Notes and Comments on Professor Carlen's Notes "Multivariable Calculus, Linear Algebra And Differential Equations"

October 23, 2023

The following notes and comments are prepared to provide some guidance for your reading of Professor Carlen's notes. We may also occasionally refer to the textbook Calculus, Early Transcendentals*, 3rd edition by Jon Rogawski and Colin Adams, to be denoted as [RC].*

Chapter 1

Geometry, Algebra and Analysis in Several Variables

We plan to cover this chapter using 4 lectures.

Lec. 1 Read 1.1.1-1.1.5.

Lec. 2 Read 1.2.1-1.2.2, 1.1.6.

Lec. 3 Read 1.1.7 - 1.1.8, 1.2.3.

Lec. 4 Read 1.3.1-1.3.4.

1.1 Algebra and Geometry in $\mathbb{R}^n$

1.1.1 Geometry, Algebra and Calculus

As remarked by Professor Carlen in the second paragraph on p.2, our strategy is to use the *visualization* in two and three dimensions to set up *algebraic algorithms* to compute quantities involving functions of multi-variables, and then *generalize* these algorithms to do computations when the number of variables is more than three.

1.1.2 Vector variables and Cartesian coordinates

Several commonly used notions in this course are introduced here:

- vectors
- reference system, and a base point

- ordered list of numbers
- Cartesian coordinates (rectangular coordinates)
- length (or magnitude, or norm) of a vector
- unit vector

The concept of vectors originates from describing *displacements, velocities*, and *forces*, which have the attributes of both *magnitude* and *direction*. Vectors of the same kind have a natural operation of *addition* and *scalar multiplication*. For instance, *displacement* 1 + *displacement* 2 can be defined as the combined effect of *displacement* 2 followed by *displacement* 1. This addition can be described as following the *parallelogram law*. Scalar multiplication of a vector by a scalar is defined in a geometrically obvious way.

Once a *rectangular coordinate system* is set up (in the three dimensional space), each vector $\mathbf{v}$ can be represented by an ordered list of three (real) numbers, called coordinates (or components), (x, y, z), and we identify $\mathbf{v} = (x, y, z)$. Then the vector addition is encoded as $(x, y, z) + (x', y', z') = (x + x', y + y', z + z')$, and the scalar multiplication of the vector (x, y, z) by the scalar c follows $c(x, y, z) = (cx, cy, cz)$. It turns out that the geometric parallelogram law of vector addition is encoded in this algebraic law of vector addition.

The geometric length of the vector $\mathbf{v}$ with coordinates (x, y, z) is given by $\sqrt{x^2 + y^2 + z^2}$ based on two successive applications of Pythagorean theorem, and we write this as $||\mathbf{v}|| = \sqrt{x^2 + y^2 + z^2}$ (please construct two right triangles associated with a given $\mathbf{v}$ to understand how Pythagorean theorem is applied to lead to $||\mathbf{v}|| = \sqrt{x^2 + y^2 + z^2}$).

Note that (geometric) vector addition and scalar multiplication are defined independent of their rectangular coordinate representation, that a vector has different coordinate representations in different rectangular coordinates (just as the movement of an aircraft has different coordinates by different flight control towers), and that the length of a vector and the angle between two vectors should be independent of the coordinates used to represent the vectors. This will be verified algebraically later in this course. In particular we will study the relations between different coordinate representations in different rectangular coordinates of the same vector.

But one can also represent vectors in terms of their polar coordinates (r, θ) (in two dimension) or spherical polar coordinates (r, θ, ϕ) (in three dimension), and vector addition would not be represented as the component-wise addition of their coordinates here. E.g., if $\mathbf{v}_1$ has polar coordinates $r = 1$ and $\theta = 0$, and $\mathbf{v}_2$ has polar coordinates $r = 1$ and $\theta = \frac{\pi}{2}$, then the vector $\mathbf{v}_1 + \mathbf{v}_2$ *does not* have polar coordinates $r = 1 + 1$ and $\theta = 0 + \frac{\pi}{2}$.

Many other quantities have the same attributes as two dimensional or three dimensional vectors, but may not have a simple geometric representation as the two dimensional or three dimensional vectors. Vectors are often defined as ordered lists of n (real) numbers, for a certain natural number n. General vector addition and scalar multiplication will be defined in 1.1.4; the length of a vector is defined in 1.1.5, using a natural generalization of the above algebraic properties.

There are also many other quantities that can be represented by ordered lists of n (real) numbers, but may not have a naturally defined addition or scalar multiplication. For example, the weather data at any location is described in terms of (temperature, barometric pressure, relative humidity), which is an ordered list of three numbers. Although there is a natural meaning to the difference of two such data points (at the same location over two different times, or at the same time over two different locations), which would measure the difference of temperature, barometric pressure, and relative humidity, how does one make sense of the addition of two such data points? — if one uses the same addition rule, what does it mean to say 55% relative humidity + 60% relative humidity = 115% relative humidity?

Math texts rarely discuss the contexts in which the vector operations are appropriate. Some texts, such as [RC], do distinguish between using n-tuple ordered numbers to represent the "state" of a certain entity and using them to represent the "change", or "increment", of the same entity. The latter has the attributes of vectors and the associated vector addition and scalar multiplication, while the former does not.

After raising awareness of this issue, we will also follow most texts not to make strict distinction between using n-tuple ordered numbers to represent the "state" of a certain entity and using them to represent the "change", or "increment", of the same entity, which would be a vector. For instance, regarding the weather data as a function of location and time, the weather data is not a vector properly, we still call it a vector-valued function, and will use vector calculus to analyze it.

Another commonly ignored issue in math texts is the use of *units* on different components of a vector. Using the same weather data example, if one would like to design a quantitative measure of difference of weather data, then it seems that

$$\sqrt{(t_1 - t_2)^2 + (p_1 - p_2)^2 + (h_1 - h_2)^2}$$

may be a good candidate for measuring how close the two pieces of weather data (t_1, p_1, h_1) and (t_2, p_2, h_2) are. But these components have different units, and quantities of different units can't be added! Using the proposed formula, a difference of $2^\circ F$ in temperature seems to be regarded as causing a bigger difference than a difference of 80% difference in relative humidity (assuming other data are kept the same). The proper way to handle this in applications is to first pre-process the data, i.e., to normalize the data. Namely, choosing appropriate unites T, P, H for the temperature,

barometric pressure, relative humidity, respectively, and use $(\frac{t}{T}, \frac{p}{P}, \frac{h}{H})$ as coordinate representation for the weather data. A mathematically equivalent way to describe this is that, in computing the "lengths", one places different *weights* on different components.

1.1.3 Parameterization

This subsection discusses how to obtain a parametrization of the round sphere, and how to use the parametrization to obtain the "tangent plane" to the round sphere at a particular point. The discussion is through a particular example, leaving out many details (e.g. equation for a plane, "best fit" plane) to be introduced soon. The example used gives rise to *spherical polar coordinates* of points in the three dimensional space.

Try to get an intuitive understanding for the concept of the tangent plane and how it is computed in this particular context. Don't get bogged down by the details of computations which you may not fully follow — we will do the discussion in more detail later on.

The main take-away is that, in multi-dimensions, there is often a need to study the geometry of an equation (or equations) via a parametric representation, even in the simplest case of a straight line in three or higher dimensions, but often one needs more than one parametric representation to describe the geometry of an equation. There will be *no systematic way* of finding a parametric representation of an arbitrary given equation, but polar spherical representation of a round sphere centered at the origin is so often used that a student needs to become proficient with using it.

Particular attention is drawn to

- equation of a plane in $\mathbb{R}^3$ on p.7 vs its parametric form on p.8.
- how a tangent plane to the round sphere at a point is obtained on pp.7-8.
- a graph as a parametric surface, as illustrated on p.9.

The earliest way to describe a surface is in terms of the graph of a function of two variables $z = f(x, y)$. Here one needs to specify the domain of definition of f in terms of (x, y). In the case of the unit round sphere centered at the origin $(0, 0, 0)$, as given by $\{(x, y, z) : x^2 + y^2 + z^2 = 1\}$, we could solve for z in terms of (x, y). It is given as $z = \pm\sqrt{1 - x^2 - y^2}$. One point of attention is the choice of $\pm$, i.e., for each (x, y) (such that $x^2 + y^2 < 1$), there are two choices of z; and when $x^2 + y^2 = 1$, the two formulae give the same points $(x, y, 0)$; furthermore, no single graph can represent the entire sphere.

In general, one single equation in three variables (x, y, z) of the form $F(x, y, z) = 0$ seems to represent a surface, as one can imagine solving one variable in terms of the

other two, so that solution is given in terms of the graph of a function of two variables. The precise conditions which make this process valid are spelled out in the *Implicit Function Theorem*. One often uses this kind of reasoning in theoretical discussions, but rarely carries out this procedure explicitly, as it is often not as easy as the case for the round sphere.

Another approach to describe a surface is that each coordinate is represented as a function of a common set of two variables on a common domain; these two variables are called parameters of the surface in this context, and such a representation of a surface is called a parametric representation.

In the case of the unit round sphere centered at the origin $(0,0,0)$, as given by $\{(x,y,z) : x^2+y^2+z^2=1\}$, the notes give a discussion which leads to

$$\begin{cases} x = x(\theta,\phi) = \cos\theta\sin\phi \\ y = y(\theta,\phi) = \sin\theta\sin\phi \\ z = z(\theta,\phi) = \cos\phi \end{cases}$$

for $(\theta,\phi) \in (-\pi,\pi] \times [0,\pi]$.

The domain of definition of this parameter representation is relatively simple, a rectangle. Each of the three functions is easy to understand and manipulate: keeping one variable fixed, each is an infinitely times differentiable function of the other variable. One technical complication of this parameter representation is that the map $X(\theta,\phi) := (x(\theta,\phi), y(\theta,\phi), z(\theta,\phi))$ is not *one-to-one* on this domain: two different sets of parameter values may map to the same point on the unit sphere. For example, $X(\theta,0) = X(\theta',0)$ for any $\theta,\theta' \in (-\pi,\pi]$. If one changes the domain of this parametric representation to $(-\pi,\pi] \times (0,\pi)$ to avoid this issue, then the modified map is no longer *onto*: no parameters in the domain $(-\pi,\pi] \times (0,\pi)$ can represent the north and south pole of the unit sphere $(x,y,z) = (0,0,\pm 1)$. This is a common issue in studying surfaces: it's often impossible to use a single parametric representation to describe a surface in a one-to-one and onto fashion.

Note that a surface given as the graph of a function $z = f(x,y)$ can also be treated as having a parametric representation, a particularly simple one in fact: $X(u,v) = (u,v,f(u,v))$, namely, $x = x(u,v) = u, y = y(u,v) = v$, and $z = z(u,v) = f(u,v)$. In order to avoid introducing unnecessary new variables, we simply identify u with x, v with y, and write $X(x,y) = (x,y,f(x,y))$.

Remark 1.1.1

One of the complications of multi-variable calculus is that, often, the domain of definition of a function, or of a parametric representation of a surface, is not as simple as a rectangle; one can sometimes restrict attention to functions defined on simple domains such as rectangles, but then one often needs multiple

parametric forms of a given surface—imagine how one would represent the surface of a donut.

Characterization of a surface and a plane Although we meet surfaces in many contexts, it is not that easy to give a good characterization of a surface. We observe two main features of a surface: (a) it is a "two-dimensional" object, and (b) it is mostly "smooth", except along some edges or corners. One possible candidate definition of a surface in $\mathbb{R}^n$ is that it is given as the graph of several coordinates as functions of two of the coordinates, say, $x_3, \ldots, x_n$ each as a function of $(x_1, x_2) : x_k = f_k(x_1, x_2)$, for $k = 3, \ldots, n$, over a certain domain D in $\mathbb{R}^2$. But this definition would be too restrictive, and would not be able to describe a sphere or a donut, which, in its entity, can't be described as a *single graph*.

A definition of surface which has the flexibility of describing commonly encountered surfaces is parametric representation, namely, it is in terms of n functions of two common parameters, say, (s, t), in a domain D of $\mathbb{R}^2$: $x_1 = x_1(s, t), \ldots, x_n = x_n(s, t)$, $(s, t) \in D$. The notion of a smooth surface is related to the notion that each of the function $x_i(s, t)$ is a (smooth) "differentiable" function of the two variables $(s, t) \in D$; however, this notion is not entirely correct, for at least two reasons: (i) even if each $x_i(s, t)$ is a "differentiable" function of the two variables $(s, t) \in D$, the resulting surface could still have sharp edges or corners, such as in the case of $x = s^2, y = s^3, z = t$; (ii) the set of functions may not represent a surface, even though there are two free parameters s and t at play, such as in the case of $x = s+t, y = 2(s+t), z = 3(s+t)$—this set of functions will represent a one-dimensional object, a straight-line, instead of a two-dimensional surface. The notion of a differentiable function of more than one variables will be discussed in Chapter 4.

However, a (two-dimensional) plane is relatively easy to characterize. It is "flat" in the sense that, given two "independent" directions $\mathbf{u}$ and $\mathbf{v}$ in a plane, and a reference point P_0 on it, any other point P in this plane can be arrived at from P_0 by following "a linear combination" of the two directions $\mathbf{u}$ and $\mathbf{v}$: $P = P_0 + s\mathbf{u} + t\mathbf{v}$, for some scalars s and t.

Example 1.1.1

Take $\mathbf{u} = (\frac{1}{2}, \frac{1}{2}, -\frac{1}{2})$, $\mathbf{v} = (-\frac{1}{2}, \frac{1}{2}, 0)$, $P_0 = (\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}})$, $P = (x, y, z)$, then

$$(x, y, z) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}}) + s(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}) + t(-\frac{1}{2}, \frac{1}{2}, 0) \quad (s, t) \in \mathbb{R}^2$$

gives a parametric representation of a plane through $P_0 = (\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}})$.

For the general situation, with $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$, and $P_0 = (x_{01}, x_{02}, x_{03})$, we then get

$$(x, y, z) = (x_{01}, x_{02}, x_{03}) + s(u_1, u_2, u_3) + t(v_1, v_2, v_3).$$

Given (x, y, z), this can be regarded as a system of 3 *linear equations* in s, t. If one can solve s, t in terms of x, y from the first two equations, then it turns out that s, t are also linear expressions in x, y: $s = \alpha x + \beta y + s_0$, $t = \gamma x + \delta y + t_0$ for some scalars $\alpha, \beta, \gamma, \delta, s_0, t_0$. Then one substitutes these linear expressions into the third equation, one would obtain z expressed as a linear expression in terms of x, y: $z = ax + by - d$ for some scalars a, b, d—this z agrees with the given z only if the given point (x, y, z) lies on the plane. This is a non-parametric form, or graph form, of the equation of a plane. It can be written in the form of $ax + by - z = d$, and is a particular case of the general graph form of the equation of an plane in $\mathbb{R}^3$.

Remark 1.1.2

The parametric form of a two-dimensional plane still works in $\mathbb{R}^n$, $n > 3$, while the graph form,

$$a_1x_1 + \cdots + a_nx_n = d,$$

represents a hyperplane, rather than a two-dimensional plane. Namely, if $\mathbf{u}$ and $\mathbf{v}$ are two "linearly independent" vectors in $\mathbb{R}^n$, $P_0 \in \mathbb{R}^n$, then

$$(s, t) \in \mathbb{R}^2 \mapsto P_0 + s\mathbf{u} + t\mathbf{v} \in \mathbb{R}^n$$

still represents a two-dimensional plane in $\mathbb{R}^n$, while an equation of the form $a_1x_1 + a_2x_2 + \ldots + a_nx_n = d$ (not all the coefficients a_i's are 0) would represent a hyperplane, instead of a plane, in $\mathbb{R}^n$, as one can solve one of them, say, x_n (assuming $a_n \neq 0$), in terms of the other variables, $x_1, \ldots, x_{n-1}$, to obtain a graph of a function of $n-1$ variables—it would represent an $(n-1)$-dimensional hyperplane, where $n - 1 > 2$ if $n > 3$.

Note that a non-parametric form of a plane (or hyperplane) is given in terms of a linear function of the coordinates $x_1, \ldots, x_n$, and a parametric form of a plane is given in terms of several "linear functions[a]" of two free parameters. However, some nonlinear functions in parametric form may also give rise to a flat plane. Here is a simple example: $(x, y, z) = s^3(1, 2, 3) + t(-1, -11)$.

[a] Informally, a linear function of several variables $s_1, \ldots, s_m$ is a function of the form $a_1s_1 + \ldots + a_ms_m + c$ for some coefficients $a_1, \ldots, a_m$, and c; but later on, we will call such functions affine functions, and reserve the name linear functions for those affine functions such that $c = 0$. We still call an equation like $a_1s_1 + \ldots + a_ms_m + c = 0$ a linear equation.

Remark 1.1.3

Likewise, a straight-line in $\mathbb{R}^n$, $n \geq 3$, is best described in parametric form: it is described as the set of points obtained from a reference point on it by adding some scalar multiple of a fixed direction vector. Namely, if $P_0 = (x_{01,}, \ldots, x_{0n})$ is a reference point on this straight-line, and $\mathbf{v} = (v_1, \ldots, v_n)$ is its direction vector, then any point $P = (x_1, \ldots, x_n)$ on this straight-line can be expressed as $P = P_0 + t\mathbf{v}$ for some scalar t. Note that, in this form, if P is given and we need to find the corresponding parameter t, then we would need to solve a system of linear equations in t. This parametric form of a straight-line is often the simplest to work with, but a straight-line can also arise from nonlinear functions of a parameter, such as when describing the motion of a particle along a straight-line whose spatial position is not a linear (or rather affine) function of the time variable t.

Reading Quizzes/Questions:

1. How do you determine a direction vector of a straight-line from knowing two points on it? If a straight-line is given as the intersection of two planes in $\mathbb{R}^3$, how would you find a direction vector and a parametric equation for the straight-line? Can you formulate a similar strategy or question in $\mathbb{R}^4$? (Don't get discouraged if you have no idea how to deal with the higher dimensional cases; we will gradually discuss the relevant ideas in this course.)

2. Do you have an interpretation for the coefficients a, b, and c in the non-parametric form of equation for a plane in $\mathbb{R}^3$? Can you determine them by knowing two points on the plane? Can you determine them from a parametric form of the plane?

3. What kind of equations, or systems of equations, would you have to solve to find the set of intersection of two or more planes in $\mathbb{R}^3$? Can you identify the possible intersections of two or more planes in $\mathbb{R}^3$? Can you adapt your analysis to $\mathbb{R}^n$ when $n > 3$? (HINT: Always start your analysis with a concrete and as simple as possible case. For example, how to adapt your analysis to the $n = 4$ case?)

4. What kind of equations, or systems of equations, would you have to solve to find the set of intersection of two or more straight-lines in $\mathbb{R}^n$ for $n = 2, 3$, or bigger? Can you identify the possible intersections of two or more straight-lines in $\mathbb{R}^n$?

5. Can you summarize Professor Carlen's discussion for computing the tangent plane of the round sphere in your own interpretation? Can you redo the computation treating the sphere as a graph over the (x, y) coordinates, when (x_0, y_0, z_0) is not on the equator? Can a portion of the sphere near (x_0, y_0, z_0) still be treated as a graph when it is on the equator?

A former TA for this course, Blair Seidler has some work sheets 2, 3 and 4 for his summer 2020 section of 251, which contain a set of good problems to test your understanding of the material in the first three subsections.

1.1.4 The vector space $\mathbb{R}^n$

Here in the absence of interpreting a vector in $\mathbb{R}^n$ as a quantity with magnitude and direction, we can still define *vector addition* and *scalar multiplication* of a vector in *purely algebraic* fashion. The key new concepts are linear combination of vectors and span of a given set of vectors. You should study carefully the proof of Theorem 1 and Example 5.

In subsection **1.2.2** we will see that a plane in $\mathbb{R}^n$ passing through the origin is characterized as the span of two vectors, none of which is a multiple of the other.

Reading Quizzes/Questions:

1. Can a ray (namely, a half-line) be the span of some vector?

2. Can the first quadrant of $\mathbb{R}^2$ be the span of some set of vectors? Is the union of the x and y axes the span of some set of vectors?

3. Is the span of two vectors in $\mathbb{R}^3$ always a plane?

4. Can every vector in $\mathbb{R}^2$ be written as a linear combination of $\{\mathbf{v}_1 := \frac{(1,1)}{\sqrt{2}}, \mathbf{v}_2 := \frac{(-1,1)}{\sqrt{2}}\}$? Given $\mathbf{x} = (x_1, x_2)$, how would you (s, t) such that $\mathbf{x} = s\mathbf{v}_1 + t\mathbf{v}_2$?

5. Can every a plane in $\mathbb{R}^n$ be characterized as the span of two vectors, none of which is a multiple of the other?

1.1.5 Geometry and the dot product

This is the first section in which the higher dimensional linear algebra enters in a significant way. We may no longer visualize the geometry of parallelogram law, or triangle inequality, or vector projection (to be discussed in the next subsection), but we should still use the two-dimensional and three-dimensional geometry to guide us. In particular, we should learn how the geometric behavior is captured using certain concepts and language of linear algebra.

Points of attention are drawn to

- the interchangeable usage of the concepts of *distance* and *metric**, and the usage of the concept of *length* of a vector (in some texts, the length of a vector is also called its *norm*).
- how Theorem 3 (Cauchy-Schwarz inequality) is needed to make sense of the notion of angle between vectors in $\mathbb{R}^n$.
- how Theorem 4 (Triangle inequality) is related to the geometry of a triangle.
- how the proof of Theorem 4 encodes the Pythagorean Theorem in $\mathbb{R}^n$ for a pair of orthogonal vectors $\mathbf{a}$ and $\mathbf{b}$: $\|\mathbf{a} \pm \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$ when $\mathbf{a} \cdot \mathbf{b} = 0$.

Remark 1.1.4

Note that the Cauchy-Schwarz inequality $|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\|\|\mathbf{b}\|$ *encodes the more complicated looking inequality*

$$|\sum_{i=1}^{n} a_i b_i| \leq \sqrt{\sum_{i=1}^{n} |a_i|^2} \sqrt{\sum_{i=1}^{n} |b_i|^2}, \tag{1.1}$$

and the triangle inequality $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$ *encodes*

$$\sqrt{\sum_{i=1}^{n} |a_i + b_i|^2} \leq \sqrt{\sum_{i=1}^{n} |a_i|^2} + \sqrt{\sum_{i=1}^{n} |b_i|^2}.$$

*In the context of points in $\mathbb{R}^n$, the distance between two points (vectors) is given in terms of the length of a vector; but the distance between two points can be defined in more general context, such as for points lying on a surface, where it may not be most appropriate to define the distance in terms of the length of a vector.

Remark 1.1.5

Note that the Euclidean distance in **Definition 7** *is derived from the Dot Product:* $\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}}$. *But there are other notions of distance on* $\mathbb{R}^n$ *which may be more suitable for certain problems, which may not be related to the notion of dot product, and in those contexts it would not have the notion of angle between vectors. One such example is the* Taxi-cab *metric* $\rho_{TC}(\mathbf{x},\mathbf{y}) := \|\mathbf{x}-\mathbf{y}\|_{TC} := \sum_{i=1}^n |x_i - y_i|$. *The name is due to the reason that a Taxi-cab can't travel from* $\mathbf{x}$ *to* $\mathbf{y}$ *along the straight line from* $\mathbf{x}$ *to* $\mathbf{y}$, *but has to travel along either horizontal or vertical streets from* $\mathbf{x}$ *to* $\mathbf{y}$. *Triangle inequality still holds for such a distance:* $\|\mathbf{x}\pm\mathbf{y}\|_{TC} \le \|\mathbf{x}\|_{TC} + \|\mathbf{y}\|_{TC}$.
As mentioned earlier, there are many contexts in which the norm of a vector and the dot product between vectors need to be weighted differently on different components. E.g. if $\mathbf{w} = (w_1,\dots,w_n)$, *where each* $w_i > 0$, *represents the weights on different components, then a dot product weighted by* $\mathbf{w}$ *can be defined as*

$$\mathbf{a}\cdot_{\mathbf{w}}\mathbf{b} := \sum_{i=1}^n w_i a_i b_i,$$

so the corresponding weighted norm is given by $\|\mathbf{a}\|_{\mathbf{w}} := \sqrt{\sum_{i=1}^n w_i|a_i|^2}$. *Note that proofs for (1.17) and (1.20), etc. in [EC] still work, simply replacing* $\cdot$ *by* $\cdot_{\mathbf{w}}$, *and the corresponding Cauchy-Schwarz inequality* $|\mathbf{a}\cdot_{\mathbf{w}}\mathbf{b}| \le \|\mathbf{a}\|_{\mathbf{w}}\|\mathbf{b}\|_{\mathbf{w}}$ *still holds, and encodes the more complicated looking inequality*

$$|\sum_{i=1}^n w_i a_i b_i| \le \sqrt{\sum_{i=1}^n w_i|a_i|^2}\sqrt{\sum_{i=1}^n w_i|b_i|^2}.$$

If you are still not very comfortable with the more abstract proofs for (1.20) in [EC], you can derive this one from the standard version (1.1) by treating $\sqrt{w_i}a_i$ *as* a_i, *and* $\sqrt{w_i}b_i$ *as* b_i *in (1.1) and applying (1.1) directly.*

Reading Quizzes/Questions:

1. Given $\mathbf{v} = (1,2,3,4)$, can you find w in $\mathbf{w} = (-2,0,2,w)$ such that $\mathbf{v}$ is orthogonal to $\mathbf{w}$? How about choosing w such that $\mathbf{v}$ and $\mathbf{w}$ form an angle of $\frac{3\pi}{4}$ radian? How about choosing w such that $\|\mathbf{v}\| = \|\mathbf{w}\|$?

2. Can you construct two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^4$ such that $\mathbf{v}\cdot\mathbf{w} = -10$, and $\|\mathbf{v}\| = \|\mathbf{w}\| = 3$?

3. Prove that if $a, b, c > 0$, then $3 \leq \sqrt{a+b+c}\sqrt{a^{-1}+b^{-1}+c^{-1}}$.

1.1.6 Parallel and orthogonal components

Understand how parallel and orthogonal components of $\mathbf{x}$ with respect to a (non-zero) vector $\mathbf{u}$ are constructed geometrically and algebraically, and then verify that the same algebraic construction produces vectors $\mathbf{x}_\parallel$ and $\mathbf{x}_\perp$ with the same property that

$$\mathbf{x} = \mathbf{x}_\parallel + \mathbf{x}_\perp \quad \text{with} \quad \mathbf{x}_\parallel \parallel \mathbf{u}, \quad \text{and} \quad \mathbf{x}_\perp \cdot \mathbf{u} = 0.$$

We will use this decomposition to provide another proof of the Cauchy-Schwarz inequality. Suppose that $\mathbf{y} \neq \mathbf{0}$, and set $\mathbf{u} = \mathbf{y}/\|\mathbf{y}\|$, then $\mathbf{x}_\parallel = (\mathbf{x} \cdot \mathbf{u})\mathbf{u}$. By the Pythagorean Theorem, $\|\mathbf{x}\|^2 = \|\mathbf{x}_\parallel\|^2 + \|\mathbf{x}_\perp\|^2 \geq \|\mathbf{x}_\parallel\|^2$. This implies

$$\|\mathbf{x}\| \geq \|\mathbf{x}_\parallel\| = |\mathbf{x} \cdot \mathbf{u}| = |\mathbf{x} \cdot \left(\frac{\mathbf{y}}{\|\mathbf{y}\|}\right)| = \frac{|\mathbf{x} \cdot \mathbf{y}|}{\|\mathbf{y}\|}.$$

We thus conclude that $\|\mathbf{x}\|\|\mathbf{y}\| \geq |\mathbf{x} \cdot \mathbf{y}|$. Equality would imply that $\mathbf{x}_\perp = \mathbf{0}$. But $\mathbf{x}_\perp = \mathbf{x} - \mathbf{x}_\parallel$, so this means that

$$\mathbf{x} = \mathbf{x}_\parallel = (\mathbf{x} \cdot \mathbf{u})\mathbf{u} = \frac{(\mathbf{x} \cdot \mathbf{y})}{\|\mathbf{y}\|^2}\mathbf{y},$$

namely, $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$.

Reading Quizzes/Questions:

1. If $\mathbf{v}$ is a non-zero multiple of $\mathbf{u}$, how do the parallel and orthogonal components of $\mathbf{x}$ with respect to $\mathbf{v}$ relate to those of $\mathbf{u}$?

2. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other, and let $\mathbf{x}_\perp$ and $\mathbf{y}_\perp$ denote, respectively, the orthogonal components of $\mathbf{x}$ and $\mathbf{y}$ with respect to some non-zero vector $\mathbf{u}$. Are $\mathbf{x}_\perp$ and $\mathbf{y}_\perp$ necessarily orthogonal to each other?

1.1.7 Orthonormal subsets of $\mathbb{R}^n$

Pay particular attention to Theorem 6 (Fundamental Theorem on Orthonormal Sets in $\mathbb{R}^n$) and its implications. One key implication is that the Pythagorean theorem in the form of $\|\mathbf{x}\|^2 = \sum_{i=1}^n (\mathbf{x} \cdot \mathbf{u}_i)^2$ holds for *any* set of n orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ in $\mathbb{R}^n$.

Reading Quizzes/Questions: Suppose that $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is a set of orthonormal vectors in $\mathbb{R}^n$, $n > k$, is it true that every vector $\mathbf{x}$ in $\mathbb{R}^n$ satisfy $\mathbf{x} = \sum_{i=1}^k (\mathbf{x} \cdot \mathbf{u}_i)\mathbf{u}_i$?

1.1.8 Householder reflections and orthonormal bases

Pay particular attention to

- How the definition of the Householder reflection (1.28) is related to the geometric mirror reflection in the plane orthogonal to $\mathbf{u}$.

- (1.27) would show a more geometric interpretation if paired with $\mathbf{x} = \mathbf{x}_\parallel + \mathbf{x}_\perp$, namely $\mathbf{h}_\mathbf{u}(\mathbf{x})$ simply flips the sign of the term $\mathbf{x}_\parallel$: $\mathbf{h}_\mathbf{u}(\mathbf{x}) = -\mathbf{x}_\parallel + \mathbf{x}_\perp$. It also follows from this that, if $\mathbf{y} = \mathbf{y}_\parallel + \mathbf{y}_\perp$, then $\mathbf{h}_\mathbf{u}(\mathbf{y}) = -\mathbf{y}_\parallel + \mathbf{y}_\perp$, and

$$\mathbf{h}_\mathbf{u}(\mathbf{x}) \cdot \mathbf{h}_\mathbf{u}(\mathbf{y}) = (-\mathbf{x}_\parallel + \mathbf{x}_\perp) \cdot (-\mathbf{y}_\parallel + \mathbf{y}_\perp) = \mathbf{x}_\parallel \cdot \mathbf{y}_\parallel + \mathbf{x}_\perp \cdot \mathbf{y}_\perp = \mathbf{x} \cdot \mathbf{y},$$

  as $\mathbf{x}_\parallel \cdot \mathbf{y}_\perp = 0 = \mathbf{x}_\perp \cdot \mathbf{y}_\parallel$.

- The most significant theoretical property of a Householder reflection, as described by (1.30); as a consequence, if $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is a set of orthonormal vectors, and $\mathbf{h}$ is any Householder reflection (in fact, any transformation of $\mathbb{R}^n$ satisfying (1.30)), then $\{\mathbf{h}(\mathbf{u}_1), \ldots, \mathbf{h}(\mathbf{u}_r)\}$ is a set of orthonormal vectors—this is used in the proof of Lemma 1.

