathbb{S}^3 := \left\{ q \in \mathbb{R}^4 \cong \mathbb{H} \,\middle|\, |q|^2 = 1 \right\}. \tag{A.11}$$

Every $q \in \mathbb{S}^3$ gives rise to a rotation in the form $\mathbf{y} \mapsto q\mathbf{y}\bar{q}$.

A.3.1 Rotation Action

Let us denote for each unit quaternion $q \in \mathbb{S}^3$ the rotation matrix by

$$R_q \in \mathbb{R}^{3\times 3}, \quad R_q\mathbf{y} := q\mathbf{y}\bar{q}. \tag{A.12}$$

The explicit formula of this rotation matrix for $q = a + b\mathbb{i} + c\mathbb{j} + d\mathbb{k}$ is given by

$$R_{a+b\mathbb{i}+c\mathbb{j}+d\mathbb{k}} = \begin{bmatrix} a^2+b^2-c^2-d^2 & 2bc-2ad & 2bd+2ac \\ 2bc+2ad & a^2-b^2+c^2-d^2 & 2cd-2ab \\ 2bd-2ac & 2cd+2ab & a^2-b^2-c^2+d^2 \end{bmatrix}. \tag{A.13}$$

Formula (A.13) (a.k.a. the **Euler–Rodrigues formula**) was discovered by Leonhard Euler in 1775, much prior to Hailton's discovery of quaternions in 1843, as a way to directly construct a rotation matrix from 4 parameters a, b, c, d with $a^2 + b^2 + c^2 + d^2 = 1$. In this context, the four numbers a, b, c, d are also called the **Euler parameters**. Note that they are *not* the more well-known *Euler angles* also pioneered by Euler.

Theorem A.9 Concatenation of rotation corresponds to quaternion multiplication

$$R_pR_q = R_{pq} \quad \text{for } p, q \in \mathbb{S}^3. \tag{A.14}$$

Proof. $R_pR_q\mathbf{y} = p(q\mathbf{y}\bar{q})\bar{p} = (pq)\mathbf{y}(\bar{q}\bar{p}) = (pq)\mathbf{y}(\overline{pq}) = R_{pq}\mathbf{y}$. □

A.3.2 Unit Quaternions Doubly Cover the Rotation Group

In Corollary A.8, we see that (nonzero) quaternions represent rotations. From the expression $\mathbf{y} \mapsto q\mathbf{y}q^{-1}$, we observe that the rotation represented by q and aq, for any scaling factor $a \in \mathbb{R}$, $a \neq 0$, yields the same rotation.

By introducing the notion of unit quaternions, we discard any q that is not norm one, as their resulting rotations are already represented by its normalized version aq, $a = 1/|q|$. But such

normalizations only exploit the positive scaling factor $a > 0$. The pair of two different unit quaternions q and $-q$ (antipodal point of the 3-sphere) still represent the same 3D rotation.

Let us clarify the distinction between the space of unit quaternions $\mathbb{S}^3$ and the space of 3D rotations.

Notation A.2 Let the collection of all 3D rotations be denoted by

$$\mathrm{SO}(3) := \left\{ A \in \mathbb{R}^{3\times 3} \,\middle|\, A^\intercal A = I, \det(A) = 1 \right\} \tag{A.15}$$

Here "O" stands for orthogonal group, *i.e.* the group of square matrices that have orthonormal column vectors $A^\intercal A = I$; "SO" stands for special orthogonal group, corresponding to the additional condition $\det(A) = 1$ which rules out the reflection matrices.

The map

$$\begin{aligned} R_{(\cdot)} \colon \mathbb{S}^3 &\to \mathrm{SO}(3) \\ q &\mapsto R_q := (\mathbf{y} \mapsto q\mathbf{y}\bar{q}) \end{aligned}$$

is a 2-to-1 map, with antipodal points $q, -q$ sent to the same element in SO(3): $R_q = R_{-q}$. In fact it is continuous 2-to-1 map, with the preimage of a sufficiently small neighborhood around R_q in SO(3) being two disjoint small neighborhoods in $\mathbb{S}^3$ respectively around q and $-q$. We call such a 2-to-1 map a **double cover**.

R In other words, SO(3) is the resulting space after identifying all antipodal points of $\mathbb{S}^3$. SO(3) is topologically equivalent (homeomorphic) to the real projective space $\mathbb{RP}^3$.

A.4 Maneuvering in 3D Rotations

In many 3D graphics framework, the 3D rotation of an instanced object is described by a unit quaternion $q \in \mathbb{S}^3$. A point $\mathbf{x} = x_1\mathbb{i} + x_2\mathbb{j} + x_3\mathbb{k}$ in the template object is rotated to $R_q\mathbf{x} = x_1 q\mathbb{i}\bar{q} + x_2 q\mathbb{j}\bar{q} + x_3 q\mathbb{k}\bar{q}$.

In such a framework, a problem one would frequently encounter is that, knowing where a few points are rotated, what is the corresponding unit quaternion that describes such a rotation?

A.4.1 Dihedral Rotation

Here we talk about the rotation between two vectors. Between two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ with unit norm $|\mathbf{u}| = |\mathbf{v}| = 1$, *i.e.* two elements on the 2-spheres $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$, there are many rotations that rotate $\mathbf{u}$ to $\mathbf{v}$. For example, if $q \in \mathbb{S}^3$ rotates $\mathbf{u}$ to $\mathbf{v}$, *i.e.* $q\mathbf{u}\bar{q} = \mathbf{v}$, then so does $e^{(\theta/2)\mathbf{v}}q$ for any θ.