  More specifically, if $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is a set of orthonormal vectors in $\mathbb{R}^n$, and if $\mathbf{u}_r = \mathbf{e}_n$, then we will take h to be the identity transformation; otherwise, one can construct a Householder transformation h such that $h(\mathbf{u}_r) = \mathbf{e}_n$. Since $h(\mathbf{u}_i) \cdot h(\mathbf{u}_j) = \mathbf{u}_i \cdot \mathbf{u}_j = 0$ for $i \neq j$; in particular, taking $i < j = r$, we find $h(\mathbf{u}_i) \cdot \mathbf{e}_n = 0$. Thus we can treat $\{\mathbf{h}(\mathbf{u}_1), \ldots, \mathbf{h}(\mathbf{u}_{r-1})\}$ as vectors in $\mathbb{R}^{n-1}$. If $r > n$, these would be $r - 1 > n - 1$ orthonormal vectors in $\mathbb{R}^{n-1}$, and an induction argument shows that this is impossible.

1.2 Lines and planes in $\mathbb{R}^3$

1.2.1 The cross product in $\mathbb{R}^3$

Pay particular attention to

- the geometric motivation of definition for the cross product in $\mathbb{R}^3$.
- the geometric interpretation of *triple product* (Theorem 8) as the *signed* volume of the parallelepiped with $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ as its adjacent edges.
- Lagrange's identity (Theorem 12).

To a physicist or an engineer, the most familiar properties of the cross product between two vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^3$ are (i) $\mathbf{a}\times\mathbf{b}\perp$ to both $\mathbf{a}$ and $\mathbf{b}$, (ii) the length of $\mathbf{a}\times\mathbf{b}$ equals the area of the parallelogram formed with $\mathbf{a}$ and $\mathbf{b}$ as its adjacent edges, and (iii) $\{\mathbf{a},\mathbf{b},\mathbf{a}\times\mathbf{b}\}$ obey the right-hand rotation rule.

For computations and derivations, the most important rules are those algebraic properties given by Theorem 7. They do not all follow easily from the physicists' definition of the cross product, but the latter follows more easily from the algebraic definition of cross product and Theorem 7.

If we reverse the logic of reasoning, and aim to define a product operation which obey the properties above, we can see that the defining formula,

$$\mathbf{a}\times\mathbf{b}=(a_2b_3-a_3b_3,a_3b_1-a_1b_3,a_1b_2-a_2b_1),$$

also follows from the above properties, for, if we write $\mathbf{a}=a_1\mathbf{e}_1+a_2\mathbf{e}_2+a_3\mathbf{e}_3$, and $\mathbf{b}=b_1\mathbf{e}_1+b_2\mathbf{e}_2+b_3\mathbf{e}_3$, then

$$\begin{aligned}\mathbf{a}\times\mathbf{b}=&(a_1\mathbf{e}_1+a_2\mathbf{e}_2+a_3\mathbf{e}_3)\times(b_1\mathbf{e}_1+b_2\mathbf{e}_2+b_3\mathbf{e}_3)\\=&a_1b_1\mathbf{e}_1\times\mathbf{e}_1+a_1b_2\mathbf{e}_1\times\mathbf{e}_2+a_1b_3\mathbf{e}_1\times\mathbf{e}_3+\\&a_2b_1\mathbf{e}_2\times\mathbf{e}_1+a_2b_2\mathbf{e}_2\times\mathbf{e}_2+a_2b_3\mathbf{e}_2\times\mathbf{e}_3+\\&a_3b_1\mathbf{e}_3\times\mathbf{e}_1+a_3b_2\mathbf{e}_3\times\mathbf{e}_2+a_3b_1\mathbf{e}_3\times\mathbf{e}_3\\=&(a_2b_3-a_3b_3)\mathbf{e}_1+(a_3b_1-a_1b_3)\mathbf{e}_2+(a_1b_2-a_2b_1)\mathbf{e}_3,\end{aligned}$$

using $\mathbf{e}_1\times\mathbf{e}_1=\mathbf{e}_2\times\mathbf{e}_2=\mathbf{e}_3\times\mathbf{e}_3=\mathbf{0}$, and $\mathbf{e}_1\times\mathbf{e}_2=-\mathbf{e}_2\times\mathbf{e}_1=\mathbf{e}_3$, $\mathbf{e}_2\times\mathbf{e}_3=-\mathbf{e}_3\times\mathbf{e}_2=\mathbf{e}_1$, and $\mathbf{e}_3\times\mathbf{e}_1=-\mathbf{e}_1\times\mathbf{e}_3=\mathbf{e}_2$.

Here is another consideration which leads to the concept of cross product (and determinant)*. Given three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in $\mathbb{R}^n$, we can construct a parallelepiped with them as adjacent edges. The volume of this parallelepiped is a function of these three vectors. How do we determine a formula for this volume in terms of the three vectors?

It turns out that the task becomes easier if we consider *signed* volume. Namely, we allow the volume to be negative; and if one of the vectors is flipped to its opposite, while the other two remain the same, then the signed volume would flip a sign as well. In other words, if we label this function as $V(\mathbf{a}, \mathbf{b}, \mathbf{c})$, then

$$V(-\mathbf{a}, \mathbf{b}, \mathbf{c}) = -V(\mathbf{a}, \mathbf{b}, \mathbf{c}), V(\mathbf{a}, -\mathbf{b}, \mathbf{c}) = -V(\mathbf{a}, \mathbf{b}, \mathbf{c}), \text{ etc.}$$

It certainly makes sense to generalize this property to

$$V(t\mathbf{a}, \mathbf{b}, \mathbf{c}) = tV(\mathbf{a}, \mathbf{b}, \mathbf{c}), V = (\mathbf{a}, t\mathbf{b}, \mathbf{c}) = tV(\mathbf{a}, \mathbf{b}, \mathbf{c}), \text{ etc.}$$

for any scalar t.

For $n = 3$, this function $V(\mathbf{a}, \mathbf{b}, \mathbf{c})$ has the additional property that

$$V(\mathbf{a} + \mathbf{a}', \mathbf{b}, \mathbf{c}) = V(\mathbf{a}, \mathbf{b}, \mathbf{c}) + V(\mathbf{a}', \mathbf{b}, \mathbf{c}), \tag{*}$$

for any vectors $\mathbf{a}, \mathbf{a}', \mathbf{b}, \mathbf{c}$, as well as similar properties for $\mathbf{b}$ and $\mathbf{c}$ (this property is not true if $n > 3$, as explained below).

This is because, if we treat the parallelogram with $\mathbf{b}, \mathbf{c}$ as adjacent edges as a base of the parallelepiped, then the dependence of $V(\mathbf{a}, \mathbf{b}, \mathbf{c})$ on $\mathbf{a}$ is only through its height with respect to the above parallelogram.

In $\mathbb{R}^3$, the plane containing the above parallelogram has a one dimensional normal, so the height with $\mathbf{a}+\mathbf{a}'$ replacing $\mathbf{a}$ = the height with $\mathbf{a}$ as edge + the height with $\mathbf{a}'$ replacing $\mathbf{a}$, if we consider signed volume, as this allows the possibility of two edges on the opposite sides of the plane containing the above parallelogram to cancel out in this consideration.

Using (*), and writing $\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3$, etc., we have

$$\begin{aligned} V(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= a_1 V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}) + a_2 V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}) + a_3 V(\mathbf{e}_3, \mathbf{b}, \mathbf{c}) \\ &= (a_1, a_2, a_3) \cdot (V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_3, \mathbf{b}, \mathbf{c})) \\ &= \mathbf{a} \cdot (V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_3, \mathbf{b}, \mathbf{c})) \end{aligned}$$

This vector $(V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_3, \mathbf{b}, \mathbf{c}))$ has the property

$$\mathbf{b} \cdot (V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_3, \mathbf{b}, \mathbf{c})) = V(\mathbf{b}, \mathbf{b}, \mathbf{c}) = 0,$$

and

$$\mathbf{c} \cdot (V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_3, \mathbf{b}, \mathbf{c})) = V(\mathbf{c}, \mathbf{b}, \mathbf{c}) = 0,$$

*The discussions in small fonts are supplemental material for interested students.

as the "heights" in both cases collapse to 0. So this vector is $\perp$ to both $\mathbf{b}$ and $\mathbf{c}$, and points in the direction (up to a sign) of the normal to the plane containing the base parallelogram. We relabel it as $\mathbf{b} \times \mathbf{c}$. Thus

$$V(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}), \text{ with } \mathbf{b} \times \mathbf{c} = (V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}), V(\mathbf{e}_3, \mathbf{b}, \mathbf{c})).$$

Since

$$V(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = |\mathbf{a}||\mathbf{b} \times \mathbf{c}| \cos\theta,$$

where θ is the angle between $\mathbf{b} \times \mathbf{c}$ (parallel to normal) and $\mathbf{a}$, and $|\mathbf{a}| \cos\theta$ gives the height of the parallelepiped with respect to the above parallelogram base, we see that $|\mathbf{b} \times \mathbf{c}|$ is the area of the parallelogram base. In conclusion, $\mathbf{b} \times \mathbf{c}$ is $\perp$ to both $\mathbf{b}$ and $\mathbf{c}$, and has length equal to the area of the parallelogram with $\mathbf{b}$ and $\mathbf{c}$ as adjacent edges. The final step is to stipulate that $V(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) > 0$, in fact $= 1$, when $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ forms a right-handed orthonormal basis. Thus

$$V(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = V(\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1) = V(\mathbf{e}_3, \mathbf{e}_1, \mathbf{e}_2) = 1.$$

We claim that this leads to

$$\mathbf{b} \times \mathbf{c} = (b_2c_3 - b_3c_2, b_3c_1 - b_1c_3, b_1c_2 - b_2c_1).$$

This follows from

$$\begin{aligned} V(\mathbf{e}_1, \mathbf{b}, \mathbf{c}) =& b_1 V(\mathbf{e}_1, \mathbf{e}_1, \mathbf{c}) + b_2 V(\mathbf{e}_1, \mathbf{e}_2, \mathbf{c}) + b_3 V(\mathbf{e}_1, \mathbf{e}_3, \mathbf{c}) \\ =& b_2 V(\mathbf{e}_1, \mathbf{e}_2, \mathbf{c}) + b_3 V(\mathbf{e}_1, \mathbf{e}_3, \mathbf{c}) \quad \text{as } V(\mathbf{e}_1, \mathbf{e}_1, \mathbf{c}) = 0 \\ =& b_2 \left[c_1 V(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1) + c_2 V(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_2) + c_3 V(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)\right] + \\ & b_3 \left[c_1 V(\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_1) + c_2 V(\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_2) + c_3 V(\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_3)\right] \\ & = b_2c_3 - b_3c_2. \end{aligned}$$

Similarly, $V(\mathbf{e}_2, \mathbf{b}, \mathbf{c}) = b_3c_1 - b_1c_3$, $V(\mathbf{e}_3, \mathbf{b}, \mathbf{c}) = b_1c_2 - b_2c_1$. We thus complete our proof of the claim.

Based on this interpretation of $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ as the signed volume of the parallelepiped with $\mathbf{a}, \mathbf{b}, \mathbf{c}$ as it adjacent edges, it follows that

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}),$$

as the latter two expressions simply compute the volume by taking the base as the parallelograms with $\mathbf{c}, \mathbf{a}$, or $\mathbf{a}, \mathbf{b}$, respectively, as its edges.

We will later identify $V(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ as the determinant of the matrix with the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$ as its rows (or columns) in that order.

When $n > 3$, vectors orthogonal to both $\mathbf{b}$ and $\mathbf{c}$ are no longer parallel to each other, namely, given in terms of scalar multiplication of one vector—this amounts to solving $\mathbf{x} \cdot \mathbf{b} = \mathbf{x} \cdot \mathbf{c} = 0$, which is a system of 2 linear equations in n unknowns; we will see that it has at least $n - 2 > 1$ free variables in its solution. Thus the heights of the parallelepipeds with

the parallelogram formed by $\mathbf{b}$ and $\mathbf{c}$ as its edges, and with $\mathbf{a}$, $\mathbf{a}'$, and $\mathbf{a}+\mathbf{a}'$ as its 3rd edge, respectively, are no longer parallel, so we may no longer have (*) in such cases!

An analogy which can be visualized is the signed area $A(\mathbf{a},\mathbf{b})$ of the parallelogram with $\mathbf{a}$ and $\mathbf{b}$ as its adjacent edges. In two dimension, vectors perpendicular to $\mathbf{b}$ are parallel to each other, we thus can establish

$$A(\mathbf{a}+\mathbf{a}',\mathbf{b}) = A(\mathbf{a},\mathbf{b}) + A(\mathbf{a}',\mathbf{b}). \qquad (**)$$

In the three dimensional space $\mathbb{R}^3$, when $\mathbf{b} \neq \mathbf{0}$, $\mathbf{a}$ is not parallel to $\mathbf{a}'$, the heights of the parallelograms formed by $\mathbf{a},\mathbf{b}$, respectively, by $\mathbf{a}',\mathbf{b}$, and $\mathbf{a}+\mathbf{a}',\mathbf{b}$ are not in the same direction, so we may no longer have (**). This explains why the algebraic behavior of $A(\mathbf{a},\mathbf{b})$ is different in $\mathbb{R}^n$ when $n > 2$.

Similarly $V(\mathbf{a},\mathbf{b},\mathbf{c})$ behaves differently in $\mathbb{R}^n$ when $n > 3$, and this also explans why cross product is not defined in $\mathbb{R}^n$ when $n > 3$ (however, a more complicated extension, *exterior product*, can be defined, for any n vectors $\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_n$ in $\mathbb{R}^n$, the signed n-dimensional volume $V(\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_n)$ is defined and is the determinant of the matrix with these vectors as columns).

Lagrange's identity may look intimidating and hard to grasp. But there is also a reasonable explanation for the appearance of the terms on the right hand side. $\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) \perp \mathbf{b}\times\mathbf{c}$, and in $\mathbb{R}^3$, vectors $\perp \mathbf{b}\times\mathbf{c}$ must be a linear combination $\beta\mathbf{b}+\gamma\mathbf{c}$ of $\mathbf{b}$ and $\mathbf{c}$ for some scalars β and γ. What remains is to find how β and γ are determined in terms of $\mathbf{a},\mathbf{b}$, and $\mathbf{c}$.

One way to make it easier is to think of two of the three vectors, say $\mathbf{b}$ and $\mathbf{c}$ as fixed, and treat both sides as a function of the remaining vector, $\mathbf{a}$, in this set up:

$$L(\mathbf{a}) := \mathbf{a}\times(\mathbf{b}\times\mathbf{c}), \quad R(\mathbf{a}) := (\mathbf{a}\cdot\mathbf{c})\mathbf{b} - (\mathbf{a}\cdot\mathbf{b})\mathbf{c}.$$

Note that

$$L(\mathbf{a}_1+\mathbf{a}_2) = L(\mathbf{a}_1)+L(\mathbf{a}_2), \quad \text{and} \quad L(t\mathbf{a}_1) = tL(\mathbf{a}_1),$$
$$R(\mathbf{a}_1+\mathbf{a}_2) = R(\mathbf{a}_1)+R(\mathbf{a}_2), \quad \text{and} \quad R(t\mathbf{a}_1) = tR(\mathbf{a}_1),$$

for any vectors $\mathbf{a}_1$ and $\mathbf{a}_2$, and scalar t. The two shared properties of L and R are called linearity properties, and they are essential for the derivation below as well as for the method in Professor Carlen's notes.

If one of $\{\mathbf{b},\mathbf{c}\}$ is a scalar multiple of the other, then $\mathbf{b}\times\mathbf{c} = \mathbf{0}$, and one can also see easily that $R(\mathbf{a}) = \mathbf{0}$, and $L(\mathbf{a}) = R(\mathbf{a})$ in such a case.

If none of vectors in $\{\mathbf{b},\mathbf{c}\}$ is a scalar multiple of the other, then a modification of the proof for Theorem 6 shows that any vector of $\mathbb{R}^3$ is a linear combination of

the vectors $\{\mathbf{b}, \mathbf{c}, \mathbf{b} \times \mathbf{c}\}$. Thus, to verify that $L(\mathbf{a}) = R(\mathbf{a})$ for any vector $\mathbf{a}$ in $\mathbb{R}^3$, it suffices to check it for $\mathbf{a}$ to be any one of the vectors $\{\mathbf{b}, \mathbf{c}, \mathbf{b} \times \mathbf{c}\}$. That is a much easier task to complete. For example, if we take $\mathbf{a} = \mathbf{b} \times \mathbf{c}$, then $L(\mathbf{b} \times \mathbf{c}) = \mathbf{0}$, while

$$R(\mathbf{b} \times \mathbf{c}) = ((\mathbf{b} \times \mathbf{c}) \cdot \mathbf{c})\mathbf{b} - ((\mathbf{b} \times \mathbf{c}) \cdot \mathbf{b})\mathbf{c} = \mathbf{0},$$

so $L(\mathbf{b} \times \mathbf{c}) = R(\mathbf{b} \times \mathbf{c})$; while if we take $\mathbf{a} = \mathbf{b}$, then $L(\mathbf{b}) = \mathbf{b} \times (\mathbf{b} \times \mathbf{c})$ is $\perp$ to both $\mathbf{b}$ and $\mathbf{b} \times \mathbf{c}$, so it must be a linear combination of $\mathbf{b}$ and $\mathbf{c}$ perpendicular to $\mathbf{b}$. Its length equals $\|\mathbf{b}\|\|\mathbf{b} \times \mathbf{c}\| = \|\mathbf{b}\|^2\|\mathbf{c}_\perp\|$, where $\mathbf{c}_\perp$ is the perpendicular component of $\mathbf{c}$ along $\mathbf{b}$. But

$$\begin{aligned} \|R(\mathbf{b})\| &= \|\,(\mathbf{b} \cdot \mathbf{c})\,\mathbf{b} - \|\mathbf{b}\|^2\mathbf{c}\| \\ &= \|\mathbf{b}\|^2\|\,(\mathbf{u} \cdot \mathbf{c})\,\mathbf{u} - \mathbf{c}\| \\ &= \|\mathbf{b}\|^2\|\|\mathbf{c}_\perp\|, \end{aligned}$$

recalling that $\mathbf{c}_\perp = \mathbf{c} - (\mathbf{u} \cdot \mathbf{c})\,\mathbf{u}$, with $\mathbf{u} := \mathbf{b}/\|\mathbf{b}\|$ being the unit vector in the direction of $\mathbf{b}$. Now that both $L(\mathbf{b})$ and $R(\mathbf{b})$ are linear combinations of $\mathbf{b}$ and $\mathbf{c}$ perpendicular to $\mathbf{b}$ and with the same length, we must have $L(\mathbf{b}) = R(\mathbf{b})$ or $L(\mathbf{b}) = -R(\mathbf{b})$. It remains to rule out the latter case. It will then establish $L(\mathbf{a}) = R(\mathbf{a})$ in all cases.

This strategy of reducing the proof of a general vector identity to simpler cases often works in other settings dealing with cross product. Professor Carlen's proof uses a somewhat different, but also very useful reduction. The key idea there is to express $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ in terms of a set of orthonormal basis $\{\mathbf{u}, \mathbf{v}, \mathbf{u} \times \mathbf{v}\}$ with $\mathbf{u}$ being the unit vector in the direction of $\mathbf{b}$, and $\mathbf{v}$ being the unit vector in the direction of $\mathbf{c}_\perp := \mathbf{c} - (\mathbf{c} \cdot \mathbf{u})\mathbf{u}$.

Reading Quizzes/Questions: Suppose that $\mathbf{x}_0$ is a solution to $\mathbf{a} \times \mathbf{x} = \mathbf{b}$. Can you describe the set of all solutions $\mathbf{x}$ to $\mathbf{a} \times \mathbf{x} = \mathbf{b}$?

1.2.2 Lines and planes in $\mathbb{R}^3$

Study carefully Examples 12-13, 15-16, 17 on the equations for a plane and a line in $\mathbb{R}^3$, in particular, the multi-viewpoints in analyzing Example 15. Keep in mind that both involve linear equation(s), and that a plane equation should have precisely two free variables(parameters), but a line should have a single free variable.

Reading Quizzes/Questions:

- How does one set up a procedure to find the intersection of two planes given by non-parametric equations?

- What if the two planes, or one of them, is given by parametric equations? E.g., what would be the intersection of two planes, one of which is given by $x+2y+3z=6$ and the other given by $(x,y,z)=(1,2,1)+s(1,-1,0)+t(1,0,-1)$?
- How does one set up a procedure to find the intersection of two lines? What do you expect for the possible set of solutions?
- To what extent can the problem and its solution be adapted to $\mathbb{R}^n$ for $n>3$?

Below are some discussions, through examples, to the setting of $\mathbb{R}^n$ for $n>3$. We will see that the solution set to the system of 2 linear equations in 4 variables

$$\begin{cases} x_1+3x_2-x_3+5x_4=6 \\ x_1+2x_2+x_3-3x_4=-2 \end{cases}$$

can be interpreted as a two-dimensional plane in $\mathbb{R}^4$, as one can eliminate x_1 first by subtracting the two sides of the equations to obtain $x_2-2x_3+8x_4=8$, thus can treat x_3 and x_4 as free variables, and solve x_2 in terms of them to obtain $x_2=2x_3-8x_4+8$, and finally substitute this into any of the two equations to get x_1:

$$\begin{aligned} x_1 &= -2x_2-x_3+4x_4-2 \\ &= -2(2x_3-8x_4+8)-x_3+4x_4-2 \\ &= -5x_3+20x_4-18. \end{aligned}$$

We can then write the solution in vector form

$$\begin{aligned} (x_1,x_2,x_3,x_4) &= (-5x_3+20x_4-18, 2x_3-8x_4+8, x_3, x_4) \\ &= (-18,8,0,0)+x_3(-5,2,1,0)+x_4(20,-8,0,1). \end{aligned}$$

Thus we can interpret $(-18,8,0,0)$ as a base point on this plane, and $(-5,2,1,0)$, $(20,-8,0,1)$ as two independent directions of this plane.

In the same vain, the solution of the system

$$\begin{cases} x_1+3x_2-x_3+5x_4=6 \\ x_1+2x_2+x_3-3x_4=-2 \\ \qquad x_2-x_3+6x_4=0 \end{cases}$$

can be interpreted as a one-dimensional line, as a similar procedure will show that $x_3=2x_4-8$, so we can treat x_4 as the free variable, and obtain the solution as

$$(x_1,x_2,x_3,x_4)=(22,-8,-8,0)+x_4(10,-4,2,1).$$

So $(22, -8, -8, 0)$ is a base point on this line, and $(10, -4, 2, 1)$ is a direction vector of this line.

So in $\mathbb{R}^n$, $n > 3$, it is natural to define a line in parametric form as $\mathbf{x} = \mathbf{x}_0 + s\mathbf{u}$ for $s \in \mathbb{R}$ and some non-zero direction vector $\mathbf{u}$ (often taken as a unit vector), and define a plane in in parametric form as $\mathbf{x} = \mathbf{x}_0 + s\mathbf{u} + t\mathbf{v}$ for $s, t \in \mathbb{R}$ and some non-zero vectors $\mathbf{u}, \mathbf{v}$ which are not multiples of each other. Each equation of the form $a_1x_1 + \cdots + a_nx_n = d$, where not all a_i are 0, represents a hyperplane in $\mathbb{R}^n$, which has $n-1$ free variables. The intersection of a number of hyperplanes may produce a line or a plane, depending on how whether we end up getting one or two free parameters in the solutions.

1.2.3 Distance problems

Three distance problems are discussed in this subsection:

(a). distance from a point to a line;

(b). distance from a point to a plane; and

(c). distance between two lines in $\mathbb{R}^3$.

The distance formula may take a different form in the three cases, but all involve the parallel or orthogonal component of a vector. The usage of cross product, which is only for $\mathbb{R}^3$, is not essential for the discussion of this subsection.

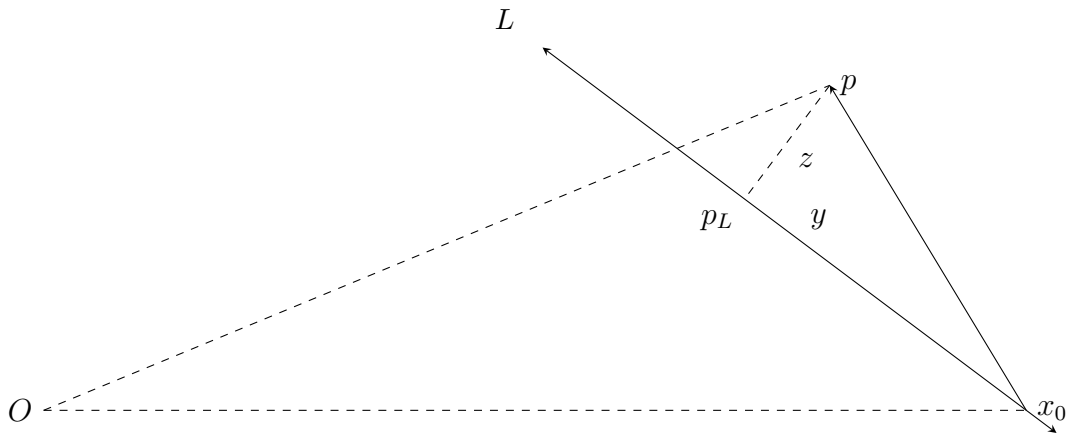
All three problems are examples of a more general minimization problem: given two sets X and Y of $\mathbb{R}^n$, find the minimum $\|\mathbf{x} - \mathbf{y}\|$ among $\mathbf{x} \in X$ and $\mathbf{y} \in Y$. The solution to this general problem requires a more careful formulation of the problem. In some problems (see the discussion on least squares problems in the next section), one is interested in not only the minimum value, but also the point(s) $\mathbf{x}_*$ in X and $\mathbf{y}_*$ in Y that attain the minimum value.

One can solve these problems by either applying tools of calculus or using geometric arguments.

In the case of (a), suppose the line is given in parametric form $\mathbf{x}(s) = \mathbf{x}_0 + s\mathbf{u}$, with $\mathbf{u}$ being a unit direction vector of the line, and $\mathbf{p}$ is a given point in $\mathbb{R}^3$. Then one needs to find the minimum distance using the relevant data $\mathbf{x}_0, \mathbf{u}, \mathbf{p}$.

Here is a geometric approach. Let $\mathbf{y}$ and $\mathbf{z}$ denote the parallel and orthogonal components of the vector $\mathbf{p} - \mathbf{x}_0$ with respect to the direction vector $\mathbf{u}$ of the line:

$$\mathbf{p} - \mathbf{x}_0 = \mathbf{y} + \mathbf{z}, \quad \mathbf{y} \parallel \mathbf{u}, \quad \mathbf{z} \perp \mathbf{u}.$$



Then $\mathbf{p}_L = \mathbf{x}_0 + \mathbf{y}$ is a point on L, and for any point $\mathbf{x}(s)$ on this line,

$$\mathbf{p} - \mathbf{x}(s) = \mathbf{p} - \mathbf{x}_0 - (\mathbf{x}(s) - \mathbf{x}_0) = \mathbf{z} - (\mathbf{x}(s) - \mathbf{x}_0 - \mathbf{y}),$$

with $\mathbf{z} \cdot \mathbf{u} = 0$, and $(\mathbf{x}(s) - \mathbf{x}_0 - \mathbf{y})$ a scalar multiple of $\mathbf{u}$, as each of $\mathbf{x}(s) - \mathbf{x}_0$ and $\mathbf{y}$ is a scalar multiple of $\mathbf{u}$. Now by the Pythagorean theorem,

$$\|\mathbf{p} - \mathbf{x}(s)\|^2 = \|\mathbf{z}\|^2 + \|\mathbf{x}(s) - \mathbf{x}_0 - \mathbf{y}\|^2 \geq \|\mathbf{z}\|^2 = \|\mathbf{p} - \mathbf{x}_0\|^2 - \|\mathbf{y}\|^2,$$

and equality is attained if and only if $\mathbf{x}(s) - \mathbf{x}_0 - \mathbf{y} = \mathbf{0}$. This shows that $p_L = \mathbf{x}_0 + \mathbf{y}$ is the unique closest point to $\mathbf{p}$ on the line, and the shortest distance is the square root of $\|\mathbf{p} - \mathbf{x}_0\|^2 - \|\mathbf{y}\|^2$, with $\mathbf{y} = (\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u}$ (using $\mathbf{u}$ as a unit vector; when $\mathbf{u}$ is not a unit vector, this should be appropriately adjusted.)

This shortest distance squared turns out to be

$$\|(\mathbf{p} - \mathbf{x}_0) - \mathbf{y}\|^2 = \|\mathbf{z}\|^2 = \|\mathbf{p} - \mathbf{x}_0\|^2 - |(-\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u}|^2.$$

Note that the derivation uses $\mathbf{y}, \mathbf{z}$, but the final solution avoids computing these vectors explicitly: one only needs to compute $\mathbf{p} - \mathbf{x}_0$ and $(\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u}$.

If we use calculus tools, we would set up the squared distance from $\mathbf{x}(s)$ to $\mathbf{p}$ as a function of s : $f(s) = \|\mathbf{p} - \mathbf{x}(s)\|^2$. If $f(s)$ attains a minimum at some s_*, then $f'(s_*) = 0$. But we can write $f(s) = (\mathbf{p} - \mathbf{x}(s)) \cdot (\mathbf{p} - \mathbf{x}(s))$, and using $\mathbf{x}'(s) = \mathbf{u}$ and product rule of differentiation, we find

$$f'(s_*) = -2(\mathbf{p} - \mathbf{x}(s_*)) \cdot \mathbf{u} = -2\left[(\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u} - s_*\right] = 0, \text{ using } \mathbf{u} \cdot \mathbf{u} = 1.$$

Geometrically, this means that $\mathbf{p} - \mathbf{x}(s_*) \perp \mathbf{u}$. One can then use the s_* found here and Pythagorean Theorem to evaluate $f(s_*)$ as above. A missing detail is to show that the $f(s)$ indeed attains a minimum value on $\mathbb{R}$.

In the case of (b), suppose that the plane is given by the equation $\mathbf{n}\cdot(\mathbf{x}-\mathbf{x}_0)=0$, where $\mathbf{n}$ is a unit vector (a normal to the plane). Consider now the parallel and orthogonal components of $\mathbf{x}_0-\mathbf{p}$ with respect to $\mathbf{n}$: $\mathbf{x}_0-\mathbf{p}=\delta\mathbf{n}+\mathbf{v}$ for some scalar δ and vector $\mathbf{v}$ such that $\mathbf{v}\cdot\mathbf{n}=0$ (I didn't have time to create a diagram to illustrate the geometry, but you should draw some diagrams to illustrate the computations below.). Then for any $\mathbf{x}$ on this plane,

$$\mathbf{x}-\mathbf{p}=\mathbf{x}-\mathbf{x}_0+\mathbf{x}_0-\mathbf{p}=\mathbf{x}-\mathbf{x}_0+\delta\mathbf{n}+\mathbf{v},$$

where we know that $(\mathbf{x}-\mathbf{x}_0)\cdot\mathbf{n}=0$. Set $\mathbf{x}-\mathbf{x}_0+\mathbf{v}$ as $\mathbf{w}$. Then $\mathbf{n}\cdot\mathbf{w}=0$, and $\mathbf{x}-\mathbf{p}=\delta\mathbf{n}+\mathbf{w}$. By Pythagorean theorem

$$\|\mathbf{x}-\mathbf{p}\|^2=\|\delta\mathbf{n}\|^2+\|\mathbf{w}\|^2\geq\|\delta\mathbf{n}\|^2=\|\mathbf{x}_*-\mathbf{p}\|^2,$$

where $\mathbf{x}_*=\mathbf{x}_0-\mathbf{v}$ is a point on the plane, as $(\mathbf{x}_*-\mathbf{x}_0)\cdot\mathbf{n}=0$. Thus the distance from $\mathbf{p}$ to the plane is given by $\|\mathbf{x}_*-\mathbf{p}\|=|\delta|$, which can be calculated as $|(\mathbf{x}_0-\mathbf{p})\cdot\mathbf{n}|$.

Suppose that in the situation of $\mathbb{R}^3$, the equation of the plane is given in the form of $ax+by+cz=d$. No base point $\mathbf{x}_0$ is given, but any point x_0 on the plane satisfies $ax_0+by_0+cz_0=d$. In vector form, we need to identify $\mathbf{n}=(a,b,c)/\sqrt{a^2+b^2+c^2}$, and the equation can be written as $\mathbf{n}\cdot\mathbf{x}=d/\sqrt{a^2+b^2+c^2}$, or $\mathbf{n}\cdot(\mathbf{x}-\mathbf{x}_0)=0$. Then the distance from $\mathbf{p}=(p_1,p_2,p_3)$ to the plane is

$$|(\mathbf{x}_0-\mathbf{p})\cdot\mathbf{n}|=\frac{|\mathbf{x}_0\cdot(a,b,c)-\mathbf{p}\cdot(a,b,c)|}{\sqrt{a^2+b^2+c^2}}=\frac{|d-\mathbf{p}\cdot(a,b,c)|}{\sqrt{a^2+b^2+c^2}}=\frac{|d-(ap_1+bp_2+cp_3)|}{\sqrt{a^2+b^2+c^2}}.$$

Remark 1.2.1

A common theme in both solutions is to construct some $\mathbf{x}_*$ in the line (plane) such that for any $\mathbf{x}$ in the line (plane), $(\mathbf{p}-\mathbf{x}_*)\cdot(\mathbf{x}-\mathbf{x}_*)=0$; in other words, $(\mathbf{p}-\mathbf{x}_*)\perp$ to every direction vector in the line (plane). This $\mathbf{x}_*$ is called the orthogonal projection of $\mathbf{p}$ in the line (plane). This idea will show up in more general problems.

The above procedure works in any dimensions. A plane in $\mathbb{R}^n$, $n>3$, can also be described as $\mathbf{x}_0+s\mathbf{v}_1+t\mathbf{v}_2$ for some vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbb{R}^n$ which are not multiples of each other. To find the distance from $\mathbf{p}$ to this plane, we need to find the minimum of

$$\|\mathbf{x}_0-\mathbf{p}+s\mathbf{v}_1+t\mathbf{v}_2\|^2 \text{ over } s,t\in\mathbb{R}.$$

When the minimum is attained at some $\mathbf{x}_*=\mathbf{x}_0+s_*\mathbf{v}_1+t_*\mathbf{v}_2$ for some s_*,t_*, then

$$(\mathbf{x}_*-\mathbf{p})\cdot\mathbf{v}_i=0 \text{ for } i=1,2. \tag{1.2}$$

This amounts to solving

$$\begin{cases} 0 = (\mathbf{x}_0 - \mathbf{p} + s_*\mathbf{v}_1 + t_*\mathbf{v}_2) \cdot \mathbf{v}_1 & = (\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{v}_1 + s_*\mathbf{v}_1 \cdot \mathbf{v}_1 + t_*\mathbf{v}_2 \cdot \mathbf{v}_1 \\ 0 = (\mathbf{x}_0 - \mathbf{p} + s_*\mathbf{v}_2 + t_*\mathbf{v}_2) \cdot \mathbf{v}_2 & = (\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{v}_2 + s_*\mathbf{v}_1 \cdot \mathbf{v}_2 + t_*\mathbf{v}_2 \cdot \mathbf{v}_2. \end{cases}$$

It is clear that this system is easy to solve when $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$ *(and* $\mathbf{v}_1 \cdot \mathbf{v}_1 = \mathbf{v}_2 \cdot \mathbf{v}_2 = 1$*, but this is less crucial). This is one reason why we are interested in constructing an orthonormal set of vectors from a spanning set.*

In the case of (c), suppose that the two lines are given by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$, and $\mathbf{x}_2(t) = \mathbf{x}_2 + t\mathbf{v}_2$. Let $f(s,t) = \|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|^2$ be the squared distance between $\mathbf{x}_1(s)$ and $\mathbf{x}_2(t)$. Observe that $f(s,t) = \|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|^2 = \|\mathbf{x}_1 - \mathbf{x}_2 + s\mathbf{v}_1 - t\mathbf{v}_2\|^2$ is the squared distance from $\mathbf{x}_2 - \mathbf{x}_1$ to $s\mathbf{v}_1 - t\mathbf{v}_2$ in the plane spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$ (assuming we are in the case that $\mathbf{v}_1$ and $\mathbf{v}_2$ are not parallel to each other; otherwise, we would be looking at the distance from $\mathbf{x}_2 - \mathbf{x}_1$ to a line). Thus the minimum possible value of $f(s,t)$ is the squared distance from $\mathbf{x}_2 - \mathbf{x}_1$ to the plane spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$, for which we already have a solution from case (b).

In $\mathbb{R}^3$, we can take the unit vector $\mathbf{n}$ in the direction of $\mathbf{v}_1 \times \mathbf{v}_2$ as the normal to the plane, and carry out the analysis. The distance is found from (b) to be $|(\mathbf{x}_2 - \mathbf{x}_1) \cdot \left(\frac{\mathbf{v}_1 \times \mathbf{v}_2}{\|\mathbf{v}_1 \times \mathbf{v}_2\|}\right)|$. But the orthogonality criterion (1.2) for the solution of the distance problem applies in any $\mathbb{R}^n$.

Professor Carlen's approach, in particular his (1.57), is a more algebraic way of describing the orthogonal projection approach as explained for case (b). We will discuss another solution after having discussed the Gram-Schmidt algorithm to produce an orthonormal set of vectors $\{\mathbf{u}_1, \mathbf{u}_2\}$ which produces the same span as $\{\mathbf{v}_1, \mathbf{v}_2\}$; that method will work for higher dimensional problems, which arise from many applications.

Example 1.2.1

Consider two lines parametrized by $\mathbf{x}_1(s) = (1,2,3,4) + s(1,4,5,6)$, and $\mathbf{x}_2(t) = (2,-1,1,3) + t(-2,-1,1,0)$. The formula in terms of cross product is not applicable here. To find the distance between these lines, we need to find s_* and t_* such that

$$\begin{cases} (1,4,5,6) \cdot [(1,2,3,4) - (2,-1,1,3) + s_*(1,4,5,6) - t_*(-2,-1,1,0)] = 0, \\ (-2,-1,1,0) \cdot [(1,2,3,4) - (2,-1,1,3) + s_*(1,4,5,6) - t_*(-2,-1,1,0)] = 0. \end{cases}$$

This amounts to

$$\begin{cases} 78s_* + t_* = -26, \\ -s_* - 6t_* = 1. \end{cases}$$

This system has a unique solution, which is used to compute the distance between these two lines.

Reading Quizzes/Questions: The discussions so far mostly focus on finding the shortest distance. Do the method also give the point(s) which attains the shortest distance? Is there an efficient way to find the shortest distance without having to compute the point(s) which attains the shortest distance?

Summary: A common theme for these distance problems is to find the minimum of

$$\|s_1\mathbf{v}_1 + \cdots + s_k\mathbf{v}_k - \mathbf{q}\| \text{ among } s_1, \cdots, s_k \in \mathbb{R},$$

where $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a set of vectors in $\mathbb{R}^n$, and $\mathbf{q}$ is some vector in $\mathbb{R}^n$. If the minimum is attained at some $(s_1^*, \cdots, s_k^*)$, then we still have the orthogonality criterion:

$$(s_1^*\mathbf{v}_1 + \cdots + s_k^*\mathbf{v}_k - \mathbf{q}) \cdot \mathbf{v}_j = 0, 1 \le j \le k.$$

We will discuss the theory for solving such system of linear equations later on.

But there is a simpler solution if $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a subset of an orthonormal basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ for $\mathbb{R}^n$. For then, by **Theorem 6** we can write

$$\mathbf{q} = q_1\mathbf{v}_1 + \cdots + q_n\mathbf{v}_n$$

for some coefficients $q_1, \cdots, q_n$, and

$$s_1\mathbf{v}_1 + \cdots + s_k\mathbf{v}_k - \mathbf{q} = (s_1 - q_1)\mathbf{v}_1 + \cdots + (s_k - q_k)\mathbf{v}_k - q_{k+1}\mathbf{v}_{k+1} - \cdots - q_n\mathbf{v}_n,$$

and by the Pythagorean Theorem,

$$\begin{aligned} &\|s_1\mathbf{v}_1 + \cdots + s_k\mathbf{v}_k - \mathbf{q}\|^2 \\ =&(s_1 - q_1)^2 + \cdots + (s_k - q_k)^2 + q_{k+1}^2 + \cdots + q_n^2 \\ \ge&q_{k+1}^2 + \cdots + q_n^2 \end{aligned}$$

with equality if and only if $s_1 - q_1 = \cdots = s_k - q_k = 0$. Thus the minimum distance squared is

$$q_{k+1}^2 + \cdots + q_n^2 = (\mathbf{q} \cdot \mathbf{v}_{k+1})^2 + \cdots + (\mathbf{q} \cdot \mathbf{v}_n)^2,$$

which can be computed directly from $\mathbf{q}$ and $\{\mathbf{v}_{k+1}, \cdots, \mathbf{v}_n\}$ without having to solve any additional equations.

For the case of $n = 3$ and $k = 2$, we just take $\mathbf{v}_3$ to be a unit normal to the plane without having to work out $\mathbf{v}_1, \mathbf{v}_2$, while for the case of $n = 3$ and $k = 1$, the above discussion relies on the construction of some orthonormal $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, but our earlier solution took advantage that $\mathbf{v}_1$ can be taken as a unit direction vector of the line, and without knowing $\mathbf{v}_2, \mathbf{v}_3$, we can still obtain

$$q_2^2 + q_3^2 = \|\mathbf{q}\|^2 - q_1^2 = \|\mathbf{q}\|^2 - (\mathbf{q} \cdot \mathbf{v}_1)^2 .$$

In any case the above discussion motivates the need for constructing appropriate set of orthonormal basis of $\mathbb{R}^n$, which the Gram-Schmidt Orthonormalization Algorithm addresses.

1.3 The Gram-Schmidt Orthonormalization Algorithm

Theorem 2 describes the fundamental property of the standard basis, and Theorem 6 describes a key property of orthonormal set. Note that a set of orthonormal vectors in the setting of Theorem 6 is as useful as the standard basis, as illustrated in proving the Lagrange identity, and in several arguments of the previous section in Carlen's notes. The Gram-Schmidt Orthonormalization Algorithm is a tool used to construct such a set of orthonormal vectors from a given set of vectors.

In the words of Professor Carlen, the algorithm is used to extract a maximal orthonormal set from any collection of m vectors in $\mathbb{R}^n$ for arbitrary m and n which will have certain relations with the given collection of vectors.

1.3.1 The Gram-Schmidt Orthonormalization Algorithm in $\mathbb{R}^3$

Actually the algorithm here works for $m = 2$ vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$ in $\mathbb{R}^n$ for any n. We first assume $\mathbf{v}_1 \neq \mathbf{0}$. We can set $\mathbf{u}_1$ to be the unit vector in the direction of $\mathbf{v}_1$: $\mathbf{u}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$. Let $\mathbf{w}_2 = \mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1$ be the orthogonal component of $\mathbf{v}_2$ with respect to $\mathbf{u}_1$. Then $\mathbf{w}_2 \cdot \mathbf{u}_1 = 0$.

If $\mathbf{w}_2 \neq \mathbf{0}$, then we can set $\mathbf{u}_2$ to be the unit vector in the direction of $\mathbf{w}_2$: $\mathbf{u}_2 = \mathbf{w}_2/\|\mathbf{w}_2\|$, then $\mathbf{u}_2 \cdot \mathbf{u}_1 = 0$, and $\{\mathbf{u}_1, \mathbf{u}_2\}$ forms an orthonormal set such that $\mathrm{Span}(\mathbf{v}_1, \mathbf{v}_2) = \mathrm{Span}(\mathbf{u}_1, \mathbf{u}_2)$.

If $\mathbf{w}_2 = \mathbf{0}$, then $\mathbf{v}_2 = (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1$, which also means that $\mathbf{v}_2$ is a scalar multiple of $\mathbf{v}_1$.

The first two steps can be best summarized as

$$\begin{cases} \mathbf{v}_1 = \|\mathbf{v}_1\|\mathbf{u}_1, \\ \mathbf{w}_2 = \mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1, \\ \mathbf{u}_2 = \mathbf{w}_2/\|\mathbf{w}_2\|, \text{ when } \mathbf{w}_2 \neq \mathbf{0}. \end{cases} \tag{1.3}$$

When this algorithm can be carried out, the relations between $\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{u}_1, \mathbf{u}_2\}$ can be best summarized as

$$\begin{cases} \mathbf{v}_1 = \|\mathbf{v}_1\|\mathbf{u}_1, \\ \mathbf{v}_2 = (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1 + \|\mathbf{w}_2\|\mathbf{u}_2. \end{cases} \tag{1.4}$$

Here the given vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$ are written on the left, and the newly constructed vectors $\{\mathbf{u}_1, \mathbf{u}_2\}$ ($\mathbf{u}_2$ can be constructed under the condition that $\mathbf{v}_2$ is not a scalar multiple of $\mathbf{v}_1$), are written on the right.

1.3.2 The Gram-Schmidt Orthonormalization Algorithm in general

This procedure can be carried out for a set of m vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$. The simplest case is when no $\mathbf{v}_j$ is a linear combination of the vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}$ for $j = 1, \ldots, m$; this implies that none of $\mathbf{v}_j$ is $\mathbf{0}$. We still take $\mathbf{u}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$. Then we construct $\mathbf{w}_2$ as above and know that $\mathbf{w}_2 \neq \mathbf{0}$, so can construct $\mathbf{u}_2$ as above.

Now let

$$\mathbf{w}_3 = \mathbf{v}_3 - (\mathbf{v}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{v}_3 \cdot \mathbf{u}_2)\mathbf{u}_2.$$

Then $\mathbf{w}_3 \cdot \mathbf{u}_1 = \mathbf{w}_3 \cdot \mathbf{u}_2 = 0$. Note that $P(\mathbf{v}_3) := (\mathbf{v}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{v}_3 \cdot \mathbf{u}_2)\mathbf{u}_2 \in \text{Span}\{\mathbf{u}_1, \mathbf{u}_2\} = \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$ is the "orthogonal projection" of $\mathbf{v}_3$ in the plane $\text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$, characterized by the two conditions:

(a) $P(\mathbf{v}_3) \in \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$;

(b) $\mathbf{v}_3 - P(\mathbf{v}_3) \perp \mathbf{v}$ for any $\mathbf{v} \in \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$.

Remark 1.3.1

When this orthogonal projection process was carried out to $\{\mathbf{v}_1, \mathbf{v}_2\}$ to produce $\mathbf{w}_2 = \mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1$, the orthogonal component of $\mathbf{v}_2$ with respect to $\mathbf{u}_1$, the parallel component $(\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1$ is the orthogonal projection of $\mathbf{v}_2$ in Span$\{\mathbf{v}_1\}$. Notice the different usage of the terminology orthogonal component and or-

thogonal projection.

Note also that $P(\mathbf{v}_3)$ is constructed using the newly constructed $\mathbf{u}_1, \mathbf{u}_2$, instead of $\mathbf{v}_1, \mathbf{v}_2$.

Under our assumption, we know $\mathbf{w}_3 \neq \mathbf{0}$, as, otherwise, $\mathbf{v}_3$ would be a linear combination of $\{\mathbf{u}_1, \mathbf{u}_2\}$, which, in turn, becomes a linear combination of $\{\mathbf{v}_1, \mathbf{v}_2\}$. So we can set $\mathbf{u}_3$ to be the unit vector in the direction of $\mathbf{w}_3$, and the relation between $\mathbf{v}_1, \mathbf{v}_2$ and $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ becomes

$$\mathbf{v}_3 = (\mathbf{v}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{v}_3 \cdot \mathbf{u}_2)\mathbf{u}_2 + \|\mathbf{w}_3\|\mathbf{u}_3.$$

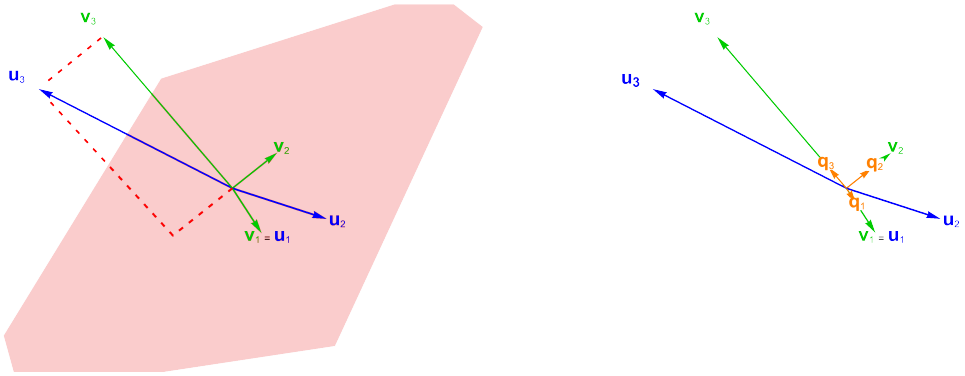


Figure 1.1: Final Outcome of the Gram-Schmidt Orthogonalization of $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. Image from the Wolfram Demonstrations Project due to Abby Brown. The code here labels the given vectors as $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, and labels the constructed orthogonal vectors as $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, and the normalized ones as $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\}$, respectively.

Reading Quizzes/Questions: Can you identify the orthogonal projection of $\mathbf{u}_3$ into the span of $\{\mathbf{u}_1, \mathbf{u}_2\}$ from the figure above?

This procedure can be carried out recursively for any $j \leq m$, namely, assuming that $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{j-1}\}$ has been constructed according to the rule that for each $1 \leq i \leq j-1$,

$$\mathbf{w}_i = \mathbf{v}_i - [(\mathbf{v}_i \cdot \mathbf{u}_1)\mathbf{u}_1 + \ldots + (\mathbf{v}_i \cdot \mathbf{u}_{i-1})\mathbf{u}_{i-1}], \tag{1.5}$$

and set $\mathbf{u}_i$ to be the unit vector in the direction of $\mathbf{w}_i$, which is not $\mathbf{0}$ under our assumption, then this relation is rewritten as

$$\mathbf{v}_i = (\mathbf{v}_i \cdot \mathbf{u}_1)\mathbf{u}_1 + \ldots + (\mathbf{v}_i \cdot \mathbf{u}_{i-1})\mathbf{u}_{i-1} + \|\mathbf{w}_i\|\mathbf{u}_i; \tag{1.6}$$

further, we set

$$\mathbf{w}_j = \mathbf{v}_j - [(\mathbf{v}_i \cdot \mathbf{u}_1)\mathbf{u}_1 + \ldots + (\mathbf{v}_i \cdot \mathbf{u}_{j-1})\mathbf{u}_{j-1}],$$

then $\mathbf{w}_j \neq \mathbf{0}$, and we set $\mathbf{u}_j$ to be the unit vector in the direction of $\mathbf{w}_j$. The outcome is a collection of orthonormal vectors $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ such that, for any $1 \leq l \leq m$, $\text{Span}(\{\mathbf{u}_1, \cdots, \mathbf{u}_l\}) = \text{Span}(\{\mathbf{v}_1, \cdots, \mathbf{v}_l\})$.

If we don't make the assumption that no $\mathbf{v}_j$ is a linear combination of the vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}$ for $j = 1, \ldots, m$, then it is possible that some $\mathbf{w}_i$ becomes $\mathbf{0}$ in this process. When this happens, we drop this $\mathbf{v}_i$ as a non-pivotal vector, as $\mathbf{w}_i = \mathbf{0}$ meant that $\mathbf{v}_i$ can be written as a linear combination of the vectors ahead of it in the list, and we simply move on to the next vector in the list to repeat this procedure.

In the end we end up with certain r indices $p_1 < p_2 < \ldots < p_r$ such that $\mathbf{w}_i \neq \mathbf{0}$ only when $i = p_l$ for some $1 \leq l \leq r$, and obtain a set of r orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ based on $\{\mathbf{w}_{p_1}, \ldots, \mathbf{w}_{p_r}\}$. The corresponding vectors $\{\mathbf{v}_{p_1}, \ldots, \mathbf{v}_{p_r}\}$ are called pivotal vectors. For any $j \notin \{p_1, p_2, \ldots, p_r\}$, assume $p_i < j < p_{i+1}$ for some i, then $\mathbf{v}_j$ is a linear combination of $\{\mathbf{v}_{p_1}, \ldots, \mathbf{v}_{p_i}\}$. We still have for any $1 \leq l \leq r$, $\text{Span}(\{\mathbf{u}_1, \cdots, \mathbf{u}_l\}) = \text{Span}(\{\mathbf{v}_{p_1}, \cdots, \mathbf{v}_{p_l}\})$.

Remark 1.3.2

Note that the number r of pivotal vectors among $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ is the same as the number of orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ that this Gram-Schmidt procedure has produced. Carlen's approach focuses on the so-called orthonormal basis such as $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$; in standard, more leisurely discussions in linear algebra, one also works with the pivotal vectors $\{\mathbf{v}_{p_1}, \ldots, \mathbf{v}_{p_r}\}$ as a basis.

Example 23 illustrates this algorithm. Take the given vectors there: $\mathbf{v}_1 = (1, 2, 3), \mathbf{v}_2 = (1, 2, 1), \mathbf{v}_3 = (2, 1, 1)$, and $\mathbf{v}_4 = (0, 1, 1)$. The algorithm finds that $p_1 = 1$, $p_2 = 2$, $\mathbf{v}_3$ non-pivotal, and $p_3 = 4$. Furthermore $\mathbf{u}_1 = \frac{1}{\sqrt{14}}(1, 2, 3)$, $\mathbf{u}_2 = \frac{1}{\sqrt{142}}(5, 4, 1)$, and $\mathbf{u}_3 = \frac{1}{\sqrt{3}}(1, 1, 1)$. The computations in the Gram-Schmidt algorithm give the relations between $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ and $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ as follows

$$\begin{cases} \mathbf{v}_1 = \sqrt{14}\mathbf{u}_1, \\ \mathbf{v}_2 = \frac{6}{\sqrt{14}}\mathbf{u}_1 + \frac{2\sqrt{42}}{7}\mathbf{u}_2, \\ \mathbf{v}_3 = \frac{3}{\sqrt{14}}\mathbf{u}_1 + \frac{15}{\sqrt{42}}\mathbf{u}_2, \\ \mathbf{v}_4 = \frac{1}{\sqrt{14}}\mathbf{u}_1 + \frac{5}{\sqrt{42}}\mathbf{u}_2 + \frac{2}{\sqrt{3}}\mathbf{u}_3. \end{cases}$$

The coefficients may look complicated to the human eyes, but this orthogonalization process turns out to be very useful in many contexts. For instance, if we need to solve x_1, x_2, x_4 such that $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_4\mathbf{v}_4 = \mathbf{f}$ for arbitrarily given $\mathbf{f}$, then using the above relations, it is equivalent to solving

$$x_1\left[\sqrt{14}\mathbf{u}_1\right] + x_2\left[\frac{6}{\sqrt{14}}\mathbf{u}_1 + \frac{2\sqrt{42}}{7}\mathbf{u}_2\right] + x_4\left[\frac{1}{\sqrt{14}}\mathbf{u}_1 + \frac{5}{\sqrt{42}}\mathbf{u}_2 + \frac{2}{\sqrt{3}}\mathbf{u}_3\right] = \mathbf{f},$$

which in turn is equivalent to

$$\left[\sqrt{14}x_1 + \frac{6}{\sqrt{14}}x_2 + \frac{1}{\sqrt{14}}x_4\right]\mathbf{u}_1 + \left[\frac{2\sqrt{42}}{7}x_2 + \frac{5}{\sqrt{42}}x_4\right]\mathbf{u}_2 + \left[\frac{2}{\sqrt{3}}x_4\right]\mathbf{u}_3 = \mathbf{f}.$$

Using the orthonormal property of $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and Theorem 6, we find that we must have

$$\begin{cases} \sqrt{14}x_1 + \dfrac{6}{\sqrt{14}}x_2 + \dfrac{1}{\sqrt{14}}x_4 = \mathbf{u}_1 \cdot \mathbf{f}, \\ \dfrac{2\sqrt{42}}{7}x_2 + \dfrac{5}{\sqrt{42}}x_4 = \mathbf{u}_2 \cdot \mathbf{f}, \\ \dfrac{2}{\sqrt{3}}x_4 = \mathbf{u}_3 \cdot \mathbf{f}. \end{cases}$$

For any given $\mathbf{f}$, we can first solve x_4 from the last equation, then substitute it into the equation above to solve for x_2, and finally substitute both x_2 and x_4 into the first equation to find x_1. One advantage of this approach* is that once the Gram-Schmidt orthogonalization is carried out, the result can be used for arbitrary $\mathbf{f}$; if one applies an elimination of variables method to solve $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_4\mathbf{v}_4 = \mathbf{f}$, then one has to repeat the elimination process for each $\mathbf{f}$.

If we include $\mathbf{v}_3$ into consideration and ask to solve $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_3\mathbf{v}_3 + x_4\mathbf{v}_4 = \mathbf{f}$, then we can pick any value of x_3 and solve $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_4\mathbf{v}_4 = \mathbf{f} - x_3\mathbf{v}_3$ as above. This leads to the above procedure is modified into

$$\begin{cases} \sqrt{14}x_1 + \dfrac{6}{\sqrt{14}}x_2 + \dfrac{1}{\sqrt{14}}x_4 = \mathbf{u}_1 \cdot (\mathbf{f} - x_3\mathbf{v}_3) & = \mathbf{u}_1 \cdot \mathbf{f} - \dfrac{3}{\sqrt{14}}x_3, \\ \dfrac{2\sqrt{42}}{7}x_2 + \dfrac{5}{\sqrt{42}}x_4 = \mathbf{u}_2 \cdot (\mathbf{f} - x_3\mathbf{v}_3) & = \mathbf{u}_2 \cdot \mathbf{f} - \dfrac{15}{\sqrt{42}}x_3, \\ \dfrac{2}{\sqrt{3}}x_4 = \mathbf{u}_3 \cdot (\mathbf{f} - x_3\mathbf{v}_3) & = \mathbf{u}_3 \cdot \mathbf{f}, \end{cases}$$

*After we introduce matrices and multiplication between matrices, this set up can be written in a compact matrix form, called the QR factorization.

using $\mathbf{u}_3 \cdot \mathbf{v}_3 = 0$. After solving for x_4 from the last equation, we can substitute it into the second equation, solve for x_2 in terms of x_3 and the value of x_4 from the last equation. The same is done to solve for x_1. In other words, x_3 is a *free variable* in this solution process. There is a general pattern: all variables corresponding to the non-pivotal vectors are free variables, and the variables corresponding to the pivotal vectors are solved in terms of them.

The outcome of the Gram-Schmidt algorithm can also be used in solving the problem of distance from a point to a plane. For instance, consider the plane Π spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$ above. In parametric form it is given by $\mathbf{x} = s\mathbf{v}_1 + t\mathbf{v}_2$. Suppose we need to find the distance from $\mathbf{p} = (0, 1, 1)$ to Π. Then similar to our discussion in the previous subsection in obtaining (†) and (‡), we need to find s and t such that

$$\begin{cases} [s\mathbf{v}_1 + t\mathbf{v}_2 - \mathbf{p}] \cdot \mathbf{v}_1 = 0, \\ [s\mathbf{v}_1 + t\mathbf{v}_2 - \mathbf{p}] \cdot \mathbf{v}_2 = 0. \end{cases}$$

But $\text{Span}(\mathbf{v}_1, \mathbf{v}_2) = \text{Span}(\mathbf{u}_1, \mathbf{u}_2)$, so $s\mathbf{v}_1 + t\mathbf{v}_2$ can also be written as $x\mathbf{u}_1 + y\mathbf{u}_2$ for some x and y, and in terms of this formulation, we would need to solve for x and y such that

$$\begin{cases} [x\mathbf{u}_1 + y\mathbf{u}_2 - \mathbf{p}] \cdot \mathbf{u}_1 = 0, \\ [x\mathbf{u}_1 + y\mathbf{u}_2 - \mathbf{p}] \cdot \mathbf{u}_2 = 0. \end{cases}$$

Recall the characterization of the orthogonal projection of $\mathbf{p}$ in Π, which identifies $x\mathbf{u}_1 + y\mathbf{u}_2$ as that projection, so

$$x\mathbf{u}_1 + y\mathbf{u}_2 = (\mathbf{p} \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{p} \cdot \mathbf{u}_2)\mathbf{u}_2.$$

This can also be obtained directly using the orthonormality of $\{\mathbf{u}_1, \mathbf{u}_2\}$: the above system can be solved directly from

$$\begin{cases} x - \mathbf{p} \cdot \mathbf{u}_1 = 0, \\ y - \mathbf{p} \cdot \mathbf{u}_2 = 0. \end{cases}$$

This leads to $x = \frac{1}{\sqrt{14}}$ and $y = \frac{5}{\sqrt{14}}$. Thus $\frac{1}{\sqrt{14}}\mathbf{u}_1 + \frac{5}{\sqrt{14}}\mathbf{u}_2$ is the point on Π closest to $\mathbf{p}$, and the distance from $\mathbf{p}$ to Π is $\|\mathbf{p} - \left[\frac{1}{\sqrt{14}}\mathbf{u}_1 + \frac{5}{\sqrt{14}}\mathbf{u}_2\right]\|$. We took $\mathbf{p}$ to be the same as $\mathbf{v}_4$ in Example 23, and it's clear from the computations using the Gram-Schmidt algorithm that $\mathbf{p} - \left[\frac{1}{\sqrt{14}}\mathbf{u}_1 + \frac{5}{\sqrt{14}}\mathbf{u}_2\right] = \frac{2}{\sqrt{3}}\mathbf{u}_3$, thus the distance from $\mathbf{p}$ to Π is $\frac{2}{\sqrt{3}}$.

If we are specifically interested in the values of s and t which gives the closes point

in Π to $\mathbf{p}$, then we can modify the procedure to get

$$\begin{cases} [s\mathbf{v}_1 + t\mathbf{v}_2 - \mathbf{p}] \cdot \mathbf{u}_1 = s(\mathbf{v}_1 \cdot \mathbf{u}_1) + t(\mathbf{v}_2 \cdot \mathbf{u}_1) - \mathbf{p} \cdot \mathbf{u}_1 & = \sqrt{14}s + \dfrac{6}{\sqrt{14}}t - \dfrac{1}{\sqrt{14}} = 0, \\ [s\mathbf{v}_1 + t\mathbf{v}_2 - \mathbf{p}] \cdot \mathbf{u}_2 = s(\mathbf{v}_1 \cdot \mathbf{u}_2) + t(\mathbf{v}_2 \cdot \mathbf{u}_2) - \mathbf{p} \cdot \mathbf{u}_2 & = \dfrac{2\sqrt{42}}{7}t - \dfrac{5}{\sqrt{42}} = 0. \end{cases}$$

We can now solve for t from the last equation, and substitute it into the first equation to solve for s.