Among all unit quaternions that rotate $\mathbf{u}$ to $\mathbf{v}$ there is one that represents the "minimal" rotation. The most efficient way of rotating $\mathbf{u}$ to $\mathbf{v}$ (with least rotation angle) is a rotation about the axis $\frac{\mathbf{u}\times\mathbf{v}}{|\mathbf{u}\times\mathbf{v}|}$ that is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$. During the rotation, the vector $\mathbf{u}$ follows a great circle on $\mathbb{S}^2$ and moves from $\mathbf{u}$ to $\mathbf{v}$. When such a rotation is represented by a quaternion, we pick between $q, -q$ the one that has positive real part, so that it is the quaternion that is closer to $\mathbb{1}$.

Knowing the rotation axis ($\frac{\mathbf{u}\times\mathbf{v}}{|\mathbf{u}\times\mathbf{v}|}$) all we need is the rotation angle. The rotation angle is given by $\phi = \cos^{-1}(\langle \mathbf{u}, \mathbf{v} \rangle)$. Note that $\cos(\phi/2) = \sqrt{\frac{1+\cos\phi}{2}}$ and $\sin(\phi/2) = \sqrt{\frac{1-\cos\phi}{2}}$.

We consider the function which takes $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$ as inputs and returns the unit quaternion described above. In Houdini VEX library, such a function is called **dihedral**. In OpenGL Mathematics (GLM), it is called **rotation**. Let us call it the dihedral rotation:

$$\begin{aligned} &\text{Dihedral}\colon \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{S}^3 \\ &\text{Dihedral}(\mathbf{u}, \mathbf{v}) \in \mathbb{S}^3, \end{aligned}$$

defined by

$$\text{Dihedral}(\mathbf{u}, \mathbf{v}) = \begin{cases} \sqrt{\frac{1+\langle \mathbf{u}, \mathbf{v} \rangle}{2}} + \sqrt{\frac{1-\langle \mathbf{u}, \mathbf{v} \rangle}{2}} \frac{\mathbf{u} \times \mathbf{v}}{|\mathbf{u} \times \mathbf{v}|} & \mathbf{u} \neq \mathbf{v} \text{ and } \mathbf{u} \neq -\mathbf{v} \\ \mathbb{1} & \mathbf{u} = \mathbf{v} \\ \text{any unit vector } \mathbf{w} \perp \mathbf{u} & \mathbf{u} = -\mathbf{v}. \end{cases} \tag{A.16}$$

A.4.2 Parallel Transports along Great Circles on the 2-Sphere

Let $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$ be two given points on the sphere. Let us suppose $\mathbf{u} \neq -\mathbf{v}$. The tangent space of $\mathbb{S}^2$ at $\mathbf{u}$ consists of vectors orthogonal to $\mathbf{u}$. Similarly, the tangent space of $\mathbb{S}^2$ at $\mathbf{v}$ consists of vectors orthogonal to $\mathbf{v}$.

The dihedral rotation rotates $\mathbf{u}$ to $\mathbf{v}$. Along with it, the orthogonal complement of $\mathbf{u}$ (*i.e.* the tangent space at $\mathbf{u}$) is rotated to the orthogonal complement of $\mathbf{v}$ (*i.e.* the tangent space at $\mathbf{v}$). For each tangent vector $\mathbf{x}$ at $\mathbf{u}$, (that is $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{x} \perp \mathbf{u}$,) we get a corresponding vector $\mathbf{y} := \text{Dihedral}(\mathbf{u}, \mathbf{v})\, \mathbf{x}\, \overline{\text{Dihedral}(\mathbf{u}, \mathbf{v})}$ orthogonal to $\mathbf{v}$.

Geometrically, this map between the tangent spaces at $\mathbf{u}$ and $\mathbf{v}$ is given by *rolling* the tangent plane along the great circle connecting $\mathbf{u}$ and $\mathbf{v}$ until the plane becomes the tangent plane of $\mathbf{v}$.

We call this map $R_{\text{Dihedral}(\mathbf{u},\mathbf{v})}$ between the tangent spaces the **parallel transport along a great circle**.

A.4.3 Finding the Rotation for a Rotated Frame

Suppose $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{S}^2$ with $\mathbf{u}_1 \perp \mathbf{u}_2$. This is called a frame, since one gets a positively-oriented orthonormal basis $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ for free by setting $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$.

Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{S}^2$ with $\mathbf{v}_1 \perp \mathbf{v}_2$ be another frame. Then there is a unique rotation that maps $\mathbf{u}_1$ to $\mathbf{v}_1$ and maps $\mathbf{u}_2$ to $\mathbf{v}_2$. The quaternion $\pm q$ that represents this rotation is constructed by the following.

First, let $q_1 := \text{Dihedral}(\mathbf{u}_1, \mathbf{v}_1)$. Then R_{q_1} is the parallel transport from $\mathbf{u}_1^\perp (\ni \mathbf{u}_2)$ to $\mathbf{v}_1^\perp (\ni \mathbf{v}_2)$. Let $\mathbf{w} = R_{q_1}\mathbf{u}_2 = q_1 \mathbf{u}_2 \overline{q_1} \in \mathbf{v}_1^\perp$. In $\mathbf{v}_1^\perp$, $\mathbf{w}$ is not necessarily $\mathbf{v}_2$. Let $q_2 = \text{Dihedral}(\mathbf{w}, \mathbf{v}_2)$, which must be a rotation about $\mathbf{v}_1$ (and thus leaves $\mathbf{v}_1$ intact); in case $\mathbf{w} = -\mathbf{v}_2$ let $q_2 = \mathbf{v}_1$.

The final quaternion is obtained by concatenation $q = q_2 q_1$.

Index