Remark 1.3.3

Although we formulated the problem as a geometric problem of finding the distance from a point to a plane (or distance between two lines or two planes), the underlying mathematics shows up in many applications. They typically show up as a problem of least squares. For instance, one may deal with three variables X, Y, Z in a certain modeling problem, and believes that there are coefficients s and t such that $sX + tY$ would predict the value of Z well, perhaps not exactly. One runs a series of n observations to obtain data points (X_i, Y_i, Z_i) for $i = 1, \ldots, n$, and would like to use the data points to find values of s and t which would give the "best prediction", namely, to make $\sum_{i=1}^{n}(sX_i + tY_i - Z_i)^2$ the smallest. Setting $\hat{X} = (X_1, \ldots, X_n)$, $\hat{Y} = (Y_1, \ldots, Y_n)$, and $\hat{Z} = (Z_1, \ldots, Z_n)$, the problem is the same as finding s and t which minimize $\|s\hat{X} + t\hat{Y} - \hat{Z}\|^2$. But $\|s\hat{X} + t\hat{Y} - \hat{Z}\|$ is the distance from $\hat{Z}$ to the point $s\hat{X} + t\hat{Y}$ in the plane spanned by $\hat{X}$ and $\hat{Y}$. So we are really solving a problem of finding the distance from a point $\hat{Z} \in \mathbb{R}^n$ to a plane.

1.3.3 Subspaces of $\mathbb{R}^n$

After introducing the definition of a *subspace* of $\mathbb{R}^n$, the focus turns to finding a set of orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ in a subspace V which is not $\{\mathbf{0}\}$, such that $V = \text{Span}(\{\mathbf{u}_1, \ldots, \mathbf{u}_r\})$, using the Gram-Schmidt algorithm.

The concept of dimension of a subspace is based on the following property*. Suppose that V a subspace which is not $\{\mathbf{0}\}$, and $V = \text{Span}(\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}) = \text{Span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_s\})$ for two sets of orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_s\}$, then $r = s$. This number r is then called the dimension of V, and each set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_s\}$ is called an orthonormal basis of V. This is a generalization of Theorem 6; it also

*The approach here is different from those in standard linear algebra textbooks, which typically develops properties of bases of a subspace which are not necessarily orthonormal, and uses those properties to define the notion of dimension. The approach here relies only on the notion of orthonormal bases.

implies that there does not exist a set of m orthonormal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ in $\mathbb{R}^n$ for some $m < n$ such that $\text{Span}(\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}) = \mathbb{R}^n$—This is part of Theorem 17.

Here is a sketch of proof of the above statement in more plain language. Since each $\mathbf{v} \in V$ can be written as $\mathbf{v} = (\mathbf{v} \cdot \mathbf{u}_1)\mathbf{u}_1 + \ldots + (\mathbf{v} \cdot \mathbf{u}_r)\mathbf{u}_r$, we map each $\mathbf{v}$ to its coordinate vector $(\mathbf{v} \cdot \mathbf{u}_1, \ldots, \mathbf{v} \cdot \mathbf{u}_r) \in \mathbb{R}^r$. This is called $C(\mathbf{v})$ in Professor Carlen's notes. The key property is (1.76), $C(\mathbf{v}) \cdot C(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$ for all $v, w \in V$. It follows from this that $\{C(\mathbf{v}_1), \ldots, C(\mathbf{v}_s)\}$ is orthonormal in $\mathbb{R}^r$. But according to Lemma 1, $\mathbb{R}^r$ can't have more than r orthonormal vectors. Thus $s \leq r$. Reversing the role of $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_s\}$, we see that $r \leq s$, and conclude that $r = s$.

Reading Quizzes/Questions:

1. Is the following statement correct? *The set of vectors in any line in $\mathbb{R}^3$ forms a subspace.*

2. Edit the following statement to make it correct: *Any two sets of orthonormal vectors in a subspace V have the same number of vectors.*

1.3.4 Orthogonal Complements

We did not have time to cover this section in the lecture. A line through the origin in $\mathbb{R}^2$ has a one dimensional set of vectors as its normal, but a line in $\mathbb{R}^3$ does not have a one dimensional set of vectors as its normal; its normals form a two dimensional plane. Likewise, a plane in $\mathbb{R}^3$ has a one dimensional set of vectors as its normal, but a plane in $\mathbb{R}^n$, $n > 3$, has an $n - 2$ dimensional set of vectors as its normal. The notion of orthogonal compliment is used to characterize these properties.

Reading Quizzes/Questions: Is it true that $S^\perp = (\text{Span}(S))^\perp$?

1.3.5 Higher dimensional analogs of lines and planes

Suppose that V is an r-dimensional subspace of $\mathbb{R}^n$, then it is an analog of a line or plane through the origin in $\mathbb{R}^3$. If we translate V by a fixed vector $\mathbf{x}_0$, namely, if we define $W = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{x}_0 + \mathbf{v} \text{ for some } \mathbf{v} \in V\}$, then we say that W is an affine space of dimension r. W is not necessarily a subspace of $\mathbb{R}^n$, because, for two vectors $\mathbf{x}, \mathbf{y} \in W$, $\mathbf{x} + \mathbf{y}$ may no longer be in W; but we always have $\mathbf{x} - \mathbf{y} \in V$, namely, the

displacement between any two vectors in W is along a vector in V. An affine space is an analog of the a line or plane in $\mathbb{R}^3$ not necessarily passing through the origin.

Such affine spaces arise when we solve systems of linear equations. The set of solutions of a given system of linear equation forms an affine space.

Chapter 2

Description of Motion

2.1 Functions from $\mathbb{R}$ to $\mathbb{R}^n$ and the description of motion

2.1.1 Continuity of functions from $\mathbb{R}$ to $\mathbb{R}^n$

The underlying mathematics of this chapter is the analysis of vector valued functions of a single variable, $t \mapsto \mathbf{x}(t) = (x_1(t), \ldots, x_n(t))$, $t \in (a, b) \subset \mathbb{R}^n$. This is done in **2.1.1**. Most of the analysis is done as for a real valued function of a single variable, with little difference.

For instance, the continuity of $\mathbf{x}(t)$ at $t = t_0$ requires that for any $\epsilon > 0$, there is a real number $\delta_\epsilon > 0$ such that

$$|t - t_0| < \delta_\epsilon \Longrightarrow \|\mathbf{x}(t) - \mathbf{x}(t_0)\| < \epsilon.$$

(Professor Carlen's notes use $\leq$ instead of $<$ in (2.2); convince yourself that the two formulations are equivalent.) Since

$$|x_i(t) - x_i(t_0)| \leq \|\mathbf{x}(t) - \mathbf{x}(t_0)\| = \sqrt{\sum_{j=1}^{n} |x_j(t) - x_j(t_0)|^2} \leq \sqrt{n} \max_{1 \leq j \leq n} |x_j(t) - x_j(t_0)|,$$

it follows from this that the continuity of $\mathbf{x}(t)$ at $t = t_0$ is equivalent to the continuity of each of its coordinate function $t \mapsto x_i(t)$ at $t = t_0$.

Here are some details for the more subtle direction: suppose that each of the coordinate functions $x_i(t)$ is continuous at $t = t_0$, and we need to show that $\mathbf{x}(t)$ is continuous at $t = t_0$ (Reflect on why this direction is considered more subtle). Using the continuity of $x_i(t)$ at $t = t_0$, we find, for a given $\epsilon > 0$, some $\delta_{\epsilon,i} > 0$, such that

when $|t-t_0| < \delta_{\epsilon,i}$, we have $|x(t)-x(t_i)| < \epsilon/\sqrt{n}$. Since there are only a finite number of such $i : 1 \le i \le n$, let $\delta_\epsilon = \min_{1\le i\le n} \delta_{\epsilon,i}$, then $\delta_\epsilon > 0$, and when $|t-t_0| < \delta_\epsilon$, we have $|x(t)-x(t_i)| < \epsilon$ for all $1 \le i \le n$, therefore $\|\mathbf{x}(t)-\mathbf{x}(t_0)\| < \epsilon$. This shows that continuity of $\mathbf{x}(t)$ at $t = t_0$.

Remark 2.1.1

When we study the continuity at a point of a function (either scalar or vector valued) of more than one variables, we will see that it is different from the continuity at that point of its restriction to any one-dimensional line through that point. For example, take

$$f(x,y) = \begin{cases} \frac{x^2y}{x^4+y^2} & (x,y) \neq (0,0), \\ 0 & (x,y) = (0,0), \end{cases}$$

then its restriction along the one-dimensional line $(x,y) = t(u,v)$, for some $(u,v) \neq (0,0)$, becomes a function of a single variable t as

$$g(t) = \begin{cases} \frac{tu^2v}{t^2u^4+v^2} & t \neq 0, \\ 0 & t = 0. \end{cases}$$

This is obviously a continuous function of t. But the continuity of f at $(0,0)$ would require

$$|f(x,y) - f(0,0)| = \frac{|x^2y|}{x^4+y^2} < \epsilon$$

for all (x,y) such that $\|(x,y)\| = \sqrt{x^2+y^2} < \delta_\epsilon$, where $\delta_\epsilon > 0$ is chosen depending on $\epsilon > 0$ and f. If we choose y according to $y = x^2$ we find

$$|f(x,x^2) - f(0,0)| = \frac{1}{2},$$

which can't satisfy $|f(x,x^2) - f(0,0)| < \epsilon$ if $\epsilon < \frac{1}{2}$, no matter how small $|x|$ is. This shows that this f is not continuous at $(0,0)$, even though its restriction to any line through $(0,0)$ is continuous. The reason is that there are infinitely many lines through $(0,0)$, and having control of the behavior along any such line, once it is fixed, is different from having control of the behavior in the two-dimensional neighborhood.

2.1.2 Differentiability of functions from $\mathbb{R}$ to $\mathbb{R}^n$

For differentiability of $\mathbf{x}(t)$ at $t = t_0$, we mimic the definition in one variable calculus: there exists a linear function of t of the form, $\mathbf{x}(t_0) + (t - t_0)\mathbf{v}$, for some vector $\mathbf{v}$, such that it approximates $\mathbf{x}(t)$ near $t = t_0$ in the sense that, the remainder

$$\mathbf{Rm}(t) := \mathbf{x}(t) - [\mathbf{x}(t_0) + (t - t_0)\mathbf{v}]$$

satisfies $\|\mathbf{Rm}(t)\|/|t - t_0| \to 0$ as $t \to t_0$. We summarize this as

$$\mathbf{x}(t) = \mathbf{x}(t_0) + (t - t_0)\mathbf{v} + \mathbf{Rm}(t); \quad \|\mathbf{Rm}(t)\|/|t - t_0| \to 0 \text{ as } t \to t_0.$$

In one variable calculus, this is often reformulated as defining the derivative through looking at the limit of the difference quotient $\dfrac{\mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0}$:

$$\left| \frac{\mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0} - \mathbf{v} \right| = \frac{\|\mathbf{Rm}(t)\|}{|t - t_0|} \to 0 \text{ as } t \to t_0.$$

As for the continuity of a vector-valued function of a single variable, $\mathbf{x}(t)$ is differentiable at $t = t_0$ if and only if each of its coordinate function $x_i(t)$ is differentiable at $t = t_0$, and $\mathbf{x}'(t_0) = (x_1'(t_0), \cdots, x_n'(t_0))$ in such a situation.

We will see soon that the differentiability of a function of more than one variables will need to be treated differently. Even for a real valued function f of $\mathbf{y} = (y_1, \ldots, y_n)$ defined in a neighborhood of $\mathbf{y}_0$, a linear function of $\mathbf{y}$ would take the form of $f_0 + a_1(y_1 - \mathbf{y}_{01}) + a_2(y_2 - \mathbf{y}_{02}) + \ldots + a_n(y_n - \mathbf{y}_{0n})$, so we would need to examine

$$Rm_f(\mathbf{y}) := f(\mathbf{y}) - [f_0 + a_1(y_1 - \mathbf{y}_{01}) + a_2(y_2 - \mathbf{y}_{02}) + \ldots + a_n(y_n - \mathbf{y}_{0n})]$$

as $\mathbf{y} \to \mathbf{y}_0$. But there is no good way of making sense of the difference quotient $\dfrac{f(\mathbf{y}) - f(\mathbf{y}_0)}{\mathbf{y} - \mathbf{y}_0}$; neither does it make sense to look at $\dfrac{Rm_f(\mathbf{y})}{\mathbf{y} - \mathbf{y}_0}$ as we can't divide by a vector. Also, $\mathbf{y} \to \mathbf{y}_0$ in *infinitely many directions*, and we may choose to let $\mathbf{y} \to \mathbf{y}_0$ only along the coordinate axis directions, namely, for each $1 \leq i \leq n$, we examine $\mathbf{y}$ such that its coordinates in the j axis for $j \neq i$ stay the same as those of $\mathbf{y}_0$, but $y_i \to \mathbf{y}_{0i}$—in the case $i = 1$, we would have $\mathbf{y} = (y_1, \mathbf{y}_{02}, \ldots, \mathbf{y}_{0n})$, then we will be looking at $\|\dfrac{Rm_f(\mathbf{y})}{y_i - \mathbf{y}_{0i}}\|$. This will give rise to the notion of *partial derivatives*. This is related to differentiability, but is not equivalent to differentiability. For (full) differentiability, what we will do, instead, is to require $\|Rm_f(\mathbf{y})\|/\|\mathbf{y} - \mathbf{y}_0\| \to 0$ as $\|\mathbf{y} - \mathbf{y}_0\| \to 0$.

The usual rules of differentiation (or derivatives) of functions of a single variable, such as for computing the derivatives of a sum or difference of two vector valued

differentiable functions, or dot product or cross product of two vector valued differentiable functions taking values in the same $\mathbb{R}^n$ (and $n = 3$ for the cross product) follow the usual rules, as given by (2.11)–(2.13). One rule not stated in **2.1.1** is a version of the chain rule.

Suppose that $\mathbf{x}(t)$ defined for $t \in (a, b)$ is differentiable at $t_0 \in (a, b)$, and $\tau \in (\alpha, \beta) \mapsto t = \phi(\tau) \in (a, b)$ is differentiable, with $\phi(\tau_0) = t_0$. Then the composition $\mathbf{x} \circ \phi(\tau)$ is differentiable at $\tau = \tau_0$, and $(\mathbf{x} \circ \phi)'(\tau_0) = \phi'(\tau_0)\mathbf{x}'(t_0)$; more appropriately,

$$\frac{d(\mathbf{x} \circ \phi)}{d\tau}(\tau_0) = \frac{d\phi}{d\tau}(\tau_0)\frac{d\mathbf{x}}{dt}(t_0).$$

A casual proof would examine

$$\frac{\mathbf{x} \circ \phi(\tau) - \mathbf{x} \circ \phi(\tau_0)}{\tau - \tau_0} = \frac{\mathbf{x}(\phi(\tau)) - \mathbf{x}(\phi(\tau_0))}{\phi(\tau) - \phi(\tau_0)} \frac{\phi(\tau) - \phi(\tau_0)}{\tau - \tau_0}.$$

But it is possible that $\phi(\tau) - \phi(\tau_0) = 0$ for some τ near τ_0, so the above analysis is flawed. A more careful proof goes by examining

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(t_0) + \mathbf{x}'(t_0)(t - t_0) + \mathbf{Rm}_{\mathbf{x}}(t), \quad \|\mathbf{Rm}_{\mathbf{x}}(t)\|/|t - t_0| \to 0, \text{ as } t - t_0 \to 0; \\ \phi(\tau) &= \phi(\tau_0) + \phi'(\tau_0)(\tau - \tau_0) + Rm_{\phi}(\tau), \quad |Rm_{\phi}(\tau)|/|\tau - \tau_0| \to 0, \text{ as } |\tau - \tau_0| \to 0; \end{aligned}$$

and substituting the second into the first to obtain

$$\begin{aligned} \mathbf{x} \circ \phi(\tau) &= \mathbf{x}(t_0) + \mathbf{x}'(t_0)(\phi(\tau) - \phi(\tau_0)) + \mathbf{Rm}_{\mathbf{x}}(\phi(\tau)) \\ &= \mathbf{x}(t_0) + \mathbf{x}'(t_0)(\phi'(\tau_0)(\tau - \tau_0) + Rm_{\phi}(\tau)) + \mathbf{Rm}_{\mathbf{x}}(\phi(\tau)) \\ &= \mathbf{x}(t_0) + \mathbf{x}'(t_0)\phi'(\tau_0)(\tau - \tau_0) + Rm_{\phi}(\tau)\mathbf{x}'(t_0) + \mathbf{Rm}_{\mathbf{x}}(\phi(\tau)). \end{aligned}$$

We see now that $\mathbf{x}(t_0) + \mathbf{x}'(t_0)\phi'(\tau_0)(\tau - \tau_0)$ is a linear function of τ, and that we should treat $\mathbf{Rm}_{\mathbf{x}\circ\phi}(\tau) := Rm_{\phi}(\tau)\mathbf{x}'(t_0) + \mathbf{Rm}_{\mathbf{x}}(\phi(\tau))$ as the remainder term, and ask whether we have $\|\mathbf{Rm}_{\mathbf{x}\circ\phi}(\tau)\|/|\tau - \tau_0| \to 0$ as $\tau \to \tau_0$. Since

$$\frac{\|\mathbf{Rm}_{\mathbf{x}\circ\phi}(\tau)\|}{|\tau - \tau_0|} \leq \frac{|Rm_{\phi}(\tau)|\|\mathbf{x}'(t_0)\|}{|\tau - \tau_0|} + \frac{\|\mathbf{Rm}_{\mathbf{x}}(\phi(\tau))\|}{|\tau - \tau_0|},$$

we see that for any $\epsilon > 0$, we can find $\delta_1 > 0$ such that

$$\frac{|Rm_{\phi}(\tau)|\|\mathbf{x}'(t_0)\|}{|\tau - \tau_0|} < \epsilon/2 \quad \text{for all } \tau \text{ such that } |\tau - \tau_0| < \delta_1;$$

and can find $\delta_2 > 0$ such that $\|\mathbf{Rm}_{\mathbf{x}}(t)\| \leq \frac{\epsilon}{2M}|t - t_0|$ for all t such that $|t - t_0| < \delta_2$, where $M > 0$ is chosen such that $M > |\phi'(\tau_0)| + 1$; and finally, using the differentiability of ϕ at $\tau = \tau_0$, can find $\delta_3 > 0$ such that $|\phi(\tau) - \phi(\tau_0)| < M|\tau - \tau_0|$ for all τ such that $|\tau - \tau_0| < \delta_3$. We now set $\delta = \min\{\delta_1, \delta_2/M, \delta_3\}$. Then $\delta > 0$, and for all τ such that $|\tau - \tau_0| < \delta$, we have $|\phi(\tau) - \phi(\tau_0)| < M|\tau - \tau_0| < \delta_2$, thus

$$\|\mathbf{Rm}_{\mathbf{x}}(\phi(\tau)\| \leq \frac{\epsilon}{2M}|\phi(\tau) - \phi(\tau_0)| \leq \frac{\epsilon}{2}|\tau - \tau_0|,$$

and finally

$$\frac{\|\mathbf{Rm}_{\mathbf{x}\circ\phi}(\tau))\|}{|\tau - \tau_0|} \leq \epsilon.$$

2.1.3 Velocity and acceleration

Some most commonly encountered vector valued functions are functions of a time variable. The remainder of this chapter carries out detailed analysis of such functions. If one treats the vector valued functions as representing the position $\mathbf{x}(t)$ of a particle as a function of time t, then one identifies $\mathbf{x}'(t)$ as the velocity, $\|\mathbf{x}'(t)\|$ as the speed, and $\mathbf{x}''(t)$ as the acceleration. If one is interested in the geometry of the path traced out by $t \mapsto \mathbf{x}(t)$, then the notion of curvature and torsion plays a prominent role if $\mathbf{x}(t)$ is $\mathbb{R}^3$-valued, and some regularity and non-deneracy conditions are assumed.

The two approaches above are related, but also differ. The curvature of the curve represented by a twice differentiable function $\mathbf{x}(t)$ should not depend on how fast a particle traverses along the curve; in other words, it should be independent of the reparameterization of the curve: if $t = \phi(\tau)$ is a (differentiable) reparameterization with $t_0 = \phi(\tau_0)$, then $\mathbf{x}(t)$ and $\mathbf{x} \circ \phi(\tau) := \mathbf{x}(\phi(\tau))$ should give the same curvature computed through $\mathbf{x}(t)$ at $t = t_0$ and $\mathbf{x} \circ \phi(\tau)$ at $\tau = \tau_0$ (discussed in **2.1.5**).

Due to time constraint, we will focus on a few main results. You would learn a lot by working through the examples worked out by Professor Carlen, but you need not be concerned with memorizing the many formulae discussed—you only need to understand well how the curvature and torsion are defined.

1. **Unit tangent vector, parallel and orthogonal components of acceleration $\mathbf{a}(t)$.**

   If the speed $v(t) := \|\mathbf{x}'(t)\| > 0$ at some t, then define $\mathbf{T}(t) := \mathbf{x}'(t)/\|\mathbf{x}'(t)\|$. $\mathbf{T}(t)$ is a unit vector, and

   $$\mathbf{x}'(t) = v(t)\mathbf{T}(t). \tag{2.1}$$

The guiding principle below will be to write all subsequent derivatives $\mathbf{x}''(t), \mathbf{x}'''(t)$, etc., in terms of a set of orthonormal frame adapted to the curve, with the first vector being $\mathbf{T}(t)$, and the second vector being the unit vector in the direction of the orthogonal component of $\mathbf{x}''(t)$ with respect to $\mathbf{x}'(t)$ obtained through the Gram-Schmidt algorithm applied to $\{\mathbf{x}'(t), \mathbf{x}''(t)\}$ when both are pivotal vectors.

Since $\mathbf{T}(t) \cdot \mathbf{T}(t) \equiv 1$, it follows by taking derivatives with respect to t on both sides that to get

$$2\mathbf{T}'(t) \cdot \mathbf{T}(t) = 0, \textit{ i.e., } \mathbf{T}'(t) \perp \mathbf{T}(t).$$

But $\mathbf{T}'(t)$ could be $\mathbf{0}$. If $\mathbf{T}'(t) \neq \mathbf{0}$, then set $\mathbf{N}(t) = \mathbf{T}'(t)/\|\mathbf{T}'(t)\|$, so

$$\mathbf{T}'(t) = \|\mathbf{T}'(t)\|\mathbf{N}(t).$$

We then define

$$\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t),$$

and call $\mathbf{N}(t)$ the principal normal vector to the curve at $\mathbf{x}(t)$, $\mathbf{B}(t)$ the binormal vector to the curve at $\mathbf{x}(t)$. We will express all vector quantities associated to the curve in terms of this set $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ of orthonormal vectors.

The first task is to look at the decomposition of $\mathbf{x}''(t)$ as its component $\mathbf{a}_\parallel$ parallel to $\mathbf{T}(t)$ (also parallel to $\mathbf{x}'(t)$), and its component $\mathbf{a}_\perp$ orthogonal to $\mathbf{T}(t)$:

$$\mathbf{x}''(t) = \mathbf{a}_\parallel + \mathbf{a}_\perp.$$

$\mathbf{a}_\parallel$ can be computed as $[\mathbf{x}''(t) \cdot \mathbf{T}(t)]\mathbf{T}(t)$, but we can also take derivative in (2.1) to obtain

$$\mathbf{a}(t) := \mathbf{x}''(t) = v'(t)\mathbf{T}(t) + v(t)\|\mathbf{T}'(t)\|\mathbf{N}(t). \tag{2.2}$$

Since $\mathbf{T}(t) \perp \mathbf{N}(t)$, we see that $v'(t)\mathbf{T}(t)$ is the parallel component $\mathbf{a}_\parallel$ of $\mathbf{a}(t)$ (along $\mathbf{T}(t)$), and $v(t)\mathbf{T}'(t) = v(t)\|\mathbf{T}'(t)\|\mathbf{N}(t)$ is the orthogonal component $\mathbf{a}_\perp$ of $\mathbf{a}(t)$.

To summarize,

$$\mathbf{a}_\parallel = v'(t)\mathbf{T}(t), \quad \mathbf{a}_\perp = v(t)\mathbf{T}'(t), \quad v'(t) = \mathbf{x}''(t) \cdot \mathbf{T}(t) = \frac{\mathbf{x}''(t) \cdot \mathbf{x}'(t)}{\|\mathbf{x}'(t)\|}. \tag{2.3}$$

Remark 2.1.2

From (2.2) we can see that the condition that $\mathbf{T}'(t) \neq \mathbf{0}$ is equivalent to the condition that $\mathbf{a}_\perp \neq \mathbf{0}$, which is also equivalent to the condition that $\mathbf{x}''(t)$ is not a scalar multiple of $\mathbf{x}'(t)$. Finally in $\mathbb{R}^3$ another equivalent condition is that $\mathbf{x}'(t) \times \mathbf{x}''(t) \neq \mathbf{0}$—this condition encodes both $\mathbf{x}'(t) \neq \mathbf{0}$ and $\mathbf{T}'(t) \neq \mathbf{0}$.

2. **Curvature.** When $v(t) = 1$, namely, when $\mathbf{x}(t)$ moves at a unit speed at t, $\|\mathbf{T}'(t)\|$ is called the curvature at $\mathbf{x}(t)$ of the curve represented by $t \mapsto \mathbf{x}(t)$, and is denoted as κ. It measures the rate of change of the unit direction of the curve per unit length along the curve; without the constraint $v(t) = 1$, $\|\mathbf{T}'(t)\|$ would measure the rate of change of the unit direction of the curve per unit time, so does not represent the "curviness" of the curve in space. For a general parametrized curve, the curvature should be defined as

$$\kappa := \frac{\|\mathbf{T}'(t)\|}{v(t)}. \tag{2.4}$$

Using (2.1), for a curve in $\mathbb{R}^3$, we have

$$\mathbf{x}'(t) \times \mathbf{x}''(t) = v(t)\|\mathbf{T}'(t)\|\mathbf{x}'(t) \times \mathbf{N}(t) = v(t)^2\|\mathbf{T}'(t)\|\mathbf{T}(t) \times \mathbf{N}(t), \tag{2.5}$$

and since $\mathbf{T}(t)$ and $\mathbf{N}(t)$ are unit vectors orthogonal to each other, we know $\|\mathbf{T}(t) \times \mathbf{N}(t)\| = 1$, so

$$\kappa = \frac{\|\mathbf{T}'(t)\|}{v(t)} = \frac{\|\mathbf{x}'(t) \times \mathbf{x}''(t)\|}{v(t)^3}. \tag{2.6}$$

The above discussion assumes the existence of $\mathbf{N}(t)$, which requires $\mathbf{T}'(t) \neq \mathbf{0}$. Even when $\mathbf{T}'(t) = 0$, we still use (2.4) to define the curvature κ (which would be 0), so we always have

$$\|\mathbf{a}_\perp\| = v(t)\|\mathbf{T}'(t)\| = v(t)^2\kappa, \quad \text{and} \quad \|\mathbf{a}\|^2 = [v'(t)]^2 + v(t)^4\kappa^2. \tag{2.7}$$

We also record

$$\mathbf{T}'(t) = v(t)\kappa\mathbf{N}(t) \quad \text{when } \mathbf{T}'(t) \neq \mathbf{0} \text{ so } \mathbf{N}(t) \text{ is defined.} \tag{2.8}$$

Remark 2.1.3

The condition $v(t) = \|\mathbf{x}'(t)\| > 0$ for $t \in (a, b)$ eliminates curves $\mathbf{x}(t)$ which are given by differentiable functions (even as many times as one would like), but which trace out a path with corners or cusps. $\mathbf{x}(t) = (t^3, t^2)$ for $t \in \mathbb{R}$ is such an example. Both t^3 and t^2 are infinitely many times differentiable, but the path traced out by $t \mapsto (t^3, t^2)$ has a cusp at $\mathbf{x}(0) = (0, 0)$. $v(t) = \|\mathbf{x}'(t)\|$ becoming 0 at $t = 0$ allows the unit tangent vector $T(t)$ to make an abrupt change of direction at $\mathbf{x}(0)$ (from pointing downward when $t < 0$ to pointing upward when $t > 0$), even though $\mathbf{x}'(t) = \mathbf{v}(t) = (3t^2, 2t^2)$ changes continuously across $t = 0$.

Remark 2.1.4

We already saw that if we allow $\mathbf{x}'(t) = \mathbf{0}$ at some t, this could allow the traced out path to have a corner or cusp (such as the case at $t = 0$ when $\mathbf{x}(t) = (t^3, t^2)$). If we allow $\mathbf{T}'(t) = \mathbf{0}$ at some t, or equivalently, $\mathbf{x}''(t)$ a scalar multiple of $\mathbf{x}'(t)$, this could cause the constructed $\mathbf{N}(t)$ to be discontinuous in t.

A simple example is $\mathbf{x}(t) = (t, t^3) \in \mathbb{R}^2$ for $t \in \mathbb{R}$. $\mathbf{x}'(t) = (1, 3t^2)$, $\mathbf{x}''(t) = (0, 6t)$. Geometrically this curve is concave downward when $t < 0$ so $\mathbf{N}(t)$ should point downward there, while it is concave upward when $t > 0$ so $\mathbf{N}(t)$ should point upward there. This will cause a discontinuity of $\mathbf{N}(t)$ at $t = 0$. Computationally, $\mathbf{N}(t)$ is to be defined as the unit vector in the direction of

$$\mathbf{x}''(t) - \left(\frac{\mathbf{x}''(t) \cdot \mathbf{x}'(t)}{\mathbf{x}'(t) \cdot \mathbf{x}'(t)}\right) \mathbf{x}'(t) = (0, 6t) - \frac{18t^3}{1 + 9t^4}(1, 3t^2) = \frac{t}{1 + 9t^4}(-18t^2, 6),$$

so we have

$$\mathbf{N}(t) = \begin{cases} -\frac{(-18t^2, 6)}{\sqrt{364t^4 + 36}}, & \text{if } t < 0, \\ \frac{(-18t^2, 6)}{\sqrt{364t^4 + 36}}, & \text{if } t > 0. \end{cases}$$

As $t \to 0-$, we see that $\mathbf{N}(t) \to -(0, 1)$, but as $t \to 0+$, $\mathbf{N}(t) \to (0, 1)$.

It is our desire to obtain a continuously varying principal normal vector $\mathbf{N}(t)$ through the Gram-Schmidt Algorithm that we require the condition that $\mathbf{x}''(t)$ not become a scalar multiple of $\mathbf{x}'(t)$, which then guarantees a continuously varying principal normal vector $\mathbf{N}(t)$, and implies that $\kappa > 0$.

For a curve in $\mathbb{R}^2$, we may choose to define $\mathbf{N}(t)$ such that $\{\mathbf{T}(t), \mathbf{N}(t)\}$ is right-handed, instead of relying on the Gram-Schmidt Algorithm as

applied to $\{\mathbf{x}'(t), \mathbf{x}''(t)\}$. We still have $\mathbf{T}'(t) = \kappa\mathbf{N}(t)$ for some scalar κ in such a formulation, but the κ here could be positive or negative, reflecting the curve being curving to the left or right as the particle moves ahead.

3. **Arc-length and arc-length parametrization.** The length of the curve $t \mapsto \mathbf{x}(t)$ when $a \le t \le b$ is $\int_a^b v(t)dt = \int_a^b \|\mathbf{x}'(t)\|dt$.

   A notion related to that of a curve is a path. A path is that traced out by a curve, removing possible redundancies when a curve traverses a portion multiple times. Analytically, a curve allows $t \mapsto \mathbf{x}(t)$ to be non-injective, while a path requires this parametrization to be injective. In other words, a curve emphasizes $\mathbf{x}(t)$ as a function of t, while a path treats $\mathbf{x}(t)$ as a geometric object in space, with t playing only the role of a parameter — other parameters can be used as well.

   For example, the curve $\mathbf{x}(t) = (\cos(2\pi t^2), \sin(2\pi t^2))$ for $0 \le t \le 2$, traverses the unit circle centered at the origin four times as t runs from 0 to 2; the latter is the path traced out by this curve. The length traveled would be $\int_0^2 \|\mathbf{x}'(t)\|\, dt$, while the length of the path is $\int_0^1 \|\mathbf{x}'(t)\|\, dt$, because this curve traces out the unit circle exactly once when t runs from 0 to 1.

   The usage of these two terminologies is not universal, so one has to watch out for the context to get the precise meaning.

   The length of the curve $\mathbf{x}(t)$ from $t = a$ to $t = t$ is

$$s(t) := \int_a^t \|\mathbf{x}'(\tau)\|d\tau. \tag{2.9}$$

   If follows that, if $v(t) = \|\mathbf{x}'(t)\| > 0$ for $t \in (a, b)$, then $t \mapsto s(t)$ is strictly increasing, differentiable, with a differentiable inverse, and it introduces a reparameterization in terms of s; furthermore,

$$\frac{d}{dt} = \frac{ds}{dt}\frac{d}{ds} = v(t)\frac{d}{ds}. \tag{2.10}$$

   Thus any derivative with respect to the arc-length parameter s is equivalent to $v(t)^{-1}$ times the derivative with respect to t, and the curvature formula becomes $\kappa = \|\mathbf{T}'(t)\|/v(t) = \|\frac{d\mathbf{T}}{ds}\|$.

   To study the geometry of a path traced out by a curve $\mathbf{x}(t)$, it is conceptually advantageous to use the arc length parametrization, although in doing computations, one rarely carries out this parametrization explicitly.

2.1.4 Torsion and the Frenet-Serret Formula for a curve in $\mathbb{R}^3$.

In (2.8) we obtain a formula for the rate of change of $\mathbf{T}(t)$. It's also interesting to obtain a formula for the rate of change of $\mathbf{N}(t)$, when it is well defined. Since $\mathbf{N}(t) \cdot \mathbf{N}(t) \equiv 1$, we also have $\mathbf{N}'(t) \cdot \mathbf{N}(t) \equiv 0$. Namely, $\mathbf{N}'(t) \perp \mathbf{N}(t)$.

Note that $\mathbf{N}'(t)$ must be a linear combination of $\mathbf{T}(t)$ and $\mathbf{B}(t)$:

$$\mathbf{N}'(t) = \alpha \mathbf{T}(t) + \beta \mathbf{B}(t) \text{ for some scalars } \alpha \text{ and } \beta.$$

Using the orthogonality between $\mathbf{T}(t)$ and $\mathbf{B}(t)$, we find $\alpha = \mathbf{N}'(t) \cdot \mathbf{T}(t)$. But $\mathbf{N}(t) \cdot \mathbf{T}(t) \equiv 0$, so taking derivatives in t gives us

$$\mathbf{N}'(t) \cdot \mathbf{T}(t) + \mathbf{N}(t) \cdot \mathbf{T}'(t) = 0,$$

which gives

$$\alpha = \mathbf{N}'(t) \cdot \mathbf{T}(t) = -\mathbf{N}(t) \cdot \mathbf{T}'(t) = -v(t)\kappa(t).$$

If t is the arc length parameter, then $v(t) = 1$, the coefficient β is called the torsion of the curve at $\mathbf{x}(t)$, and is denoted as $\tau(t)$; so in the general case we have $\beta = v(t)\tau(t)$. Furthermore

$$\begin{aligned} \mathbf{B}'(t) &= \mathbf{T}'(t) \times \mathbf{N}(t) + \mathbf{T}(t) \times \mathbf{N}'(t) \\ &= v(t)\kappa(t)\mathbf{N}(t) \times \mathbf{N}(t) + \mathbf{T}(t) \times v(t) \left[-\kappa(t)T(t) + \tau \mathbf{B}(t)\right] \\ &= v(t)\tau(t)\mathbf{T}(t) \times \mathbf{B}(t) \\ &= -v(t)\tau(t)\mathbf{N}(t). \end{aligned} \tag{2.11}$$

To summarize, we have obtain the formulae for the rate of change of $\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)$, under the assumption that $v(t), \kappa(t) > 0$ (to guarantee the existence of $\mathbf{T}(t), \mathbf{N}(t)$):

$$\begin{cases} \mathbf{T}'(t) = v(t)[\kappa(t)\mathbf{N}(t)] \\ \mathbf{N}'(t) = v(t)[-\kappa(t)\mathbf{T}(t) + \tau(t)\mathbf{B}(t)] \\ \mathbf{B}'(t) = v(t)[-\tau(t)\mathbf{N}(t)]. \end{cases} \tag{2.12}$$

These are called the Frenet-Serret equations.

The plane through $\mathbf{x}(t)$ with $\mathbf{B}(t)$ as its normal is called the osculating plane of the curve $\mathbf{x}(t)$ at $\mathbf{x}(t)$. Since $\|\mathbf{B}'(t)\| = |v(t)\tau(t)|$, the torsion $\tau(t)$ is the rate of turning of the unit normal $\mathbf{B}(t)$ of the osculating plane per unit length along the curve.

Note that if $\tau(t) = 0$ for $t \in (a, b)$, then $\mathbf{B}'(t) \equiv 0$ for $t \in (a, b)$. It follows that $\mathbf{B}(t)$ is a constant vector for $t \in (a, b)$. This implies that the curve $\mathbf{x}(t)$ stays in a fixed plane with $\mathbf{B}(t)$ as its normal.

It is not easy to compute $\tau(t)$ from the discussion above. We now derive a formula for computing $\tau(t)$. From

$$\mathbf{x}''(t) = v'(t)\mathbf{T}(t) + v(t)^2\kappa(t)\mathbf{N}(t)$$

we differentiate in t one more time to obtain

$$\begin{aligned}\mathbf{x}'''(t) &= v''(t)\mathbf{T}(t) + v'(t)\mathbf{T}'(t) + [v(t)^2\kappa(t)]'\mathbf{N}(t) + v(t)^2\kappa(t)\mathbf{N}'(t)\\ &= [v''(t) - v(t)^3\kappa(t)^2]\mathbf{T}(t) + [v(t)v'(t) + [v(t)^2\kappa(t)]']\mathbf{N}(t) + v(t)^3\tau(t)\mathbf{B}(t)\end{aligned}$$

and take the dot product of both sides above with $\mathbf{B}(t)$, to obtain, using (2.5)

$$\tau(t) = \frac{\mathbf{B}(t)\cdot\mathbf{x}'''(t)}{v(t)^3} = \frac{(\mathbf{x}'(t)\times\mathbf{x}''(t))\cdot\mathbf{x}'''(t)}{v(t)^6\kappa^2(t)}. \tag{2.13}$$

Remark 2.1.5

It is not so much interesting to compute $\tau(t)$; rather it is more interesting to learn what role it plays in determining the geometry of a curve. Professor Carlen discusses this issue via **Example 34** *and* **Theorems 26, 28**, *as well the Darboux vector.*
(2.12) indicates that the movement of the orthonormal frame $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ is determined by $v(t), \kappa(t)$, and $\tau(t)$.
A deeper theorem in differential geometry says that two thrice differentiable curves with their curvature non-vanishing anywhere, both parametrized by their arc length parameter, such that they start at the same point in the same direction, and their curvatures and torsions agree with each other at their respective points corresponding to the same arc length parameter ($\kappa_1(s) = \kappa_2(s)$, $\tau_1(s) = \tau_2(s)$), then the two curves coincide.
Conversely, given any continuous $\kappa(s) > 0$ and $\tau(s)$, and any initial point and initial tangent and principal normal directions, one can construct a (unique) curve, parametrized by arc length parameter, such that it starts at the initial point and its initial tangent and principal normal directions agree with the assigned ones, and its curvature and torsion at the point corresponding to the arc length parameter value s are equal to $\kappa(s) > 0$ and $\tau(s)$, respectively. In other words, curvature and torsion together determine the geometry of a curve in $\mathbb{R}^3$ completely, after the initial tangent and principal normal directions are assigned.

Remark 2.1.6

One motivation for defining the osculating plane to the curve $\mathbf{x}(t)$ *at* $\mathbf{x}(t_0)$ *to be the plane through* $\mathbf{x}(t_0)$ *containing the vectors* $\mathbf{T}(t_0)$ *and* $\mathbf{N}(t_0)$ *is the following consideration. If one does a Taylor expansion of* $\mathbf{x}(t)$ *at* $t = t_0$*, which can be done as in one variable calculus,*

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \mathbf{x}'(t_0)(t - t_0) + \frac{1}{2}\mathbf{x}''(t_0)(t - t_0)^2 + \mathbf{Rm}_{\mathbf{x}}(t),$$

where $\|\mathbf{Rm}_{\mathbf{x}}(t)\|/|t - t_0|^2 \to 0$ *as* $t \to t_0$*, one notes that*

$$\begin{aligned}
&\mathbf{x}(t_0) + \mathbf{x}'(t_0)(t - t_0) + \frac{1}{2}\mathbf{x}''(t_0)(t - t_0)^2 \\
=&\mathbf{x}(t_0) + v(t_0)(t - t_0)\mathbf{T}(t_0) + \frac{(t - t_0)^2}{2}\left[v'(t_0)\mathbf{T}(t_0) + v(t_0)^2\kappa(t_0)\mathbf{N}(t_0)\right] \\
=&\mathbf{x}(t_0) + \left[v(t_0)(t - t_0) + \frac{(t - t_0)^2 v'(t_0)}{2}\right]\mathbf{T}(t_0) + \left[\frac{(t - t_0)^2 v(t_0)^2\kappa(t_0)}{2}\right]\mathbf{N}(t_0).
\end{aligned}$$

This is a plane curve in the plane through $\mathbf{x}(t_0)$ *and containing the vectors* $\mathbf{T}(t_0)$ *and* $\mathbf{N}(t_0)$ *(the same plane as containing* $\mathbf{x}'(t_0)$ *and* $\mathbf{x}''(t_0)$*). In other words, up to order* $(t - t_0)^2$*, the behavior of the curve* $\mathbf{x}(t)$ *is that of the plane curve above; any spatial behavior of* $\mathbf{x}(t)$ *is reflected only through* $\mathbf{Rm}_{\mathbf{x}}(t)$*, which is of order higher than* $(t - t_0)^2$*. It is for this reason that the plane through* $\mathbf{x}(t_0)$ *containing the vectors* $\mathbf{T}(t_0)$ *and* $\mathbf{N}(t_0)$ *is called the osculating plane to the curve* $\mathbf{x}(t)$ *at* $\mathbf{x}(t_0)$*.*

Remark 2.1.7

The Frenet-Serret formulas has a geometric interpretation in terms of the Darboux vector[a]*. Let* $x(t)$ *be a twice differentiable curve with non-zero speed and curvature at each* t *in some open interval so that* $\mathbf{T}(t), \mathbf{N}(t)$ *and* $\mathbf{B}(t)$ *are all defined on this interval. The Darboux vector* $\boldsymbol{\omega}(t)$ *is defined on this interval by*

$$\boldsymbol{\omega}(t) = \tau(t)\mathbf{T}(t) + \kappa(t)\mathbf{B}(t).$$

Carlen states that for small $h > 0$*, the orthonormal basis* $\{\mathbf{T}(t+h), \mathbf{N}(t+h), \mathbf{B}(t+h)\}$ *is, up to errors of size* h^2 *, what one would get by applying a rotation of angle* $v(t)\|\boldsymbol{\omega}(t)\|$ *about the axis of rotation in the direction of* $\boldsymbol{\omega}(t)$*.*
This is seen based on the following

- *When a vector* $\mathbf{x}(t)$ *is moving along a circle of radius* r *at an angular speed of* ω*, then its instantaneous rate of motion is* $r\omega$ *and its direction of motion is orthogonal to the radius vector from the center of the circle to* $\mathbf{x}(t)$*. This is seen by setting up coordinates such that* $\mathbf{x}(t) = (r\cos(\omega t), r\sin(\omega t), 0)$*, so* $\mathbf{x}'(t) = r\omega(-sin(\omega t), \cos(\omega t), 0)$*.*

- $\boldsymbol{\omega}(t)$ *has the property that*

  $$\boldsymbol{\omega}(t)\times\mathbf{T}(t) = \kappa\mathbf{N}(t), \quad \boldsymbol{\omega}(t)\times\mathbf{N}(t) = -\kappa\mathbf{T}(t)+\tau\mathbf{B}(t), \quad \boldsymbol{\omega}(t)\times\mathbf{B}(t) = -\tau\mathbf{N}(t),$$

  so

  $$\mathbf{T}'(t) = v(t)\boldsymbol{\omega}(t)\times\mathbf{T}(t), \quad \mathbf{N}'(t) = v(t)\boldsymbol{\omega}(t)\times\mathbf{N}(t), \quad \mathbf{B}'(t) = v(t)\boldsymbol{\omega}(t)\times\mathbf{B}(t).$$

- $\boldsymbol{\omega}(t) \times \mathbf{T}(t)$ *is perpendicular to both* $\boldsymbol{\omega}(t)$ *and* $\mathbf{T}(t)$*, with its magnitude equal to* $\|\boldsymbol{\omega}(t)\|\sin\theta$*, where* θ *is the angle between* $\boldsymbol{\omega}(t)$ *and* $\mathbf{T}(t)$*. Since* $\|\mathbf{T}(t)\| = 1$ *for all* t*,* $\mathbf{T}(t)$ *moves as a vector on the unit sphere in* $\mathbb{R}^3$ *with* 0 *as center; and since* $\mathbf{T}'(t) \perp \boldsymbol{\omega}(t)$ *by the equations above, this means that* $\mathbf{T}(t)$ *is moving in a circle with* $\boldsymbol{\omega}(t)$ *as its normal. And* $\mathbf{T}(t)$ *rotating with* $\boldsymbol{\omega}(t)$ *would have* $\sin\theta$ *as the radius of the circle in which the rotation takes place, and since* $\|\mathbf{T}'(t)\| = v(t)\|\boldsymbol{\omega}(t)\|\sin\theta$*, this means that* $\mathbf{T}(t)$ *is rotating with* $\boldsymbol{\omega}(t)$ *as its axis of rotation and at an angular speed* $v(t)\|\boldsymbol{\omega}(t)\|$*. The same interpretation applies to* $\mathbf{N}(t)$ *and* $\mathbf{B}(t)$*.*

[a] This discussion is optional.

2.1.5 Integration of vector valued functions of a single variable.

Integration of vector valued functions of a single variable is not discussed formally, except in the discussion on geodesics in **2.1.8**. If $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))$ is continuous on $t \in [a, b]$, then we define the integration of $\mathbf{x}(t)$ over $[a, b]$ component wise:

$$\int_a^b \mathbf{x}(t)dt = \left(\int_a^b x_1(t)dt, \ldots, \int_a^b x_n(t)dt\right).$$

The usual properties such at $\int_a^b [c\mathbf{x}(t) + d\mathbf{y}(t)]dt = c\int_a^b \mathbf{x}(t)dt + d\int_a^b \mathbf{y}(t)dt$ hold. The following inequality looks similar to the one for integration of real valued functions,

but requires a more careful proof.

$$\| \int_a^b \mathbf{x}(t)dt\| \leq \int_a^b \|\mathbf{x}(t)\|dt. \tag{2.14}$$

When written out in the components, in encodes the following

$$\sqrt{\Big| \int_a^b x_1(t)dt\Big|^2 + \ldots + \Big| \int_a^b x_n(t)dt\Big|^2} \leq \int_a^b \sqrt{|x_1(t)|^2 + \ldots + |x_n(t)|^2}dt,$$

which is not easily seen to hold without using the underlying dot product structure here. If one interprets $\mathbf{x}(t)$ as the velocity at time t of a particle moving in $\mathbb{R}^n$, then $\int_a^b \mathbf{x}(t)dt = (\int_a^b x_1(t)dt, \cdots, \int_a^b x_n(t)dt)$ stands for the actual displacement of the particle from $t = a$ to $t = b$, which is effectively measuring displacement along straight lines, but $\int_a^b \|\mathbf{x}(t)\|dt$ accounts for the total length of the path that the particle has traveled, so (2.14) has a geometric interpretation that *displacement measured along straight lines is shorter than along any other paths*—although an analytical proof, to be given below, is not that easy to conceive.

Lemma. *Suppose* $\mathbf{y} \in \mathbb{R}^n$ *is such that* $\mathbf{y} \cdot \mathbf{v} \leq A\|\mathbf{v}\|$ *for all* $\mathbf{v} \in \mathbb{R}^n$*, then* $\|\mathbf{y}\| \leq A$.

The proof is simply by taking $\mathbf{v} = \mathbf{y}$ to obtain $\mathbf{y} \cdot \mathbf{y} \leq A\|\mathbf{y}\|$, from which the conclusion follows.

Proof of (2.14). Set $\mathbf{y} = \int_a^b \mathbf{x}(t)dt$ and take any $\mathbf{v} \in \mathbb{R}^n$. Then

$$\mathbf{y} \cdot \mathbf{v} = \int_a^b \mathbf{x}(t) \cdot \mathbf{v}dt \leq \int_a^b |\mathbf{x}(t) \cdot \mathbf{v}|dt \leq \int_a^b \|\mathbf{x}(t)\|\|\mathbf{v}\|dt \leq \left(\int_a^b \|\mathbf{x}(t)\|dt\right)\|\mathbf{v}\|,$$

where we have used Cauchy-Schwarz inequality $|\mathbf{x}(t) \cdot \mathbf{v}| \leq \|\mathbf{x}(t)\|\|\mathbf{v}\|$. It now follows from the Lemma that $\| \int_a^b \mathbf{x}(t)dt\| = \|\mathbf{y}\| \leq \int_a^b \|\mathbf{x}(t)\|dt$. Note how we avoid dealing with $\| \int_a^b \mathbf{x}(t)dt\|$ directly, but reduce this problem of estimating the length of a vector into a scalar problem by taking a dot product with a (test) vector $\mathbf{v}$. □

Chapter 3

CONTINUOUS FUNCTIONS

3.1 Continuity in several variables

As explained in Professor Carlen's notes, one of the basic motivations for considering continuous functions of several variables is that for this class of functions it is meaningful to look for schemes to solve equations of the kind $\mathbf{f}(\mathbf{x}) = \mathbf{b}$ by some approximation algorithm: one looks for $\mathbf{x}_k$ such that

(a) $\mathbf{f}(\mathbf{x}_k) = \mathbf{b}_k \to \mathbf{b}$ as $k \to \infty$, and

(b) $\mathbf{x}_k \to \mathbf{x}$ for some $\mathbf{x}$ as $k \to \infty$.

Then the continuity of $\mathbf{f}$ at $\mathbf{x}$ would imply $\mathbf{f}(\mathbf{x}) = \mathbf{b}$ and that $\mathbf{x}_k$, for k large, is a good approximation for a solution to $\mathbf{f}(\mathbf{x}) = \mathbf{b}$.

The above property is in terms of the behavior of $\mathbf{f}$ along a sequence of $\mathbf{x}_k \to \mathbf{x}_0$. The formal definition of continuity of a function of several variables at $\mathbf{x}_0$, as given in **Definition 34**, is in terms of the behavior of $\mathbf{f}$ on the continuum set of $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}_0\| \leq \delta$, which is almost the same as that for a function of a single variable, only replacing the absolute values by the lengths (also called norms) of the vectors. These two characterizations of continuity are equivalent, as proved in **Theorem 35**.

Definition 34 also defines a function $\mathbf{f}$ to be a continuous function on a set S if it is continuous at every point $\mathbf{x}_0$ of S. Note that the $\delta > 0$ in the requirement (3.8) depends on ϵ as well as on $\mathbf{x}_0$. If one can choose $\delta > 0$ to depend only on $\epsilon > 0$ but independent of $\mathbf{x}_0 \in S$, then we say $\mathbf{f}$ is uniformly continuous on S.

A first property of this definition is reflected in **Theorem 33**, the vector-valued function $\mathbf{f}(\mathbf{x}) := (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ is continuous iff each of $f_j(\mathbf{x})$ is continuous (either at a specific point or point wise on a set).

The next property of continuous functions is **Theorem 34**, the sum and product of continuous functions are continuous functions, as are a quotient of two such

functions $f(x)/g(x)$ whenever $g(x) \neq 0$, and the composition of two such functions, $h \circ f$.

Based on this property, functions constructed using a finite number of these procedures are continuous at an appropriate domain.

It follows that the one variable function $f(x) = \frac{1}{x}$ defined on $(0, \infty)$ is continuous at every point of $(0, \infty)$. But it fails to be uniformly continuous on $(0, \infty)$, for, in order for $|f(x) - f(x_0)| = \frac{|x-x_0|}{xx_0} \leq \epsilon$, we need $|x - x_0| \leq xx_0\epsilon$, but no matter how small $\delta > 0$ is taken, we can always find $x_0 > 0$ sufficiently close to 0 such that even if $|x - x_0| \leq \delta$, we can still find $x \in (0, \infty)$ such that $xx_0\epsilon < |x - x_0|$: just take $x = x_0 + \delta$, and take $0 < \delta < \frac{1}{2}$, $0 < x_0 < \min\{\delta/\epsilon, \frac{1}{2}\}$.

A useful property of a continuous function on a bounded and closed interval is that it is uniformly continuous there and attains the maximum and minimum values on that interval. One main goal in the several variable setting is to find an analogue of this property.

Using the same approach we can argue that $f(\mathbf{x}) := \|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$ is a continuous function of $\mathbf{x} \in \mathbb{R}^n$ by recognizing it as the composition of $u = g(\mathbf{x}) = x_1^2 + \cdots + x_n^2 : \mathbb{R}^n \mapsto \mathbb{R}_{\geq 0}$ with $y = \sqrt{u} : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$, where g is the sum of n continuous functions, and the square root function is also a continuous function on $\mathbb{R}_{\geq 0}$. Of course one could argue directly by examining

$$|\|\mathbf{x}\| - \|\mathbf{x}_0\|| \leq \|\mathbf{x} - \mathbf{x}_0\|,$$

which follows as a consequence of the triangle inequality:

$$\|\mathbf{x}\| \leq \|\mathbf{x} - \mathbf{x}_0\| + \|\mathbf{x}_0\|.$$

It then follows that for any $\epsilon > 0$, as long as we take $\delta = \epsilon$ and $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, we would get $|\|\mathbf{x}\| - \|\mathbf{x}_0\|| < \epsilon$, proving the continuity of $f(\mathbf{x}) = \|\mathbf{x}\|$ at $\mathbf{x}_0$.

When the construction involves a quotient of two continuous functions or a root or logarithm of a continuous function, then special attention needs to be paid in neighborhoods of points where the denominator or the function under the root or logarithm become 0, as the resulting function may fail to be continuous there.

Since a quotient of two continuous functions or a root or logarithm of a continuous function may produce points the denominator or the function under the root or logarithm become 0, we need to investigate the behavior of the resulting function in a neighborhood of such points in more detail.

For this, one often needs to work with the negation of continuity of a function $\mathbf{f}$ at a point $\mathbf{x}_0$. The formal way to describe $\mathbf{f}$ not continuous at $\mathbf{x}_0$ is that, for some $\epsilon > 0$, no matter how small a $\delta > 0$ one takes, one can always find some $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{x}_0\| \leq \delta$, such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| > \epsilon$.

Taking a sequence $\delta_k > 0$, $\delta_k \to 0$, as $k \to \infty$, one finds a sequence $\mathbf{x}_k$ such that $\|\mathbf{x}_k - \mathbf{x}_0\| \leq \delta_k$, but $\|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}_0)\| > \epsilon$. This produces a sequence $\mathbf{x}_k \to \mathbf{x}_0$ with $\|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}_0)\| > \epsilon$. This is what one needs to do in order to show that some $\mathbf{f}$ not continuous at $\mathbf{x}_0$.

One simple example is

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases} \tag{3.1}$$

where if $(x_k, y_k) = (\frac{1}{k}, \frac{1}{k})$, we would get $f(x_k, y_k) = \frac{1}{2}$, so $|f(x_k, y_k) - f(0,0)| \geq \frac{1}{2} > \epsilon$, if we take some $0 < \epsilon < \frac{1}{2}$.

In the other direction, **Theorem 35** characterizes the continuity of $\mathbf{f}$ at a point $\mathbf{x}_0$ through $\lim_{\mathbf{x}_k \to \mathbf{x}_0} \mathbf{f}(\mathbf{x}_k) = \mathbf{f}(\mathbf{x}_0)$ for *any sequence* $\{\mathbf{x}_k\}$ such that $\mathbf{x}_k \to \mathbf{x}_0$. This gives another criterion for f to be not continuous at $\mathbf{x}_0$: if there are two sequences $\mathbf{x}_k \to x_0$ and $\mathbf{y}_l \to x_0$ such that $\lim_{k\to\infty} f(\mathbf{x}_k) \neq \lim_{l\to\infty} f(y_l)$.

One can use this approach to see that the f above is not continuous at $(0,0)$, as for $(x_k, y_k) = (\frac{1}{k}, \frac{1}{k})$, we would get $f(x_k, y_k) = \frac{1}{2}$, while for $x_k, y_k) = (\frac{1}{k}, 0)$, we would get $f(x_k, y_k) = 0$.

For a function of one variable of the form $f(x)/g(x)$, where both $f(x)$ and $g(x)$ are continuous, but $f(x_0) = g(x_0) = 0$, whether $\lim_{x\to x_0} f(x)/g(x)$ exists (therefore making it possible to extend $f(x)/g(x)$ as a continuous function on a domain including x_0), one can apply l'Hospital rule or Taylor expansions to both the numerator and denominator to try to draw conclusions.

But for a function of several variables of the form $f(\mathbf{x})/g(\mathbf{x})$, the above method is not directly applicable. One often carries out some preliminary study by following one or both of the following approaches:

(i) Study $f(\mathbf{x})/g(\mathbf{x})$ along any straight-line passing through $\mathbf{x}_0$, namely, consider the line $\mathbf{x} = \mathbf{x}_0 + \mathbf{v}t$ for some fixed direction $\mathbf{v}$, and study $f(\mathbf{x}_0 + \mathbf{v}t)/g(\mathbf{x}_0 + \mathbf{v}t)$ as a function of one variable t; one can then apply tools from one variable calculus; in the case of $\mathbf{x} \in \mathbb{R}^2$ and when $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{v}$ is a unit vector, this amounts to studying the function $f(\mathbf{x})/g(\mathbf{x})$ in polar coordinates in $\mathbf{z} := \mathbf{x} - \mathbf{x}_0 = (t\cos\theta, t\sin\theta) = t(\cos\theta, \sin\theta)$.

(ii) Freeze all the components of $\mathbf{x}$ but one of them, and study $f(\mathbf{x})/g(\mathbf{x})$ as a function of the (unfrozen) free single variable. This is related to the notion of separate continuity given in **Definition 35**.

However, the above two approaches are merely preliminary examinations; a function can have separate continuity at a specific point or even if at every point in a

neighborhood of a point but fails to be continuous as a function of the joint variables. The function in (3.1) is one simple example.

Furthermore, even if $f(\mathbf{x})/g(\mathbf{x})$ converges to $f(\mathbf{x}_0)/g(\mathbf{x}_0)$ (or has a common limit) as $\mathbf{x} \to \mathbf{x}_0$ along *each* such path, it does not guarantee that $f(\mathbf{x})/g(\mathbf{x})$ is continuous at $\mathbf{x}_0$. One needs to find a way to verify the condition in the definition of continuity which is not affected by how $\mathbf{x} \to \mathbf{x}_0$.

Example 46 is such an example. It shows that, even if $f(\mathbf{v}r)/g(\mathbf{v}r) \to 0$ as $r \to 0$, for each fixed unit direction vector $\mathbf{v}$, the function $f(\mathbf{x})/g(\mathbf{x})$ may not have 0 as its limit as $\mathbf{x} \to 0$. For, with $\mathbf{v} = (a, b)$ and $(x, y) = r(a, b)$,

$$\frac{x^2 y}{x^4 + y^2} = r\frac{a^2 b}{a^4 r^2 + b^2},$$

which $\to 0$ as $r \to 0$: when $b \neq 0$, $\frac{a^2 b}{a^4 r^2 + b^2} \to \frac{a^2}{b}$ and the factor r would make the fraction $\to 0$; when $b = 0$, the fraction $= 0$ for all r, so certainly $\to 0$ as $r \to 0$. But, if one does not fix (a, b), allowing its adjustment as $r \to 0$, e.g., taking $b = rsa^2$ for some fixed s, then

$$r\frac{a^2 b}{a^4 r^2 + b^2} = \frac{s}{1 + s^2},$$

which does converge to 0 as $r \to 0$ for $s \neq 0$. The condition $b = rsa^2$ is equivalent to $y = sx^2$, so if $(x, y) \to (0, 0)$ along the parabola $y = sx^2$, the function $f(x, y)$ would take the constant value $\frac{s}{1+s^2} \neq 0$ for $s \neq 0$. Thus this $f(x, y)$ is not continuous at $(0, 0)$.

The condition $b = rsa^2$ is motivated by examining how big the fraction $\frac{a^2 b}{a^4 r^2 + b^2}$ can attain as $r \to 0$ and adjusting (a, b): the smallness of $a^4 r^2$ tends to make the fraction large, but it plays that role only when b^2 is comparable to $a^4 r^2$; equivalently when b is comparable to ra. More formally, for any $r > 0$ small but fixed, by varying $(a, b) = (\cos\theta, \sin\theta)$, one examines the maximum value of

$$\frac{a^2 b}{a^4 r^2 + b^2} = \frac{\cos^2\theta \sin\theta}{r^2 \cos^4\theta + \sin^2\theta} = \frac{\sin\theta \cos^{-2}(\theta)}{r^2 + (\sin\theta \cos^{-2}(\theta))^2},$$

which equals $\frac{1}{2r}$ for each $r > 0$ by taking $\sin\theta \cos^{-2}(\theta) = r$, thus $r\frac{a^2 b}{a^4 r^2 + b^2} = \frac{1}{2}$ along such a path, and fails to $\to 0$.

Remark 3.1.1

Examples illustrating possible behavior of the quotient of two continuous functions both vanishing at a common point are mostly based on modifying the

function in Example 45:

$$f(x,y)=\begin{cases}\frac{xy}{x^2+y^2} & (x,y)\neq(0,0),\\ 0 & (x,y)=(0,0).\end{cases}$$

For example, the function in Example 46 is obtained by replacing x with x^2, so the behavior along $y=kx$ there is now reflected as along $y=kx^2$; although this modified function now approaches 0 along each straight line segment of the form $y=kx$.
The central requirement for continuity of $f(\mathbf{x})$ at $\mathbf{x}_0$ is, for appropriate $\delta(\epsilon)>0$ depending on $\epsilon>0$,

$$\|\mathbf{x}-\mathbf{x}_0\|<\delta(\epsilon)\implies|f(\mathbf{x})-f(\mathbf{x}_0)|<\epsilon,$$

so it is natural to try to express $f(\mathbf{x})-f(\mathbf{x}_0)$ in terms the polar coordinates in $\mathbf{z}:=\mathbf{x}-\mathbf{x}_0$, namely, using $r=\|\mathbf{z}\|$ and $\boldsymbol{\omega}=(\cos\theta,\sin\theta)=\mathbf{z}/r$.
Sometimes it may not be easy to see directly that one can make $|f(\mathbf{x})-f(\mathbf{x}_0)|<\epsilon$; but if it is possible to get $|f(\mathbf{x})-f(\mathbf{x}_0)|\le g(\mathbf{x})$ for some function $g(\mathbf{x})$, and it is easier to insure that $g(\mathbf{x})<\epsilon$, we can draw the needed conclusion on f. This approach is called the Squeeze Principle.

In some situation, such as in Example 48, the function $\dfrac{x^5}{x^4+y^6}$ becomes $r\left(\dfrac{\cos^5\theta}{\cos^4\theta+r^2\sin^6\theta}\right)$, and the fraction $\left|\dfrac{\cos^5\theta}{\cos^4\theta+r^2\sin^6\theta}\right|\le|\cos\theta|\le 1$ after dropping the $r^2\sin^6\theta$ term in the denominator. Because this bound is uniform independent of θ, unlike in the situation of Example 46, we can conclude, using the Squeeze Principle, that $\dfrac{x^5}{x^4+y^6}\to 0$ as $(x,y)\to(0,0)$.

Another similar example is $\frac{x^2y}{x^2+y^2}$ for $(x,y)\neq(0,0)$. Since $0\le|\frac{x^2y}{x^2+y^2}|\le|y|$ (by dropping the y^2 in the denominator), and $|y|\to 0$ as $(x,y)\to(0,0)$, we conclude that $\frac{x^2y}{x^2+y^2}\to 0$ as $(x,y)\to(0,0)$.

If we modify the function in Example 46 into

$$f(x,y)=\begin{cases}\frac{1}{\sqrt[4]{x^2+y^2}}\frac{x^2y}{x^4+y^2} & \text{if } (x,y)\neq(0,0)\\ 0 & \text{if } (x,y)=(0,0),\end{cases}$$

then in polar coordinate, when $(x,y)=r(\cos\theta,\sin\theta)\neq(0,0)$, $f(x,y)=\sqrt{r}\left(\dfrac{\cos^2\theta\sin\theta}{r^2\cos^4\theta+\sin^2\theta}\right)$, which still $\to 0$ as $r\to 0$, for each fixed θ. However, along the curve $r\cos\theta\cot\theta=s$,

$f(x,y) = \dfrac{s}{\sqrt{r}(s^2+1)}$, which $\to \infty$ as $r \to 0$ for $s > 0$, and $\to -\infty$ as $r \to 0$ for $s < 0$. This function $f(x,y)$ is separately continuous along each vertical line $x = a$ or each horizontal line $y = b$, as well as on each straight-line through the origin, yet it fails to be bounded near $(0,0)$. In particular, considering it as a function defined in the disc $\{(x,y) : x^2 + y^2 \leq 1\}$, it does not attain a maximum or minimum value.

We would like a continuous function of several variables to maintain an important and useful property of a continuous function of a single variable defined on a closed interval: it attains its maximum and minimum value in that interval. It is for this reason that we work with the definition of continuous function as given, instead of the separate continuity, and we need to find the right notion in multi-dimensions which corresponds to a closed interval in one-dimension. The appropriate notion turns out to be that of a bounded closed set of $\mathbb{R}^n$—the formal name for a bounded closed set of $\mathbb{R}^n$ is a compact set.

Reading Quizzes/Questions:

1. If $f(x,y)$ is continuous at (x_0,y_0), does it imply that for any $x_k \to x_0$, $f(x_k,y_0) \to f(x_0,y_0)$?

2. If for any $x_k \to x_0$, $f(x_k,y_0) \to f(x_0,y_0)$, and for any $y_k \to y_0$, $f(x_0,y_k) \to f(x_0,y_0)$, does this imply that $f(x,y)$ is continuous at (x_0,y_0)?

3. If for any $x_k \to x_0$, $f(x_k,y_0) \to f(x_0,y_0)$, and there exists some $L > 0$ such that $|f(x,y) - f(x,y_0)| < L|y - y_0|$ for all (x,y) near (x_0,y_0), does this imply that $f(x,y)$ is continuous at (x_0,y_0)?

3.2 Continuity, compactness and maximizers*

In one variable calculus, we mostly deal with functions defined on an open interval, or a closed interval, or half-open and half-closed intervals. The most direct generalizations of these objects to multi-dimensions would be an open rectangular box of the form $\{(x_1,\ldots,x_n) : a_i < x < b_i, \text{ for each } 1 \leq i \leq n\}$, or a closed rectangular box of the form $\{(x_1,\ldots,x_n) : a_i \leq x \leq b_i, \text{ for each } 1 \leq i \leq n\}$, or generalization of the notion of half-open and half-closed intervals. However, these domains are too restrictive for applications; they don't even include domains in the shape of balls.

*Due to time constraint, our discussion of the material of this section will be very selective.

The appropriate extensions which include most domains we encounter in applications are the open or closed subsets in $\mathbb{R}^n$.

The following are three basic properties from one-variable calculus* that will be the basis for their extensions to multi-dimensions.

(I) (Bolzano-Weierstrass Theorem) Any bounded sequence in $\mathbb{R}$ has a convergent subsequence.

(II) Any Cauchy sequence in $\mathbb{R}$ is convergent. (A sequence $\{x_i\}$ is called a Cauchy sequence, or simply Cauchy, if for any $\epsilon > 0$ there exists N such that for all $i, j \geq N$, $|x_i - x_j| < \epsilon$. This notion can be easily extended to $\mathbb{R}^n$, simply replacing $|x_i - x_j| < \epsilon$ by $\|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon$.)

(III) Any continuous function of a single variable defined on a closed intelrval attains its maximum and minimum value in that interval.

A bounded and closed set in $\mathbb{R}^n$ is called a compact set of $\mathbb{R}^n$.

Theorem 38 is a generalization of (I) to multi-variables in the context of a bounded closed set.

Theorem. 38 (The Bolzano-Weierstrass Theorem) *Let K be a compact subset of $\mathbb{R}^n$. Then for every sequence $\{\mathbf{x}_k\}$ in K, there is a subsequence $\{\mathbf{x}_{k_l}\}$ and a $\mathbf{z} \in K$ such that $\lim_{l\to\infty} \mathbf{x}_{k_l} = \mathbf{z}$. On the other hand, if K is not compact, then there exists a sequence $\{\mathbf{x}_k\}$ in K that has no subsequence convergent in K.*

Theorem 39 (Continuity and Maximizers) is a generalization of (III).

The proof for **Theorem 38** in Professor Carlen's notes uses (II) while his proof for **Theorem 39** uses the notion of the least upper bound of a subset of $\mathbb{R}$.

A subset X of $\mathbb{R}$ is said to be bounded from above if there exists a real number b such that $x \leq b$ for all $x \in X$; such a number b is called an upper bound of X.

*These properties are properties of the set of all real numbers, and are usually discussed in details only in undergraduate analysis courses, not in a single variable course. These properties are subtle in that none of them would be valid if we confine ourselves working with functions defined on the set of rational numbers or taking values among rational numbers—even if we enlarge the set of rational numbers by including roots of polynomials with rational numbers as coefficients. Even though our daily usage of numbers are almost always limited to decimal numbers that have a finite number of digits, theoretical discussions require that limits of sequences of such numbers (such as when taking the integral of a function) are accounted for; without including these numbers and properties such as (I) and (II), we wouldn't be able to discuss the integral of some of the basic functions such as $\int_1^2 \frac{1}{x}\,dx$ or solutions of the one of the most basic differential equations $y'(x) = y(x)$.

Note that if b is an upper bound of X, then any number greater than b is also an upper bound of X, but a number smaller than b may no longer be an upper bound of X.

For any subset X of $\mathbb{R}$ which is bounded from above, as a consequence of (II)*, there exists a real number β such that (i) $x \leq \beta$ for all $x \in X$ (namely, β is an upper bound of X), and (ii) for any $\epsilon > 0$, $\beta - \epsilon$ is no longer an upper bound of X, namely, there exists some $x' \in X$ such that $\beta - \epsilon < x'$. This β is called the least upper bound of X (also commonly called the supremum of X), and is denoted as either $l.u.b(X)$ or $\sup X$—when such a number β exists, properties (i) and (ii) imply that it is unique. When a subset X of $\mathbb{R}$ which is not bounded from above, it makes sense to define its least upper bound as ∞.

This is a property of the set of real numbers; if one restricts oneself to working with only rational numbers, for instance, then a subset of rational numbers may not have a rational number as its least upper bound. In fact, none of (I–III) holds if one restricts oneself to working only with (continuous) functions defined on some set of rational numbers or taking values only on the set of rational numbers.

The notion of the least upper bound of a set of numbers is a proper generalization of the notion of the maximum of a finite set of numbers. A set of infinite numbers may not have a maximum value within the set, even if it is bounded from above; but always has a well defined least upper bound. For example, if we define $X = \{x \text{ is a rational number such that } x^2 < 2\}$, then we can see that X is bounded from above. But X has no value which is greater than any of the other values of X: for any rational x such that $x^2 < 2$, we can always find some other rational number y such that $y^2 < 2$ and $x < y$. X does have a least upper bound in the set of real

*The assertion that the least upper bound of X exists is a crucial property of the set of real numbers, whose proof relies on the Bolzano-Weierstrass property of real numbers, or an equivalent formulation. Carlen begins his proof of Theorem 39 with "Let B be the least upper bound of f on C", which means that he is using this assertion, although his notes are not spelling this out explicitly. Here is a sketch of above assertion in the case that f is bounded from above on C. Let b_1 be an upper bound of $\{f(\mathbf{x}) : \mathbf{x} \in C\}$. If there exists some $\mathbf{x}_0 \in C$ such that $f(\mathbf{x}_0) = b_1$, then b_1 is the least upper bound of f on C. It remains to study the case where this does not happen. Set $a_1 = f(\mathbf{x}_1)$ for some $\mathbf{x}_1 \in C$. Then $a_1 < b_1$. Let $c = (a_1 + b_1)/2$. If c is still an upper bound of $\{f(\mathbf{x}) : \mathbf{x} \in C\}$, we define $b_2 = c$ and $a_2 = a_1$, then repeat this procedure with a_2 and b_2; otherwise, there must be some $\mathbf{x}_2 \in C$ such that $c < f(\mathbf{x}_2)$, then we let $a_2 = c$, $b_2 = b_1$ and repeat this procedure with a_2 and b_2. This generates a sequence of intervals $[a_k, b_k]$ for $k \in \mathbf{N}$ such that $a_k \leq a_{k+1} < b_{k+1} \leq b_k$, and $(b_k - a_k) = (b_1 - a_1)/2^{k-1} \to 0$ as $k \to \infty$. The sequence of numbers $\{a_k\}$ is monotone non-decreasing, and bounded. By the Bolzano-Weierstrass property, it has a subsequence $\{a_{k_l}\}$ and a limit B such that $a_{k_l} \to B$ as $l \to \infty$. It follows also that $b_{k_l} \to B$ as $l \to \infty$. This B is then the least upper bound of f on C. For, any $\mathbf{x} \in C$ satisfies $f(\mathbf{x}) \leq b_{k_l}$, so $f(\mathbf{x}) \leq B = \lim_{l\to\infty} b_{k_l}$; and for any $\epsilon > 0$ and all sufficiently large l, $B - \epsilon < a_{k_l}$, but by construction of a_k, there exists some $\mathbf{x}_{k_l}$ such that $a_{k_l} < f(\mathbf{x}_{k_l})$, making $B - \epsilon$ not an upper bound of f on C. Thus B satisfies both criteria of the least upper bound of f on C.

numbers, and that number is labeled as $\sqrt{2}$.

If X is not bounded from above, namely, for any real number b, there exists some $x \in X$ such that $x > b$. By taking $b = k \in \mathbb{N}$, we would get a sequence $\{x_k\}$ in X such that $x_k > k$ for all $k \in \mathbb{N}$. In this way, we get a sequence $\{x_k\}$ in X which tends to ∞ as $k \to \infty$.

If X has a finite least upper bound β, then, similarly, one can find a sequence $\{x_k\}$ in X such that $x_k > \beta - \frac{1}{k}$ for all $k \in \mathbb{N}$, while $x \leq \beta$ for all $x \in X$.

Similarly, any subset X of $\mathbb{R}$ which is bounded from below has a greatest lower bound, denoted as either $g.l.b(X)$ or $\inf X$, and characterized by the properties (a) $x \geq \inf X$ for all $x \in X$, and (b) for any $\epsilon > 0$, $\inf X + \epsilon$ is no longer a lower bound of X, namely, there exists some $x' \in X$ such that $\inf X + \epsilon > x'$.

Below is a different proof of **Theorem 38** using (I). Let $\{\mathbf{x}_i\}$ be a sequence in the compact set K of $\mathbb{R}^n$. Since K is compact, it must be bounded, namely, there exists some $R > 0$ such that $\|\mathbf{x}\| \leq R$ for all $\mathbf{x} \in K$. In particular, $\|\mathbf{x}_i\| \leq R$ for all i. Write each $\mathbf{x}_i = ((\mathbf{x}_i)_1, \ldots, (\mathbf{x}_i)_n)$, then $|(\mathbf{x}_i)_k| \leq R$ for all i and $1 \leq k \leq n$. First we apply (I) to $\{(\mathbf{x}_i)_1\}$, the sequence consisting of the first components of $\mathbf{x}_i$, to obtain a subsequence $\{(\mathbf{x}_{i_{j_1}})_1\}$ which converges to some z_1 ($\{i_{j_1}\}$ is a subsequence of $\{1, 2, 3, \ldots\}$ indexed by j_1, which goes from $1, 2, \ldots$); then we apply (I) to $\{(\mathbf{x}_{i_{j_1}})_2\}$ to obtain a subsequence which converges to some z_2. The notation is becoming cumbersome; let's agree to use the generic notion $\{\mathbf{x}_{i_j}\}$ to denote this new subsequence. By now we have $(\mathbf{x}_{i_j})_1 \to z_1$ and $(\mathbf{x}_{i_j})_2 \to z_2$ as $j \to \infty$. We do this for $(n-2)$ more times to obtain a final subsequence, which we still denote as $\{\mathbf{x}_{i_j}\}$, which has the property that $(\mathbf{x}_{i_j})_k \to z_k$ for each $1 \leq k \leq n$ as $j \to \infty$. It remains to prove that $\mathbf{x}_{i_j} \to \mathbf{z} = (z_1, \ldots, z_n)$. But this follows from

$$\|\mathbf{x}_{i_j} - \mathbf{z}\| = \sqrt{\sum_{k=1}^{n} [(\mathbf{x}_{i_j})_k - z_k]^2} \leq \sqrt{n \max_k [(\mathbf{x}_{i_j})_k - z_k]^2} \leq \sqrt{n} \max_k |(\mathbf{x}_{i_j})_k - z_k|,$$

and each $|(\mathbf{x}_{i_j})_k - z_k| \to 0$ as $j \to \infty$.

Theorem 39 is the main property of continuous functions of interest to us. It does not give us an algorithm to find the maximum value or locations where the maximum values may be attained; it merely assures that the maximum value is attained somewhere in the compact domain of a continuous function.

Example 3.2.1

Let $I = \{(s, 0) \in \mathbb{R}^2 : 0 \leq s < 1\}$ and define $\rho(\mathbf{x}, I) = \inf\{|\mathbf{x} - (s, 0)| : (s, 0) \in I\}$ be the distance from $\mathbf{x} = (x, y) \in \mathbb{R}^2$ to the set I. Then for each $\mathbf{x} \in \mathbb{R}^2$, $\rho(\mathbf{x}, I)$ is well-defined, although, due to I being non-compact (why?),

for some particular $\mathbf{x} \in \mathbb{R}^2$, there may not be some $(s^*, 0) \in I$ attaining $\rho(\mathbf{x}, I)$ in the sense that $\rho(\mathbf{x}, I) = |\mathbf{x} - (s^*, 0)|$ (Can you identify such $\mathbf{x}$?—for this part, $0 \leq s < 1$ is the variable).
One the other hand, we can prove that this $\rho(\mathbf{x}, I)$ is a continuous function of $\mathbf{x} \in \mathbb{R}^2$, without having any formula for this function. For, given any $\mathbf{x}_0 \in \mathbb{R}^2$ and any $\epsilon > 0$, according to the property of $\rho(\mathbf{x}_0, I)$, first of all, there exists some $(s^*, 0) \in I$ such that

$$|\mathbf{x}_0 - (s^*, 0)| < \rho(\mathbf{x}_0, I) + \frac{\epsilon}{2}.$$

Then for any $\mathbf{x} \in \mathbb{R}^2$, by the triangle inequality, we have

$$|\mathbf{x} - (s^*, 0)| \leq |\mathbf{x}_0 - (s^*, 0)| + |\mathbf{x} - \mathbf{x}_0| < \rho(\mathbf{x}_0, I) + \frac{\epsilon}{2} + |\mathbf{x} - \mathbf{x}_0|,$$

which implies that

$$\rho(\mathbf{x}, I) \leq |\mathbf{x} - (s^*, 0)| < \rho(\mathbf{x}_0, I) + \frac{\epsilon}{2} + |\mathbf{x} - \mathbf{x}_0|; \tag{3.2}$$

secondly, for any $(s, 0) \in I$, by the triangle inequality, we have

$$|\mathbf{x} - (s, 0)| \geq |\mathbf{x}_0 - (s, 0)| - |\mathbf{x} - \mathbf{x}_0| \geq \rho(\mathbf{x}_0, I) - |\mathbf{x} - \mathbf{x}_0|,$$

which implies that

$$\rho(\mathbf{x}, I) \geq \rho(\mathbf{x}_0, I) - |\mathbf{x} - \mathbf{x}_0|. \tag{3.3}$$

Combining this inequality with (3.2), we get

$$\frac{\epsilon}{2} + |\mathbf{x} - \mathbf{x}_0| > \rho(\mathbf{x}, I) - \rho(\mathbf{x}_0, I) \geq -|\mathbf{x} - \mathbf{x}_0|,$$

thus for $\mathbf{x}$ such that $|\mathbf{x} - \mathbf{x}_0| < \epsilon$, we would get $|\rho(\mathbf{x}, I) - \rho(\mathbf{x}_0, I)| < \epsilon$, proving the continuity of $\rho(\mathbf{x}, I)$ at $\mathbf{x}_0$.
In fact, we can reverse the role between $\mathbf{x}_0$ and $\mathbf{x}$ in proving (3.3) to get

$$\rho(\mathbf{x}_0, I) \geq \rho(\mathbf{x}, I) - |\mathbf{x} - \mathbf{x}_0|, \tag{3.4}$$

Another way to prove (3.4) is to note that, for any $\mathbf{x}$ and $\mathbf{x}_0$, (3.2) holds for arbitrary ϵ, so as a consequence, we must have

$$\rho(\mathbf{x}, I) \leq \rho(\mathbf{x}_0, I) + |\mathbf{x} - \mathbf{x}_0|.$$

(Why? Why can't we keep the strict $<$?) Combining (3.4) with (3.3), we get

$$|\rho(\mathbf{x}, I) - \rho(\mathbf{x}_0, I)| \leq |\mathbf{x} - \mathbf{x}_0|.$$

This proves the Lipschitz continuity of $\rho(\mathbf{x}, I)$.

Question. *Although in general it is impossible to find an explicit formula for the distance from a point to a general set, in this particular case, one can use geometric arguments to find an explicit formula for* $\rho(\mathbf{x}, I)$*. Can you work this out?*

Chapter 4

DIFFERENTIABLE FUNCTIONS

Continuous functions have many desired properties, but in applications we often demand more: we want to be able to approximate a function by the simplest possible functions: linear functions in the independent variables. In the one variable setting, a linear function takes the form $a + b(x - x_0)$—written this way with the $b(x - x_0)$ term so that itself is a linear approximation to the constant term a when x is near x_0, and a function $f(x)$ can be approximated by $a + b(x - x_0)$ near x_0 if

$$\frac{|f(x) - [a + b(x - x_0)]\,|}{|x - x_0|} \to 0 \quad \text{as } x \to x_0,$$

namely, the error $|f(x) - [a + b(x - x_0)]\,|$ is vanishingly small compared with $|x - x_0|$ as $x \to x_0$.

In the one variable setting, this notion is equivalent to the existence of the derivative $f'(x_0)$ and $b = f'(x_0)$; geometrically, it means that the graph of $y = f(x)$ has a tangent line at $(x_0, f(x_0))$.

In the multi-variable setting, the linear approximations should be

$$a + b_1(x_1 - (\mathbf{x}_0)_1) + \cdots + b_n(x_n - (\mathbf{x}_0)_n) = a + \mathbf{b} \cdot (\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{b} = (b_1, \cdots, b_n)$ and $\mathbf{x}_0 = ((\mathbf{x}_0)_1, \cdots, (\mathbf{x}_0)_n)$.

It turns out that this notion by linear approximation (called differentiability) is different from the notion that the function has partial derivatives in each variable, namely, when treated as a function of a single variable if holding all variables fixed except for one, it has derivative in that variable. But this latter notion of partial derivatives is easier to work with in practice. We will discuss what knowledge of these partial derivatives would make a function differentiable, which would allow us to use linear approximation on such a function (see **Theorems 41, 42**).

4.1 Vertical slices and directional derivatives

4.1.1 Directional derivatives and partial derivatives

Carlen's description of the idea of "slicing" means that we study a function f of several variables by first focusing on its behavior on any one-dimensional slices (i.e., a one-dimensional line) so that we can apply our tools from one variable calculus. The concept of directional derivative comes out of this approach. The directions along the coordinate axes are special directions, and the directional derivative in the direction parallel to the x_i axis, if it exists, is call the partial derivative of the function f at $\mathbf{x}$ with respect to x_i.

There are multiple notations in usage for this quantity. Carlen uses $\frac{\partial}{\partial x_i} f(\mathbf{x})$; one also often sees $\frac{\partial f}{\partial x_i}(\mathbf{x})$, or $\partial_{x_i} f(\mathbf{x})$, of $f_{x_i}(\mathbf{x})$. In call these notations, one should treat the part before $(\mathbf{x})$ as a new function constructed out of f, and the part $(\mathbf{x})$ means that we are evaluating this new function at $\mathbf{x}$. In such a convention, $\frac{\partial f}{\partial x_i}(2\mathbf{x})$ or $\partial_{x_i} f(2\mathbf{x})$ would mean evaluating that function at $2\mathbf{x}$; if we need to work with the composite function $f(2\mathbf{x})$ and take its partial derivative at $\mathbf{x}$, our notation needs to make it clear that it is taking the partial derivative of this function at $\mathbf{x}$, not the partial derivative of $f(\mathbf{x})$ and evaluating it at $2\mathbf{x}$. In such a setting $\frac{\partial}{\partial x_i}[f(2\mathbf{x})]$ or $\partial_{x_i}[f(2\mathbf{x})]$ would be a better notation for the former, although this convention is not universally accepted.

Carlen does not introduce a specific notation for the directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{v}$, if it exists. A commonly used notation for this quantity is $\nabla_{\mathbf{v}} f(\mathbf{x})$ or $D_{\mathbf{v}} f(\mathbf{x})$.

Example 4.1.1

Define $f(x, y) = x^y$ for $x, y > 0$. When taking the partial derivative with respect to x variable, we hold y as a constant, thus $\partial_x f(x, y) = yx^{y-1}$; while when taking the partial derivative with respect to y variable, we hold x as a constant, thus $\partial_y f(x, y) = x^y \ln x$.

Question. *For a given function $f(\mathbf{x})$, suppose that its directional derivative exists for every non-zero (directional) vector $\mathbf{v} \in \mathbb{R}^n$, are these directional derivatives related to each other? —Bear in mind that there are infinitely many possible directional vectors at any $\mathbf{x}$. Theorem 41 gives an answer under some additional conditions on the partial derivatives of f.*

Remark 4.1.1

The existence of all partial derivatives does not necessarily guarantee the existence of all directional derivatives, as demonstrated by

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Both partial derivatives $\frac{\partial f}{\partial x}(x,y)$ *and* $\frac{\partial f}{\partial y}(x,y)$ *certainly exist for* $(x,y) \neq (0,0)$. *At* $(0,0)$, *to examine the existence of* $\frac{\partial f}{\partial x}(0,0)$, *we need to examine* $g(t) := f(t,0)$ *according to the definition, which is* 0 *for all* t, *so* $\frac{\partial f}{\partial x}(0,0) = g'(0) = 0$. *The same is true for* $\frac{\partial f}{\partial y}(0,0) = 0$. *Yet for* $\mathbf{v} = (v_1, v_2)$, *where* $v_1, v_2 \neq 0$, *we need to examine*

$$g(t) := f(tv_1, tv_2) = \begin{cases} \frac{v_1 v_2}{v_1^2+v_2^2}, & \text{if } t \neq 0, \\ 0, & \text{if } t = 0. \end{cases}$$

This $g(t)$ *is not even continuous as a function of* t *at* $t = 0$, *let alone having derivative at* $t = 0$! *So for such directions* $\mathbf{v}$, *the directional derivative of* f *at* $(0,0)$ *does not exist.*

4.1.2 The gradient and a chain rule for functions of a vector variable

Remark 4.1.2

The gradient vector $\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}))$ *is only valid in the rectangular coordinates. We can transform* $\mathbf{x}$ *into polar coordinates when* $\mathbf{x} \in \mathbb{R}^2$, *and define* $\frac{\partial f}{\partial r}(\mathbf{x}), \frac{\partial f}{\partial \theta}(\mathbf{x})$ *in a similar fashion, but the gradient vector in such transformed coordinates need to be defined differently.*

Remark 4.1.3

(4.6) and (4.7) require that the partial derivatives be continuous near the point where the formulae are to be applied. When that condition fails, the formulae (4.6) and (4.7) may not hold, as illustrated by

$$f(x,y) = \begin{cases} \frac{x^2y}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Similar to an example in the previous subsection, $\dfrac{\partial f}{\partial x}(0,0) = \dfrac{\partial f}{\partial y}(0,0) = 0$, *and for any* $\mathbf{v} = (v_1, v_2)$, *by looking at*

$$g(t) := f(tv_1, tv_2) = \begin{cases} \frac{tv_1^2v_2}{v_1^2+v_2^2}, & \text{if } t \neq 0, \\ 0, & \text{if } t = 0. \end{cases}$$

we see that $g'(0) = \dfrac{v_1^2v_2}{v_1^2+v_2^2}$, *which is* $\nabla_{\mathbf{v}} f(0,0)$, *and not equal to* $\mathbf{v} \cdot \nabla f(0,0) = \mathbf{v} \cdot (0,0)$.

Question. *Suppose that we know that the partial derivatives of* $f(\mathbf{x})$ *are continuous for* $\mathbf{x} \in \mathbb{R}^2$ *near* $\mathbf{x}_0$, *and, instead of knowing the values of all the partial derivatives of* $f(\mathbf{x})$ *at* $\mathbf{x}_0$, *we know the values of* $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ *for two directional vectors* $\mathbf{v}$ *which are not multiples of each other. Can we figure out the values of all the partial derivatives of* $f(\mathbf{x})$ *at* $\mathbf{x}_0$*? What about the situation for higher dimensions?*

4.1.3 The geometric meaning of the gradient

The discussion of (4.8) uses the geometric suggestive notion of angle θ. The underlying property is the Cauchy-Schwarz inequality: $-\|\mathbf{a}\|\|\mathbf{b}\| \leq \mathbf{a} \cdot \mathbf{b} \leq \|\mathbf{a}\|\|\mathbf{b}\|$.

Based on (4.8), if $f(\mathbf{x})$ has continuous partial derivatives near $\mathbf{x}_0$, then for any unit directional vector $\mathbf{u}$, we have

$$-\|\nabla f(\mathbf{x}_0)\| \leq \nabla_{\mathbf{u}} f(\mathbf{x}_0) = \mathbf{u} \cdot \nabla f(\mathbf{x}_0) \leq \|\nabla f(\mathbf{x}_0)\|,$$

and $\nabla_{\mathbf{u}} f(\mathbf{x}_0) = \mathbf{u} \cdot \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\|$ if and only if $\mathbf{u}$ points in the same direction as $\nabla f(\mathbf{x}_0)$. Thus $\nabla f(\mathbf{x}_0)$ points in the direction of steepest increase of f at $\mathbf{x}_0$.

4.1.4 Critical points

In looking for a minimum or maximum point of a function f defined in a domain D, we need to generalize a theorem from one variable calculus. For the case of $D = [a, b]$, we have

(A) If f is continuous on $[a, b]$, then f must attain a minimum and a maximum value on $[a, b]$, as a consequence of the Bolzano-Weierstrass property of the set of real numbers; but this theorem does not tell us how to find the minimum or maximum value.

(B) If f is continuous on $[a, b]$ and differentiable in (a, b), and if f attains its minimum or maximum value in $[a, b]$ at an interior point c: $a < c < b$, then $f'(c) = 0$. In such a situation, we can find the minimum and maximum values of f in $[a, b]$ by finding all its critical points c in (a, b), and compare the values of $f(c)$ with $f(a)$ and $f(b)$.

If D is in multi-dimensions, the first task is to define interior and boundary point of D. A point $\mathbf{x} \in D$ is an interior point of D, if there exists some $\delta > 0$ such that the ball $B_\delta(\mathbf{x}) \subset D$. A point $\mathbf{x}$ is a boundary point of D, if for any $\delta > 0$, the ball $B_\delta(\mathbf{x})$ contains points in D and also points in the complement D^c of D. Note that a boundary point of D may not be in D itself. The set of all boundary points of D is denoted as ∂D. When D is itself an open ball $B_r(\mathbf{x}_0)$, then every point of $B_r(\mathbf{x}_0)$ is an interior point of $B_r(\mathbf{x}_0)$, while a point $\mathbf{x}$ is a boundary point of $B_r(\mathbf{x}_0)$ iff $\|\mathbf{x} - \mathbf{x}_0\| = r$, namely, iff $\mathbf{x}$ is on the sphere $\partial B_r(x_0)$ of radius r centered at $\mathbf{x}_0$. It is possible for a set to have no interior point at all, as in the case of the sphere $\partial B_r(x_0)$ of radius r centered at $\mathbf{x}_0$.

When we try to generalize (A)-(B) to multi-dimensions, (A) still holds as given by **Theorem 39**; but the generalization of (B) would need to study the values of f on its boundary points, which are often consisting of infinitely many points!—take the case when D is the closed ball $\overline{B_r(x_0)}$ of radius r centered at $\mathbf{x}_0$.

Here we need to

(i). identify all interior critical points $\mathbf{c}$,

(ii). identify the minimum and maximum values of f on the boundary ∂D of D, and

(iii). compare the values of $f(\mathbf{c})$ for all interior critical points $\mathbf{c}$ with the minimum and maximum values of f on the boundary ∂D of D.

(ii) can often be solved using Lagrange multipliers of **section 5.2.1**.

Example 4.1.2

To find the minimum and maximum of the function $f(x,y) = x^2 + xy$ on the closed disc $D = \{(x,y) : x^2 + y^2 \leq 1$, we first find solutions in the interior of D of

$$\begin{aligned} \partial_x f(x,y) = 2x + y &= 0 \\ \partial_y f(x,y) = x &= 0 \end{aligned}$$

namely, the critical points in the interior of D. The only critical point here is $(x,y) = (0,0)$, where $f(0,0) = 0$.
Next we need to identify the minimum and maximum of the function $x^2 + xy$ on the boundary $\partial D = \{(x,y) : x^2 + y^2 = 1\}$. We will develop the method of Lagrange multipliers in **section 5.2.1**, but we can solve this particular problem by noting that any point $(x,y) \in \partial D$ can be parametrized as $(\cos\theta, \sin\theta)$ for some $\theta \in [0, 2\pi]$, and $f(x,y) = \cos^2\theta + \cos\theta\sin\theta = \frac{\cos(2\theta)+\sin(2\theta)+1}{2}$. We can find the minimum and maximum of this function of θ over $[0, 2\pi]$ using one variable calculus: the minimum is attained when $\cos(2\theta) = \sin(2\theta) = -\frac{1}{\sqrt{2}}$, so $\theta = \frac{5\pi}{8}$ or $\frac{13\pi}{8}$, with the minimum value $= \frac{1}{2} - \sqrt{2}$; the maximum is attained when $\cos(2\theta) = \sin(2\theta) = \frac{1}{\sqrt{2}}$, so $\theta = \frac{\pi}{8}$ or $\frac{9\pi}{8}$, with the maximum value $= \frac{1}{2} + \sqrt{2}$.
Finally, we compare the extremal value $f(0,0)$ in the interior of D with those on the boundary ∂D to conclude that the minimum of $x^2 + xy$ on the closed disc D is $\frac{1}{2} - \sqrt{2}$, and the maximum is $\frac{1}{2} + \sqrt{2}$.

Exercise 4.1.1. *Find the minimum and maximum of the function* $f(x,y) = x^2 + xy + \frac{x+y}{2}$ *on the closed square* $S = \{(x,y) : |x|, |y| \leq 1\}$.

Example 63 involves finding the minimum and maximum of the function $f(x,y) = x^4 + y^4 + 4xy$ on the unbounded domain $\mathbb{R}^2$. We can't apply **Theorem 39** directly; but find that we can reduce the problem to a situation where **Theorem 39** is applicable when we recognize that $f(x,y) \geq \frac{1}{2}(\|\mathbf{x}\|^2 - 2)^2 - 2$, which implies that $f(x,y) \geq 0$ when $\|\mathbf{x}\| \geq 2$, but $f(\epsilon, -\epsilon) = -\epsilon^2(4 - \epsilon^2) < 0$, for $\epsilon > 0$ small. So if a minimum of f exists, it must be < 0, and can't occur for $\|\mathbf{x}\| \geq 2$. Thus we can reduce the problem to looking at f on $C = \{\mathbf{x} : \|\mathbf{x}\| \leq 2\}$. On its boundary ∂C, $f(x,y) \geq 0$. All three critical points of f, $(0,0)$, $(-1,1)$, $(1,-1)$ are interior points of C, and we find $f(0,0) = 0$, $f(-1,1) = -2$, $f(1,-1) = -2$. This allows us to conclude that the minimum of f on $\mathbb{R}^2$ is -2, and it is attained at $(1,-1)$ and $(-1,1)$.

On the other hand, when we examine the maximum possible value of f on $\mathbb{R}^2$, we find, using $f(x,y) \geq \frac{1}{2}(\|\mathbf{x}\|^2 - 2)^2 - 2$, that $f(x,y)$ can be greater than any preassigned value M, by taking $\|\mathbf{x}\|$ sufficiently large: as long as $\|\mathbf{x}\|^2 > 2 + \sqrt{4 + 2M}$, then $f(x,y) \geq \frac{1}{2}(\|\mathbf{x}\|^2 - 2)^2 - 2 > M$. So f has no finite maximum value on $\mathbb{R}^2$.

If we change the function f to $f(x,y) = x^4 - y^4 + 4xy$. Then the only critical point is $(x,y) = (0,0)$, and $f(0,0) = 0$. We can still apply **Theorem 39** to any closed ball $\overline{B_r(\mathbf{0})}$. But the maximum and minimum values of f on $\overline{B_r(\mathbf{0})}$ all occur at a boundary point of $\overline{B_r(\mathbf{0})}$, instead of at its only critical point $(0,0)$; and the maximum and minimum values of f on $\overline{B_r(\mathbf{0})}$ grow unbounded with R, so this f has no finite maximum or minimum value on $\mathbb{R}^2$.

Reading Quizzes/Questions:

(i) If $f(\mathbf{x})$ has continuous partial derivatives in a domain U, and attains its maximum on a sub-domain D of U at a point $\mathbf{x}_0 \in D$, then $\nabla_{\mathbf{v}} f(\mathbf{x}_0) = 0$ for all $\mathbf{v}$?

(ii) If $f(\mathbf{x})$ attains its maximum on a domain D at an interior point $\mathbf{x}_0$, and $f(\mathbf{x})$ has a well-defined directional derivative $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ for some $\mathbf{v}$, then $\nabla_{\mathbf{v}} f(\mathbf{x}_0) = 0$?

4.1.5 The gradient and tangent planes

It is worthwhile to examine **Example 65**. One important concept of this subsection is **Definition 48** (Differentiability of functions from $\mathbb{R}^n$ to $\mathbb{R}$)—in one variable calculus the differentiability of a function f at some x is equivalent to the existence of the derivative of f at x; but for a function of more than one variables, the existence of all partial derivatives of f at $\mathbf{x}$ is *not equivalent* to the differentiability of a function f at $\mathbf{x}$.

A key feature of the notion of differentiability of f at $\mathbf{x}_0$ is that, instead of focusing on partial derivatives or directional derivatives of f at $\mathbf{x}_0$, we focus on finding a "best linear approximation" of $f(\mathbf{x}) - f(\mathbf{x}_0)$ at $\mathbf{x}_0$, namely, a "linear function"

$$L(\mathbf{x}) = b_1(x_1 - (\mathbf{x}_0)_1) + \cdots + b_n(x_n - (\mathbf{x}_0)_n) = \mathbf{b} \cdot (\mathbf{x} - \mathbf{x}_0)$$

such that the remainder $Rm_f(\mathbf{x}, \mathbf{x}_0) := f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x})$ vanishes at "higher order than linear rate of $\mathbf{x} - \mathbf{x}_0$" in the sense that

$$\frac{|f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|} \to 0 \quad \text{as } \mathbf{x} \to \mathbf{x}_0.$$

A basic consequence of $f(\mathbf{x})$ being differentiable at $\mathbf{x}_0$ is that all directional derivatives of $f(\mathbf{x})$ at $\mathbf{x}_0$ exist, and $L(\mathbf{x})$ is uniquely determined as $L(\mathbf{x}) = \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$.

This follows by setting $\mathbf{x} = \mathbf{x}_0 + h\mathbf{e}_i$ in $\mathbf{x} \to \mathbf{x}_0$, where $1 \le i \le n$ is fixed, then $L(\mathbf{x}) = b_i h$, and

$$\frac{|f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|} = \frac{|f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0) - b_i h|}{|h|} = \left| \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h} - b_i \right|,$$

so it follows that

$$\lim_{h \to 0} \left| \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h} - b_i \right| = 0,$$

and $b_i = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h} = \partial_{x_i} f(\mathbf{x}_0)$, as a result $\mathbf{b} = \nabla f(\mathbf{x}_0)$.

When $\mathbf{x} \in \mathbb{R}^2$, the graph of $f(\mathbf{x}_0) + L(\mathbf{x})$ is a plane passing through $(\mathbf{x}_0, f(\mathbf{x}_0))$, and tangent to the graph of $z = f(\mathbf{x})$ at $(\mathbf{x}_0, f(\mathbf{x}_0))$. When $\mathbf{x} \in \mathbb{R}^n$, $n > 2$, the graph of $f(\mathbf{x}_0) + L(\mathbf{x})$ is a hyperplane in $\mathbb{R}^{n+1}$ passing through $(\mathbf{x}_0, f(\mathbf{x}_0))$, which is also called the tangent (hyper)plane to the graph of $z = f(\mathbf{x})$ at $(\mathbf{x}_0, f(\mathbf{x}_0))$. Note that the equation of the tangent plane $y = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$ is equivalent to $((\mathbf{x}, y) - (\mathbf{x}_0, f(\mathbf{x}_0))) \cdot (\nabla f(\mathbf{x}_0), -1) = 0$, from which we see that $(\nabla f(\mathbf{x}_0), -1)$ is a normal vector to the tangent plane.

Theorem 42 says that if all the partial derivatives of $f(\mathbf{x})$ exist for $\mathbf{x} \in U$ and these partial derivatives are continuous for $\mathbf{x} \in U$, then $f(\mathbf{x})$ is differentiable at every $\mathbf{x} \in U$. In fact, a more precise statement holds: if all the partial derivatives of $f(\mathbf{x})$ exist for $\mathbf{x}$ in some ball $B(\mathbf{x}_0)$ centered at $\mathbf{x}_0$, and that these partial derivatives are continuous at $\mathbf{x}_0$, then $f(\mathbf{x})$ is differentiable at $\mathbf{x}_0$.

Example 4.1.3

Here is an illustration of a basic usage of the linear approximation. For $f(x, y) = x^y$ for $x, y > 0$, we are interested in finding the linear approximation to $f(x, y)$ near $(2, 1)$ and use it to estimate $f(2.1, 0.9)$. Recall that $\partial_x f(x, y) = yx^{y-1}$ and $\partial_y f(x, y) = x^y \ln x$, so they are continuous near $(2, 1)$, as a result, x^y is differentiable at $(2, 1)$, and its linear approximation is given by

$$f(2, 1) + \partial_x f(2, 1)(x - 2) + \partial_y f(2, 1)(y - 1) = 2 + 1(x - 2) + 2 \ln 2(y - 1),$$

so $f(2.1, 0.9) = 2.1^{0.9}$ can be approximated by $2 + 0.1 - 0.2 \ln 2 = 1.96$, up to an error which is (vanishingly) small compared with $\sqrt{(2.1 - 2)^2 + (0.9 - 1)^2} = 0.1\sqrt{2} \approx 0.14$.
Also the equation $x^y = 2$ can be approximated near $(2, 1)$ by $2 + 1(x - 2) + 2 \ln 2(y - 1) = 2$, which is actually the tangent line to the curve $x^y = 2$ at $(2, 1)$.

Example 4.1.4

Take $f(x,y)$ to be defined by

$$f(x,y) = \begin{cases} \frac{x^2y}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

If this $f(x,y)$ is differentiable at $\mathbf{x}_0 = (0,0)$, we would have $L(x,y) = f_x(0,0)x + f_y(0,0)y = 0$, and

$$\frac{|f(x,y) - f(0,0) - L(x,y)|}{||(x,y)||} = \frac{|f(x,y)|}{||(x,y)||} \to 0 \quad \text{as } (x,y) \to (0,0).$$

But

$$\frac{|f(x,y)|}{||(x,y)||} = \frac{|x^2y|}{(x^2+y^2)^{3/2}} = \cos^2(\theta)|\sin(\theta)|$$

in terms of the polar coordinates (r,θ) of (x,y), which does not tend to 0 as $r = \sqrt{x^2+y^2} \to 0$. This shows that this given $f(x,y)$ is not differentiable at $(0,0)$ (even though all directional derivatives of $f(x,y)$ exist at $(0,0)$).
If we define $g(x,y) = yf(x,y)$, then we also have $g(0,0) = 0$, and $g_x(0,0) = g_y(0,0) = 0$, so if we choose $g(0,0) + g_x(0,0)x + g_y(0,0)y = 0$ as the "best linear approximation" to $g(x,y)$ at $(0,0)$, we find

$$\frac{|g(x,y) - 0|}{||(x,y)||} = \frac{x^2y^2}{(x^2+y^2)^{3/2}} = r\cos^2(\theta)\sin^2(\theta),$$

which tends to 0 as $r = \sqrt{x^2+y^2} \to 0$ by the Squeeze Principle. So this $g(x,y)$ is differentiable at $(0,0)$.
The differentiability of both $f(x,y)$ and $g(x,y)$ at $(x_0,y_0) \neq (0,0)$ follows from Theorem 42.

Here are the relations among the four related notions: (a) $f(\mathbf{x})$ is differentiable at $\mathbf{x}_0$; (b) all directional derivatives of $f(\mathbf{x})$ exist at $\mathbf{x}_0$; (c) all the partial derivatives of $f(\mathbf{x})$ exist at $\mathbf{x}_0$; (d) all the partial derivatives of $f(\mathbf{x})$ exist in some ball containing $\mathbf{x}_0$ and are continuous at $\mathbf{x}_0$.

$$\text{(d)} \implies \text{(a)} \implies \text{(b)} \implies \text{(c)}.$$

Theorem 41 was formulated assuming a stronger version of (d); the same conclusion actually holds only assuming (a).

Example 4.1.5

Here is an example of a function $f(x,y)$ which is differentiable at some (x_0,y_0), in fact differentiable at all nearby points, so its partial derivatives exist at all nearby points, but the partial derivatives are not all continuous at (x_0,y_0).

$$f(x,y)=\begin{cases}(x^2+y^2)\sin\frac{1}{(x^2+y^2)^a}, & \text{if } (x,y)\neq(0,0),\\ 0, & \text{if } (x,y)=(0,0),\end{cases}$$

where $a\geq 1/2$. This example is based on a function in one-variable calculus which is differentiable everywhere, but has discontinuous derivatives.
The partial derivatives of this $f(x,y)$ are continuous at any $(x_0,y_0)\neq(0,0)$, so its differentiability at any $(x_0,y_0)\neq(0,0)$ follows from Theorem 41. It only remains to check the differentiability at $(0,0)$. We know that when a function is differentiable at some (x_0,y_0), then its linear approximation there is given by $f(x_0,y_0)+\nabla f(x_0,y_0)\cdot(x-x_0,y-y_0)$. In the case here, we can compute $\partial_x f(0,0)=\partial_y f(0,0)=0$ by restricting f to the x and y axis. For example

$$\partial_x f(0,0)=\lim_{h\to 0}\frac{h^2\sin\frac{1}{h^{2a}}}{h}=\lim_{h\to 0}h\sin\frac{1}{h^{2a}}=0.$$

Thus we only need to examine whether

$$\lim_{(x,y)\to(0,0)}\frac{\left|(x^2+y^2)\sin\frac{1}{(x^2+y^2)^a}-0-(0,0)\cdot(x,y)\right|}{\sqrt{x^2+y^2}}=0.$$

But

$$\frac{\left|(x^2+y^2)\sin\frac{1}{(x^2+y^2)^a}-0-(0,0)\cdot(x,y)\right|}{\sqrt{x^2+y^2}}=\left|(x^2+y^2)^{1/2}\sin\frac{1}{(x^2+y^2)^a}\right|\to 0$$

as $(x,y)\to(0,0)$, so we conclude that this f is differentiable at $(0,0)$.
For any $(x,y)\neq(0,0)$, we find

$$\partial_x f(x,y)=2x\sin\frac{1}{(x^2+y^2)^a}-\frac{2ax}{(x^2+y^2)^a}\cos\frac{1}{(x^2+y^2)^a},$$

and as $(x,y)\to(0,0)$, $2x\sin\frac{1}{(x^2+y^2)^a}\to 0$, but $\frac{2ax}{(x^2+y^2)^a}$ may approach ∞ when $2a>1$. When $2a=1$, we restrict $\partial_x f(x,y)$ to the x-axis to find

$$\partial_x f(x,0)=2x\sin\frac{1}{|x|}-\frac{x}{|x|}\cos\frac{1}{|x|},$$

which does not converge to $\partial_x f(0,0)=0$ as $x\to 0$. Thus $\partial_x f(x,y)$ is not continuous at $(0,0)$.

Reading Quizzes/Questions:

(i) If $f(\mathbf{x})$ has well-defined directional derivative $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ for every direction vector $\mathbf{v}$, is $f(\mathbf{x})$ necessarily continuous at $\mathbf{x}_0$?

(ii) If $f(\mathbf{x})$ is differentiable at $\mathbf{x}_0$, is $f(\mathbf{x})$ necessarily continuous at $\mathbf{x}_0$?

(iii) If $f(\mathbf{x})$ has well-defined directional derivative $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ for every direction vector $\mathbf{v}$, does $f(\mathbf{x})$ necessarily increase in the direction of $\nabla f(\mathbf{x}_0)$?

(iv) If $f(\mathbf{x})$ has well-defined directional derivative $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ for every direction vector $\mathbf{v}$, does there necessarily exist a direction $\mathbf{v}$ such that $\nabla_{\mathbf{v}} f(\mathbf{x}_0) = 0$?

(v) If $f(\mathbf{x})$ is differentiable at $\mathbf{x}_0$, is $\nabla f(\mathbf{x}_0)$ a normal vector to the tangent plane of the graph of $f(\mathbf{x})$ at $(\mathbf{x}_0, f(\mathbf{x}_0))$?

4.2 Linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$

To generalize the notion of the best linear approximation to a function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, for $m > 1$, the first step is to study linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$.

In earlier contexts, we tend to call a function such as $y = ax+b$, or $z = ax+by+c$, a linear function, as these functions involve only "linear terms" of the unknown(s). More properly, these functions should be called affine functions. We will reserve the name of *linear functions* to functions such as $y = ax$ or $z = ax + by$; these functions exhibit two simple and useful properties that functions such as $y = ax + b$, or $z = ax + by + c$ do not quite have. They are summarized in Professor Carlen's notes as (4.22)

$$\mathbf{f}(s\mathbf{x} + t\mathbf{y}) = s\mathbf{f}(\mathbf{x}) + t\mathbf{f}(\mathbf{y}) \quad \text{for all } s, t \in \mathbb{R} \text{ and } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{4.22}$$

(4.22) implies that $\mathbf{f}$ is determined completely by knowing $\mathbf{f}(\mathbf{e}_i)$, $i = 1, 2, \ldots, n$—review (4.24), and these n vectors in $\mathbb{R}^m$ can be organized as an $m \times n$ matrix, as described below.

4.2.1 The matrix representation of linear functions

We could describe a linear function in terms of its components, by writing $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, then $\mathbf{f}$ is a linear function when each of its components, $f_i(\mathbf{x})$ is

a linear function of $\mathbf{x}$ of the form $a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n$. Putting these together, we have

$$\begin{cases} f_1(\mathbf{x}) = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ f_2(\mathbf{x}) = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \quad \vdots \qquad\qquad \vdots \\ f_m(\mathbf{x}) = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{cases}$$

So such an $\mathbf{f}$ is determined by the $m \times n$ coefficients $a_{ij}, 1 \leq i \leq m, 1 \leq j \leq n$. It is conceptually more clear to organize these coefficients as an $m \times n$ array, called a matrix:

$$A_{\mathbf{f}} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

Since the matrix $A_{\mathbf{f}}$ encodes all information about $\mathbf{f}$, it is natural to associate $\mathbf{f}(\mathbf{x})$ with $A_{\mathbf{f}}\mathbf{x}$, where we think of the matrix $A_{\mathbf{f}}$ acting, or multiplying, on $\mathbf{x}$ to produce $\mathbf{f}(\mathbf{x})$:

$$\begin{aligned} \mathbf{f}(\mathbf{x}) = A_{\mathbf{f}}\mathbf{x} = &\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{bmatrix} \\ = &x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix}. \end{aligned}$$

Even if an $m \times n$ matrix A is not necessarily associated with a function $\mathbf{f}$, we still define its product with a vector $\mathbf{x} \in \mathbb{R}^n$, $A\mathbf{x}$, as a vector in $\mathbb{R}^m$, which is the linear combination $x_1A_1 + x_1A_1 + \ldots + x_nA_n$ of the columns of A:

$$A\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \mathbf{x} = x_1A_1 + x_1A_1 + \ldots + x_nA_n.$$

Note that this requires that the number of columns of A matches the number of components of $\mathbf{x}$. This also implies that $A\mathbf{e}_j = A_j$, namely, the jth column of A is the output when A acts on (or multiplies to) $\mathbf{e}_j$.

Remark 4.2.1

Most textbooks make a clear distinction between denoting a vector as a row vector or as a column vector. Professor Carlen's notes do not make this distinction. He treats the row vector $(x_1, \ldots, x_n)$ *and the column vector* $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ *as different representations of the same vector! So you would see in his notes that for an* $m \times n$ *matrix* A*, the matrix product* $A(x_1, \ldots, x_n)$*, which, in most standard texts, would be denoted as* $A\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ *or as* $A\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$.

Note that if we take $\mathbf{e}_j = (0, \ldots, 1, \ldots, 0)$ (with the only 1 on the jth slot), then

$$\mathbf{f}(\mathbf{e}_j) = \begin{bmatrix} f_1(\mathbf{e}_j) \\ \vdots \\ f_m(\mathbf{e}_j) \end{bmatrix} = A\mathbf{e}_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix},$$

which is the jth column A_j of the matrix A. In general, for $\mathbf{x} = (x_1, \ldots, x_n) = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n$, we have, by (4.22),

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n) = x_1\mathbf{f}(\mathbf{e}_1) + \ldots + x_n\mathbf{f}(\mathbf{e}_n) = x_1A_1 + \ldots + x_nA_n,$$

so the following equalities summarize the different representations $A\mathbf{x}$, $x_1A_1 + x_2A_2 + \ldots + x_nA_n$ of $\mathbf{f}(\mathbf{x})$:

$$\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{bmatrix} = \mathbf{f}(\mathbf{x}) = x_1A_1 + x_2A_2 + \ldots + x_nA_n$$

$$= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

The main theorem of this subsection is **Theorem 45**: A function $\mathbf{f}$ from $\mathbb{R}^n$ to $\mathbb{R}^m$ is linear if and only if for some $m \times n$ matrix A, $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$.

4.2.2 Composition of linear functions and matrix multiplication

Suppose that $\mathbf{z}$ itself is a linear function of another vector variable $\mathbf{y} \in \mathbb{R}^m$: $\mathbf{z} = g(\mathbf{y}) = (g_1(\mathbf{y}), \ldots, g_p(\mathbf{y})) \in \mathbb{R}^p$ for $\mathbf{y} \in \mathbb{R}^m$. Then we can write out

$$\begin{cases} g_1(\mathbf{y}) = b_{11}y_1 + b_{12}y_2 + \ldots + b_{1m}y_m \\ g_2(\mathbf{y}) = b_{21}y_1 + a_{22}y_2 + \ldots + b_{2m}y_m \\ \quad \vdots \qquad\quad \vdots \\ g_p(\mathbf{y}) = b_{p1}y_1 + b_{p2}y_2 + \ldots + b_{pm}y_m \end{cases}$$

and have a corresponding matrix

$$B = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & b_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ b_{p1} & b_{p2} & \ldots & b_{pm} \end{bmatrix}$$

We can verify that the composition $g \circ f(\mathbf{x})$ is a linear function of $\mathbf{x}$ in two ways: First, we work with the definition of linear functions directly:

$$(g\circ f)(s\mathbf{x}+t\mathbf{y}) = g(f(s\mathbf{x}+t\mathbf{y})) = g(sf(\mathbf{x})+tf(\mathbf{y})) = sg(f(\mathbf{x}))+tg(f(\mathbf{y})) = s(g\circ f)(\mathbf{x})+t(g\circ f)(\mathbf{y});$$

second, we work with the concrete expressions for each component:

$$\begin{aligned} (g \circ f)_i(\mathbf{x}) =& b_{i1}y_1 + b_{i2}y_2 + \ldots + b_{im}y_m \\ =& b_{i1}[a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n] + b_{i2}[a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n] + \ldots \\ & + b_{im}[a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n] \\ & = c_{i1}x_1 + c_{i2}x_2 + \ldots + c_{in}x_n, \end{aligned}$$

where

$$c_{ik} = b_{i1}a_{1k} + b_{i2}a_{2k} + \ldots + b_{im}a_{mk} = \sum_{j=1}^{m} b_{ij}a_{jk}.$$

If C is the $p \times n$ matrix $(c_{ik})_{i=1,\ldots,p;k=1,\ldots,n}$, then we define the multiplication of B and A (in that order) as $C = BA$, and the above rule gives how the multiplication should be carried out.

A short-hand way to view the above defining relation is that c_{ik} is given by the dot product between the ith row of B and the kth column of A; again the number of columns of B has to match the number of rows of A. Denoting the ith row of B by $\mathbf{b}_i$, the jth column of B by B_j, the ith row of C by $\mathbf{c}_i$, and the kth column of C by C_k, we now have three different ways of doing the matrix multiplication $C = BA$:

(I) $c_{ik} = \mathbf{b}_i \cdot \mathbf{A}_k$;

(II) $C_k = BA_k = a_{1k}B_1 + a_{2k}B_2 + \ldots + a_{mk}B_m$; (Column k of BA is a linear combination of columns of B, using the entries in the kth column of A as coefficients in the linear combination;)

(III) $\mathbf{c}_i = \mathbf{b}_i A = \begin{bmatrix} b_{i1} & b_{i2} & \ldots & b_{im} \end{bmatrix} A = b_{i1}\mathbf{a}_1 + b_{i2}\mathbf{a}_2 + \ldots + b_{im}\mathbf{a}_m$. (Row i of BA is a linear combination of rows of A, using the entries in the ith row of B as coefficients in the linear combination.)

Note that AB may not equal BA, even if both are defined!

Example 4.2.1

Suppose that

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -4 & 5 \end{bmatrix}, \quad A \text{ is a } 3 \times 2 \text{ matrix.}$$

If $C = BA$, then C is a 2×2 matrix, and according to (II), row 1 of C $\mathbf{c}_1 = 1\mathbf{a}_1 + 2\mathbf{a}_2 + 3\mathbf{a}_3$, row 2 of C $c_2 = 0\mathbf{a}_1 - 4\mathbf{a}_2 + 5\mathbf{a}_3$; but according to (III), the two columns of C are

$$C_1 = BA_1 = a_{11}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_{21}\begin{bmatrix} 2 \\ -4 \end{bmatrix} + a_{31}\begin{bmatrix} 3 \\ 5 \end{bmatrix},$$
$$C_2 = BA_2 = a_{12}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_{22}\begin{bmatrix} 2 \\ -4 \end{bmatrix} + a_{32}\begin{bmatrix} 3 \\ 5 \end{bmatrix},$$

Note also that AB is a 3×3 matrix, with its first row equal to

$$\left(\mathbf{a}_1 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{a}_1 \cdot \begin{bmatrix} 2 \\ -4 \end{bmatrix}, \mathbf{a}_1 \cdot \begin{bmatrix} 3 \\ 5 \end{bmatrix}\right)$$

and its first column equal to $1\text{Column}_1(A) + 0\text{Column}_2(A)$, etc. AB and BA are not even of the same size in this case!

Reading Quizzes/Questions:

(i) True or False: If A and B are two matrices such that AB is defined, and $AB = O$, where O is a matrix all of whose entries are 0, then either $A = O$ or $B = O$. (Here the three matrices O's could be matrices of different sizes).

(ii) True or False: If A, B, and C are matrices such that AB and AC are defined, $AB = AC$, and $A \neq O$, then it must be true that $B = C$.

Remark 4.2.2

Matrix multiplication is defined to describe the composition of linear functions, so properties of special linear functions are reflected in their matrix representations. E.g., a Householder reflection $h_{\mathbf{u}}$ has two special properties: (a) it preserves the dot product of vectors, (b) $h_{\mathbf{u}} \circ h_{\mathbf{u}} =$ the identity function $I : \mathbb{R}^n \to \mathbb{R}^n$ defined as $I(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$. As a consequence of (a), $h_{\mathbf{u}}(\mathbf{e}_i)$ is a unit vector for each i, and $h_{\mathbf{u}}(\mathbf{e}_i) \cdot h_{\mathbf{u}}(\mathbf{e}_j) = 0$ for $i \neq j$. Thus the columns of the matrix representation $A_{h_{\mathbf{u}}}$ are orthonormal. (b) implies that $A_{h_{\mathbf{u}}}^2 = I_n$, the identity matrix whose columns are $\mathbf{e}_1, \ldots, \mathbf{e}_n$.
Note also, since $h_{\mathbf{u}}(\mathbf{u}) = -\mathbf{u}$, and $h_{\mathbf{u}}(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v}$ such that $\mathbf{v} \perp \mathbf{u}$, we must have $A_{h_{\mathbf{u}}}\mathbf{u} = -\mathbf{u}$, and $A_{h_{\mathbf{u}}}\mathbf{v} = \mathbf{v}$ for all $\mathbf{v}$ such that $\mathbf{v} \perp \mathbf{u}$.

4.2.3 Solving the equation $A\mathbf{x} = \mathbf{b}$

An $m \times n$ matrix A defines a linear function $f : \mathbf{x} \in \mathbb{R}^n \mapsto A\mathbf{x} \in \mathbb{R}^m$. Here are some basic questions that we need to address:

(a) Given a specific vector $\mathbf{b} \in \mathbb{R}^m$, how do we determine whether there is a solution $\mathbf{x} \in \mathbb{R}^n$ to $A\mathbf{x} = \mathbf{b}$? If there is one, is there a unique one or there are multiple, or perhaps infinitely many solutions? Is there a formula or algorithm to find the solutions?

(b) Is there a criterion, or criteria, that guarantees the solvability of $A\mathbf{x} = \mathbf{b}$ for *every* $\mathbf{b} \in \mathbb{R}^m$?

Recall that $A\mathbf{x} = \mathbf{b}$ is equivalent to

$$x_1 \operatorname{Col}_1(A) + x_2 \operatorname{Col}_2(A) + \cdots + x_n \operatorname{Col}_n(A) = \mathbf{b},$$

namely, $\mathbf{b}$ is in the Column Space $\operatorname{Col}(A)$ of matrix A consisting of the span of the column vectors of the matrix A. So question (a) above is equivalent to whether $\mathbf{b}$ is in the Column Space A, and question (b) is equivalent to whether the Column Space A is the entire $\mathbb{R}^m$.

The question of uniqueness of solution is related to the Null Space $\operatorname{Null}(A)$ of matrix A consisting of all vectors $\mathbf{z} \in \mathbb{R}^n$ such that $A\mathbf{z} = \mathbf{0}$. If $\mathbf{x}_1 \neq \mathbf{x}_2$ are different

solutions to $A\mathbf{x} = \mathbf{b}$, then $A(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$, making $\mathbf{x}_1 - \mathbf{x}_2$ a non-zero vector in Null(A); conversely, if $\mathbf{z}$ is a non-zero vector in Null(A), then whenever $\mathbf{x}$ solves $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} + \mathbf{z}$ is also a solution:

$$A(\mathbf{x} + \mathbf{z}) = A\mathbf{x} + A\mathbf{z} = \mathbf{b} + \mathbf{0} = \mathbf{b}.$$

Thus $A\mathbf{x} = \mathbf{b}$ has non-unique solutions if and only if the null space Null(A) contains non-zero vectors. As a consequence, if we know all the solutions $\mathbf{z}$ to $A\mathbf{z} = \mathbf{0}$, and a particular solution $\mathbf{x}_0$ to $A\mathbf{x}_0 = \mathbf{b}$, then the general solution of $A\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x}_0 + \mathbf{z}$, where $\mathbf{z}$ is the general solution to $A\mathbf{z} = \mathbf{0}$.

Note also that $\mathbf{x} \in$ Null(A) if and only if $\mathbf{x}$ is orthogonal to each row of matrix A, namely, $\mathbf{x}$ is in the orthogonal complement of the row space of matrix A.

For an $m \times n$ matrix A, the main properties of the solvability of $A\mathbf{x} = \mathbf{b}$ is summarized in the following.

(i). $A\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b}$ lies in the column space of A.

(ii). $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$ if and only if the column space of A is the full $\mathbb{R}^m$, namely, A has m pivotal columns (in the Gram-Schmidt Orthogonalization process).

(iii). $A\mathbf{x} = \mathbf{b}$ has at most one solution if and only if $A\mathbf{x} = \mathbf{0}$ has $\mathbf{x} = \mathbf{0}$ as the only solution, namely, the null space of A is $\{\mathbf{0}\}$.

(iv). If A is $n \times n$, then $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$ if and only if the null space of A is $\{\mathbf{0}\}$.

(v). If $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$, and $\mathbf{z}$ satisfies $A\mathbf{z} = \mathbf{0}$, then $\mathbf{x}+\mathbf{z}$ satisfies $A(\mathbf{x}+\mathbf{z}) = \mathbf{b}$.

(vi). If $\mathbf{x}$ and $\mathbf{y}$ satisfy $A\mathbf{x} = A\mathbf{y} = \mathbf{b}$, then $\mathbf{x} - \mathbf{y}$ satisfies $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$.

(iv) follows from the earlier parts as follows. If the null space of A is $\{\mathbf{0}\}$, then every column of A is a pivotal column, for, otherwise, say column j is not a pivotal column, then $\text{Col}_j(A) = x_1\text{Col}_1(A) + \ldots + x_{j-1}\text{Col}_{j-1}(A)$ for some coefficients $x_1, \ldots, x_{j-1}$.

This then implies that

$$A\begin{bmatrix} x_1 \\ \vdots \\ x_{j-1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

namely, there is a non-zero vector in the null space of A. Now that we know every column of A is a pivotal column, and A is assumed to be an $n \times n$ square matrix, it follows that A has exactly n pivotal columns, which span $\mathbb{R}^n$. Thus $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$.

Conversely, if $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$, it means that A has n pivotal columns, so for a square $n \times n$ matrix A, every column is a pivotal column and $A\mathbf{x} = \mathbf{0}$ can't have a non-zero solution: if $\mathbf{x} = (x_1, \ldots, x_n)$ is a non-zero solution. Let x_j be the last entries among $x_1, \ldots, x_n$ to be non-zero, i.e., $x_j \neq 0$, but $x_k = 0$ for all $k > j$. This then implies that $x_1\text{Col}_1(A) + \ldots + x_{j-1}\text{Col}_{j-1}(A) + x_j\text{Col}_j(A) = \mathbf{0}$. Since $x_j \neq 0$, we can solve for $\text{Col}_j(A)$ in terms of $\text{Col}_1(A), \ldots, \text{Col}_{j-1}(A)$ in the form of

$$\text{Col}_j(A) = -\frac{x_1}{x_j}\text{Col}_1(A) - \ldots - \frac{x_{j-1}}{x_j}\text{Col}_{j-1}(A),$$

which then means that $\text{Col}_j(A)$ is not a pivotal column of A.

Remark 4.2.3

The discussion here for solving $A\mathbf{x} = \mathbf{b}$ is at a conceptual level; it does not give an algorithm for computing $\mathbf{x}$ when the system has a solution. We will describe an algorithm in the next subsection using the QR factorization of matrix A. It is different from the algorithm of doing row reductions to the augmented matrix of the system, which is often the first algorithm introduced in an elementary linear algebra course.

When A is either of the following two types, then there are simple algorithms to find the solutions of $A\mathbf{x} = \mathbf{b}$:

(I). *A is a square matrix whose columns are orthonormal;*

(II). *A is an $m \times n$ matrix in a row echelon form, namely, for each $1 \leq i \leq m$, either all entries of row i of A are zero, or the first non-zero entry, call the pivot of this row, occurs in a column j with $j \geq i$, and if $i < i'$, then the pivot of row i' occurs at a later column than that of row i. This forces such a matrix to be of upper triangular form, namely, the entries in (i, j)*

for $j < i$ are all 0. Another consequence is that, if a_{ij} is the pivot of row i of A, then $a_{i'j} = 0$ for all $i' > i$, namely, all entries in positions below a pivot are 0. Here are some simple examples of such matrices

$$\begin{bmatrix} 2 & -1 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & -1 & 0 & 1 & 2 \\ 0 & 0 & -1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix}$$

The highlighted entries are the pivots.

If each row of an upper triangular $m \times n$ matrix R has a pivot, then $n \geq m$, and the columns of R containing a pivot span $\mathbb{R}^m$, so we can solve $R\mathbf{x} = \mathbf{b}$ for any $\mathbb{R}^m$ in such a situation. Furthermore, if, in addition, R has any column that does not contain a pivot (this would imply $n > m$), we can set that variable as a free variable and solve the pivot variables in terms of these free variables, then there will be no unique solution to $R\mathbf{x} = \mathbf{b}$.

If some row of an upper triangular $m \times n$ matrix R has no pivot, then this row must be 0, and for any $\mathbf{b} \in \mathbb{R}^m$, whose entry in this row is not zero, there will be no solution to $R\mathbf{x} = \mathbf{b}$.

The second matrix above as coefficient matrix would correspond to a linear system with 5 unknowns of the form

$$\begin{bmatrix} 2 & -1 & 0 & 1 & 2 \\ 0 & 0 & -1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

We can treat x_5 as a free variable and use the last equation to solve for x_4 in terms of x_5: $x_4 = b_3 - 3x_5$, then plug this into the previous equation to solve for x_3: $x_3 = -2x_4 + 2x_5 - b_2 = -b_2 - 2b_3 + 8x_5$. Finally we treat x_2 as a free variable also, and solve for x_1 from the first equation to get $x_1 = \frac{1}{2}(x_2 - x_4 - 2x_5 + b_1) = \frac{1}{2}(b_1 - b_3 + x_2 + x_5)$. We can write the solution in vector form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(b_1 - b_3 + x_2 + x_5) \\ x_2 \\ -b_2 - 2b_3 + 8x_5 \\ b_3 - 3x_5 \\ x_5 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(b_1 - b_3) \\ 0 \\ -b_2 - 2b_3 \\ b_3 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} \frac{1}{2} \\ 0 \\ 8 \\ -3 \\ 1 \end{bmatrix},$$

which shows clearly the roles of the free variables x_2 and x_5. In particular, setting

$x_2 = x_5 = 0$, we get a particular solution, while the vectors

$$\begin{bmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \frac{1}{2} \\ 0 \\ 8 \\ -3 \\ 1 \end{bmatrix}$$

correspond to the case of $b_1 = b_2 = b_3 0$ and setting $x_2 = 1$ and $x_5 = 0$, and respectively, $x_2 = 0$ and $x_5 = 1$, which both satisfy

$$\begin{bmatrix} 2 & -1 & 0 & 1 & 2 \\ 0 & 0 & -1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 & 1 & 2 \\ 0 & 0 & -1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ 0 \\ 8 \\ -3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

This is consistent with (v) and (vi) at the beginning of this subsection. In fact, the above two vectors form a basis of the null space of the coefficient matrix.

The matrix

$$\begin{bmatrix} 2 & -1 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

has two pivots in rows 1 and 2, but no pivot in row 3. The system

$$\begin{bmatrix} 2 & -1 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

would have no solution unless $b_3 = 0$. When $b_3 = 0$, the system would have infinitely many solutions, as in such a case, x_3 can freely take any value, and we can solve for x_1 and x_2 in terms of x_3.

Reading Quizzes/Questions: Investigate the following True or False questions.

(a). If $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$, then the null space of A is $\{\mathbf{0}\}$.

(b). If the null space of A is $\{\mathbf{0}\}$, then $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$.

(c). If $\mathbf{x}$ and $\mathbf{y}$ satisfy $A\mathbf{x} = A\mathbf{y} = \mathbf{b}$, then any linear combination of $\mathbf{x}$ and $\mathbf{y}$ also solves the same system.

(d). If $A\mathbf{x} = \mathbf{b}$ has more than one solutions, then it has infinitely many solutions.

(e). The matrix

$$\begin{bmatrix} 2 & -1 & 0 & 1 & 2 \\ 0 & 0 & -1 & -2 & 2 \\ 0 & 0 & 1 & 1 & 3 \end{bmatrix}$$

is not upper triangular.

(f). The matrix

$$\begin{bmatrix} 2 & -1 & 0 & 1 & 2 \\ 0 & 0 & -1 & -2 & 2 \\ 0 & 1 & 0 & 1 & 3 \end{bmatrix}$$

is in a row echelon form.

For the case of (I), where A is a *square* matrix whose columns are orthonormal, the system $A\mathbf{x} = \mathbf{b}$ is equivalent to

$$x_1\text{Col}_1(A) + x_2\text{Col}_2(A) + \ldots + x_n\text{Col}_n(A) = \mathbf{b}.$$

Since $\{\text{Col}_1(A), \text{Col}_2(A), \ldots, \text{Col}_n(A)\}$ is a set of n orthonormal vectors in $\mathbb{R}^n$, it forms a basis of $\mathbb{R}^n$ by the Fundamental Theorem on Orthonormal Sets in $\mathbb{R}^n$, and $x_i = \text{Col}_i(A) \cdot \mathbf{b}$. So the vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \text{Col}_1(A) \cdot \mathbf{b} \\ \vdots \\ \text{Col}_n(A) \cdot \mathbf{b} \end{bmatrix}$$

is the unique solution to $A\mathbf{x} = \mathbf{b}$ in such a case. Note that in this case, base on matrix multiplication rules, we can also write $\mathbf{x}$ as

$$\mathbf{x} = A^T\mathbf{b},$$

where A^T is the transpose of matrix A so that the jth row of A^T equals the jth column of A.

Reading Quizzes/Questions: Verify that with $A = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$, the columns of A form an orthonormal set of vectors in $\mathbb{R}^2$, and that

$$\begin{bmatrix} \text{Col}_1(A) \cdot \mathbf{b} \\ \text{Col}_2(A) \cdot \mathbf{b} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix} \mathbf{b} = A^T\mathbf{b}.$$

When A is a *not* square matrix whose columns are still orthonormal, we must have $m > n$; we could still form the vector

$$\mathbf{x} = \begin{bmatrix} \mathrm{Col}_1(A) \cdot \mathbf{b} \\ \vdots \\ \mathrm{Col}_n(A) \cdot \mathbf{b} \end{bmatrix}$$

but this $\mathbf{x}$ may not solve $A\mathbf{x} = \mathbf{b}$! It turns out that

$$A\mathbf{x} = \sum_{i=1}^{n} (\mathrm{Col}_i(A) \cdot \mathbf{b}) \, \mathrm{Col}_i(A)$$

is the orthogonal projection of $\mathbf{b}$ in the columns space of A, and the vector on the right is closest vector in $\mathrm{Col}(A)$ to $\mathbf{b}$. The solution $\mathbf{x}$ here is called a least square (approximate) solution to $A\mathbf{x} = \mathbf{b}$. This notion will be explored in a challenge problem set. Note also that

$$\sum_{i=1}^{n} (\mathrm{Col}_i(A) \cdot \mathbf{b}) \, \mathrm{Col}_i(A) = AA^T\mathbf{b}.$$

Reading Quizzes/Questions: Verify that the columns of the following matrices are orthonormal, then discuss the solvability $Q\mathbf{x} = \mathbf{b}$ for various choices of $\mathbf{b}$ (e.g. $\mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ in the case of Q_2, Q_3).

$$Q_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \quad Q_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ 0 & -\frac{2}{\sqrt{6}} \end{bmatrix}.$$

For each Q, compute Q^TQ and QQ^T, then discuss features of these matrices.

The next subsection will discuss how to factorize any non-zero $m \times n$ matrix A as QR, where Q is some $m \times r$ matrix whose columns are orthonormal (though r may be $< m$), and R is an $r \times n$ matrix in row echelon form with a pivot in each row. We will discuss how to combine the above two solution algorithms in the simpler cases (I) and (II) to study the solvability of $A\mathbf{x} = \mathbf{b}$.

4.2.4 QR factorization

This is simply a way of organizing the computational outcome of the Gram-Schmidt Algorithm. Given $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \subset \mathbb{R}^m$. Apply the Gram-Schmidt Algorithm to $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$. Let $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ be the resulting orthonormal vectors, with r being the number of pivotal vectors in $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$. Then $r \leq k$, and each $\mathbf{u}_i = R_{1i}\mathbf{w}_1 + \ldots + R_{ii}\mathbf{w}_i$ for some coefficients $R_{1i}, \ldots, R_{ii}$; furthermore, $R_{ii} = 0$ if $\mathbf{u}_i$ is not a pivotal column (as in such a case, $\mathbf{u}_i$ would be a linear combination of $\{\mathbf{w}_1, \ldots, \mathbf{w}_{i-1}\}$). But these relations are simply encoded in the matrix equation

$$\begin{bmatrix}\mathbf{u}_1 & \ldots & \mathbf{u}_k\end{bmatrix} = \begin{bmatrix}\mathbf{w}_1 & \ldots & \mathbf{w}_r\end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & \ldots & R_{1r} & \ldots & R_{1k} \\ 0 & R_{22} & \ldots & R_{2r} & \ldots & R_{2k} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & R_{rr} & \ldots & R_{rk} \end{bmatrix}$$

If $\mathbf{u}_j$ is the l_jth pivotal column (e.g. if $\mathbf{u}_3$ is the 2nd pivotal column, then $l_3 = 2$), then $l_j \leq j$, $R_{l_j j} > 0$, and $R_{ij} = 0$, for $i > l_j$, and $R_{l_i k} = 0$ for $k < j$ (i.e., $R_{l_j j}$ is the first non-zero entry of R in the l_jth row). This shows that R is an echelon form. In the situation here, each row of R has a pivotal entry. Compare against **Lemma 13**. Set $Q = \begin{bmatrix}\mathbf{w}_1 & \ldots & \mathbf{w}_r\end{bmatrix}$, and

$$R = \begin{bmatrix} R_{11} & R_{12} & \ldots & R_{1r} & \ldots & R_{1k} \\ 0 & R_{22} & \ldots & R_{2r} & \ldots & R_{2k} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & R_{rr} & \ldots & R_{rk} \end{bmatrix},$$

then the columns of Q are orthonormal, which can be encoded in the matrix equation $Q^TQ = I_{r\times r}$, and R is an echelon form. This matrix product QR provides a QR-factorization of $\begin{bmatrix}\mathbf{u}_1 & \ldots & \mathbf{u}_k\end{bmatrix}$: set $A = \begin{bmatrix}\mathbf{u}_1 & \ldots & \mathbf{u}_k\end{bmatrix}$, then $A = QR$.

The factorization discussed above applies to any non-zero matrix A, and we will use this factorization $A = QR$ to solve for $A\mathbf{x} = \mathbf{b}$. If $A\mathbf{x} = \mathbf{b}$ and $A + QR$, then we set $R\mathbf{x} = \mathbf{y}$, and it follows that $Q\mathbf{y} = \mathbf{b}$. Conversely, if $\mathbf{y}$ satisfies $Q\mathbf{y} = \mathbf{b}$, and we can find $\mathbf{x}$ such that $R\mathbf{x} = \mathbf{y}$, then $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$. Thus we have reduced the solvability of $A\mathbf{x} = \mathbf{b}$ to that of two (simpler) systems: $Q\mathbf{y} = \mathbf{b}$ and $R\mathbf{x} = \mathbf{y}$. Note that once the first system has a solution, the second one will have a solution using the echelon form of R, so the key is the solvability of $Q\mathbf{y} = \mathbf{b}$.

Note that if A is $m \times k$, then Q is $m \times r$, $\mathbf{y} \in \mathbb{R}^r$. $Q\mathbf{y} = \mathbf{b}$ if and only if $\mathbf{b}$ is in the column space of Q, so the question boils down to whether a given $\mathbf{b}$ is in the column space of Q. Suppose it is: $\mathbf{b} = Q\mathbf{y}$, then, using $Q^TQ = I$ and multiplying to the left of both sides by Q^T, we find $\mathbf{y} = Q^TQ\mathbf{y} = Q^Tb$. We can then solve for $\mathbf{x}$ from $R\mathbf{x} = \mathbf{y}$ by back substitution.

Example 4.2.2

To solve

$$\begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix} \mathbf{x} = \mathbf{b},$$

we apply the Gram-Schmidt algorithm to the columns of A to obtain

$$\left\{ \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}, \begin{bmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{bmatrix}, \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix} \right\}$$

as the output orthonormal vectors. And the algorithm gives the relations

$$\begin{aligned} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} &= 3 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}, \\ \begin{bmatrix} -3 \\ 0 \\ 6 \end{bmatrix} &= 3 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} + 6 \begin{bmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{bmatrix}, \\ \begin{bmatrix} -2 \\ 5 \\ 5 \end{bmatrix} &= 6 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} + 3 \begin{bmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{bmatrix} + 3 \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix}. \end{aligned}$$

This gives the QR factorization:

$$A = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \\ 0 & 0 & 3 \end{bmatrix}.$$

Thus, to solve

$$A\mathbf{x} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \\ 0 & 0 & 3 \end{bmatrix} \mathbf{x} = \mathbf{b},$$

we first solve $\mathbf{y}$ from

$$Q\mathbf{y} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \mathbf{y} = \mathbf{b},$$

then solve for $\mathbf{x}$

$$\begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \\ 0 & 0 & 3 \end{bmatrix} \mathbf{x} = \mathbf{y}.$$

Since the columns of Q are orthonormal in $\mathbb{R}^3$, we obtain

$$\mathbf{y} = Q^T \mathbf{b} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \mathbf{b}.$$

This can also be seen, noting

$$y_1 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} + y_2 \begin{bmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{bmatrix} + y_3 \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix} = \mathbf{b},$$

so

$$y_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \cdot \mathbf{b}, \quad y_2 = \begin{bmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{bmatrix} \cdot \mathbf{b}, \quad y_3 = \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix} \cdot \mathbf{b}.$$

Once $\mathbf{y}$ is found, we can find $\mathbf{x}$ easily by backward substitution, starting from solving x_3 first.

If $k > r$, then some of the columns of A are non-pivotal. In solving $R\mathbf{x} = \mathbf{y}$, each variable x_i corresponding to a non-pivotal column can be assigned a value arbitrarily, so is called a free variable, and we solve for variables corresponding to pivotal columns in terms of these variables corresponding to a non-pivotal column. In the end we would obtain a solution containing a certain number of free variables, the number of which is the number of non-pivotal columns of A.

Example 4.2.3

If we modify the system in the previous example into

$$\begin{bmatrix} 1 & -3 & -2 & 0 \\ 2 & 0 & 5 & 3 \\ 2 & 6 & 5 & 6 \end{bmatrix} \mathbf{x} = \mathbf{b},$$

then the QR factorization of the coefficient matrix would give

$$\begin{bmatrix} 1 & -3 & -2 & 0 \\ 2 & 0 & 5 & 3 \\ 2 & 6 & 5 & 6 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & 3 & 6 & 6 \\ 0 & 6 & 3 & 3 \\ 0 & 0 & 3 & 0 \end{bmatrix},$$

and the system

$$\begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \mathbf{b}$$

is still solved in the same way. Once we have found y_1, y_2, y_3, in solving

$$\begin{bmatrix} 3 & 3 & 6 & 6 \\ 0 & 6 & 3 & 3 \\ 0 & 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix},$$

we can take x_4 to be a free variable, and solve for x_1, x_2, x_3 in terms of x_4 so this system has a one-parameter family of solutions for any given $\mathbf{b}$.

Example 4.2.4

If we modify the system from the earlier example into

$$\begin{bmatrix} 1 & -3 & 0 \\ 2 & 0 & 3 \\ 2 & 6 & 6 \end{bmatrix} \mathbf{x} = \mathbf{b},$$

then the QR factorization of the coefficient matrix would give

$$\begin{bmatrix} 1 & -3 & 0 \\ 2 & 0 & 3 \\ 2 & 6 & 6 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \end{bmatrix},$$

and we would need to solve

$$\begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{b}.$$

If we take $\mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix}$, the above system would have no solution (why?), so the

system

$$\begin{bmatrix} 1 & -3 & 0 \\ 2 & 0 & 3 \\ 2 & 6 & 6 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix}$$

has no solution.

The above discussion of solving $A\mathbf{x} = \mathbf{b}$ using the QR factorization of A is conditioned on the existence of a solution to $Q\mathbf{y} = \mathbf{b}$. If $\mathbf{b}$ is not in the column space of Q, then $Q\mathbf{y} = \mathbf{b}$ has no solution. Recall that our logic goes as follows. If $\mathbf{y}$ solves $Q\mathbf{y} = \mathbf{b}$, then $Q^TQ\mathbf{y} = Q^T\mathbf{b}$, which would lead to $\mathbf{y} = Q^T\mathbf{b}$ using $Q^TQ = I$. But, if we set $\mathbf{y} = Q^T\mathbf{b}$, does it solve $Q\mathbf{y} = \mathbf{b}$ automatically? Not necessarily! If we compute $Q\mathbf{y} = QQ^T\mathbf{b}$, so unless $QQ^T\mathbf{b} = \mathbf{b}$, we would not know that $Q\mathbf{y} = \mathbf{b}$. In such a case, what do we know about $\mathbf{b} - QQ^T\mathbf{b}$? It turns out that this vector is orthogonal to each column of Q! This is seen by

$$Q^T[\mathbf{b} - QQ^T\mathbf{b}] = Q^T\mathbf{b} - Q^TQQ^T\mathbf{b} = Q^T\mathbf{b} - Q^T\mathbf{b} = \mathbf{O}.$$

Also note that $QQ^T\mathbf{b}$ is in the column space of Q (why?), so it is the orthogonal projection of $\mathbf{b}$ in the column space of Q. In summary,

If the columns of the $m \times r$ matrix Q are orthonormal, then, for any given $\mathbf{b} \in \mathbb{R}^m$, $QQ^T\mathbf{b}$ is in the column space of Q, and $\mathbf{b} - QQ^T\mathbf{b}$ is orthogonal to each column of Q. As a consequence, for any vector $\mathbf{v}$ in the column space of Q,

$$\|\mathbf{b} - \mathbf{v}\| \geq \|\mathbf{b} - QQ^T\mathbf{b}\|.$$

In other words, $\mathbf{b}_\perp := QQ^T\mathbf{b}$ is the best approximation to $\mathbf{b}$ among vectors in the column space of Q, and $\mathbf{y} = Q^T\mathbf{b}$ is a solution to $Q\mathbf{y} = \mathbf{b}_\perp$, instead of to $Q\mathbf{y} = \mathbf{b}$. This $\mathbf{y}$ is also called the least square solution to $Q\mathbf{y} = \mathbf{b}$, with the square here referring to $\|\mathbf{b} - \mathbf{b}_\perp\|$ defined as $\sqrt{(\mathbf{b} - \mathbf{b}_\perp) \cdot (\mathbf{b} - \mathbf{b}_\perp)}$. The rule for finding this $\mathbf{y}$ is also simple: it is such that $Q\mathbf{y} - \mathbf{b}$ is orthogonal to each column of Q; and this condition is encoded in the matrix equation:

$$Q^T(Q\mathbf{y} - \mathbf{b}) = \mathbf{0}.$$

This analysis works when Q is replaced by any matrix whose rank equals the number of its columns.

Example 4.2.5

Suppose $Q = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix}$ and $R = \begin{bmatrix} 1 & 4 \\ 0 & -2 \end{bmatrix}$, and we are interested in solving

$$\begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 0 & -2 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

We need to set up $\mathbf{y} = \begin{bmatrix} 1 & 4 \\ 0 & -2 \end{bmatrix} \mathbf{x}$, and try to first solve for $\mathbf{y}$ from

$$\begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \mathbf{y} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}. \tag{4.1}$$

This is a system of three linear equations for the two components of $\mathbf{y}$. It has a solution only if the vector on the right hand side is in the column space of Q. The columns of Q are orthonormal, but there is no direct way to tell whether a particular vector is in the column space of Q, unless another method is employed. If the system has a solution $\mathbf{y}$, then multiplying by Q^T on the left side of both sides of the equation, and using $Q^T Q = I_2$, we would get

$$\mathbf{y} = Q^T \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ -\frac{7}{3} \end{bmatrix}.$$

We can now check whether $Q\mathbf{y} - \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \mathbf{0}$. But we find

$$Q\mathbf{y} = Q \begin{bmatrix} -\frac{1}{3} \\ -\frac{7}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} -\frac{1}{3} \\ -\frac{7}{3} \end{bmatrix} = \begin{bmatrix} \frac{13}{9} \\ -\frac{16}{9} \\ \frac{5}{9} \end{bmatrix} \neq \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

So the system $Q\mathbf{y} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$ has no solution. The best we can do is to make $\| Q\mathbf{v} - \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \|$ as small as possible. Our analysis shows that for any vector

$\mathbf{v} \in \mathbb{R}^2$,

$$\|Q\mathbf{v} - \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}\| \geq \| \begin{bmatrix} \frac{13}{9} \\ -\frac{16}{9} \\ \frac{5}{9} \end{bmatrix} - \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \|,$$

and equality occurs when $\mathbf{v} = \mathbf{y} = Q^T \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ -\frac{7}{3} \end{bmatrix}$. In one needs to find $\mathbf{x}$ which gives rise to this least square approximation, one can solve for $\mathbf{x}$ from $\mathbf{y} = R\mathbf{x}$, and it will be called the least square approximate solution to the given system.

The computation above amounts to checking whether

$$Q\mathbf{y} = QQ^T \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

But

$$QQ^T = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{5}{9} & -\frac{2}{9} & \frac{4}{9} \\ -\frac{2}{9} & \frac{8}{9} & \frac{2}{9} \\ \frac{4}{9} & \frac{2}{9} & \frac{5}{9} \end{bmatrix},$$

and

$$\begin{bmatrix} \frac{5}{9} & -\frac{2}{9} & \frac{4}{9} \\ -\frac{2}{9} & \frac{8}{9} & \frac{2}{9} \\ \frac{4}{9} & \frac{2}{9} & \frac{5}{9} \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{13}{9} \\ -\frac{16}{9} \\ \frac{5}{9} \end{bmatrix} \neq \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

Note that $QQ^T \neq I_3$, as a result $QQ^T\mathbf{b} \neq \mathbf{b}$ for some vectors $\mathbf{b}$. However, $QQ^T\mathbf{b} = \mathbf{b}$ when $\mathbf{b}$ is a vector in the column space of Q (check this when $\mathbf{b}$ is a column vector of Q). QQ^T is an example of an orthogonal projection matrix.

When a system $A\mathbf{x} = \mathbf{b}$ has no solution, it is called inconsistent. In such a case, one would like to find some $\mathbf{x}$ to make $\|A\mathbf{x} - \mathbf{b}\|$ as small as possible.

Inconsistent systems arise in many applications, in particular in data fitting. One often expects two variables to have a linear relation; specifically, one proposes a linear relation of the form $y = a + bx$ between two variables x and y. But real time data do not satisfy such a relation strictly; instead, one may obtain a collection of *observed data* $(x_1, y_1^{\text{obs}}), \cdots, (x_n, y_n^{\text{obs}})$, where the observed values of y_j^{obs} likely differ from the predicted values $a + bx_j$. Our task is to find the values of the parameters a and b which gives the **least error sum of squares**

$$E = [y_1^{\text{obs}} - (a + bx_1)]^2 + \cdots + [y_n^{\text{obs}} - (a + bx_n)]^2$$

between the observed data and the data *predicted* by the linear relation $y = a + bx$.

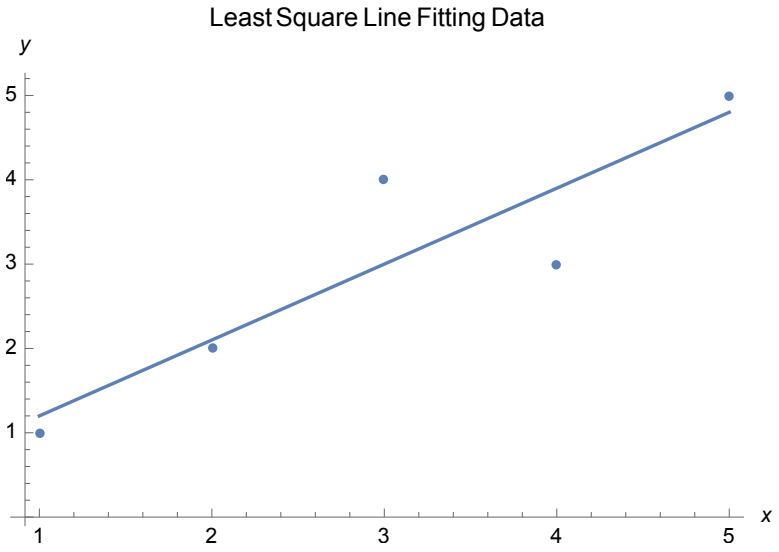


Figure 4.1: Given some data points, find a line which produces the least error sum of squares

Setting

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1^{\text{obs}} \\ y_2^{\text{obs}} \\ \vdots \\ y_n^{\text{obs}} \end{bmatrix},$$

the above question becomes one of finding a and b which minimizes $||\mathbf{y} - (a\mathbf{v}_1 + b\mathbf{v}_2)||$, the solution to which is given by

$$A \begin{bmatrix} a \\ b \end{bmatrix} = P_W(\mathbf{y}), \text{ the orthogonal projection of } \mathbf{y} \text{ in } W, \tag{4.2}$$

where $A = [\mathbf{v}_1\, \mathbf{v}_2]$ and $W = \text{Col}(A) = \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$, and based on our discussion, $P_W(\mathbf{y}) = A \begin{bmatrix} a \\ b \end{bmatrix}$ can be given by the condition

$$A^T \left(A \begin{bmatrix} a \\ b \end{bmatrix} - \mathbf{y} \right) = \mathbf{0},$$

which is the same type of equation that we discussed above.

From the point of view of solving a system of linear equations, it is unlikely for $a\mathbf{v}_1 + b\mathbf{v}_2 = \mathbf{y}$ to have a solution, as a solution would mean that $E = 0$ and *all* observed values of y_j^{obs} match exactly the predicted values $a + bx_j$.

We now discuss how to find $\mathbf{x}$ to make $\|A\mathbf{x} - \mathbf{b}\|$ smallest possbiel. Note that $A\mathbf{x}$ is a column vector of A, so if there exists some $\mathbf{x}$ such that $\|A\mathbf{x} - \mathbf{b}\| \leq \|A\mathbf{v} - \mathbf{b}\|$ for any choice of $\mathbf{v}$, the vector $A\mathbf{x} - \mathbf{b}$ must be orthogonal to all columns of A. This can be encoded into the matrix equation $A^T(A\mathbf{x} - \mathbf{b}) = \mathbf{0}$. In summary, if $\mathbf{x}$ solves $A^TA\mathbf{x} = A^T\mathbf{b}$, then $A\mathbf{x}$ provides the least square approximation to $\mathbf{b}$ among vectors of the form $A\mathbf{v}$.

What remains is to develop a good algorithm to solve $A^TA\mathbf{x} = A^T\mathbf{b}$. There are different approaches to solving this system. For a small matrix A, one can solve $A^TA\mathbf{x} = A^T\mathbf{b}$ directly. Below we discuss how to use the QR factorization of $A = QR$ to solve $A^TA\mathbf{x} = A^T\mathbf{b}$. Since the column space of A is the same as the column space of Q, the condition that $A\mathbf{x} - \mathbf{b}$ must be orthogonal to all columns of A is equivalent to $A\mathbf{x} - \mathbf{b}$ must be orthogonal to all columns of Q, which can be written in the matrix equation $Q^T(A\mathbf{x} - \mathbf{b}) = \mathbf{0}$. Using $Q^TQ = I$, we see that $Q^TA = Q^TQR = R$, so the above system is equivalent to $R\mathbf{x} = Q^T\mathbf{b}$. Since R is an upper triangular matrix, this system is easily solved.

Example 4.2.6

The above discussion is for a general matrix with a QR factorization. For the specific problem of least square fitting line above, the matrix A is an $n \times 2$ matrix with specific column vectors $\mathbf{v}_1, \mathbf{v}_2$, with the properties that

$$\|\mathbf{v}_1\| = \sqrt{n}, \quad \mathbf{v}_1 \cdot \mathbf{v}_2 = x_1 + \cdots + x_n = n\bar{x},$$

where $\bar{x} = (x_1 + \cdots + x_n)/n$ is the average of the input data $x_1, \cdots, x_n$. We could work directly with the 2×2 system $A^TA\mathbf{x} = A^T\mathbf{b}$. But we choose to illustrate how to use the QR factorization to solve the problem.

$$\mathbf{u}_1 = \frac{1}{\sqrt{n}}\mathbf{v}_1 \text{ is the unit vector in the direction of } \mathbf{v}_1,$$

and

$$\mathbf{u}_2 \text{ is the unit vector in the direction of } \mathbf{v}_2 - \frac{\mathbf{v}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 = \mathbf{v}_2 - \bar{x}\mathbf{v}_1.$$

Since

$$\|\mathbf{v}_2 - \bar{x}\mathbf{v}_1\| = \sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2} = \sigma_{\mathbf{x}},$$

where $\sigma_{\mathbf{x}}$ is the standard deviation of the input data $x_1, \cdots, x_n$. $\sigma_{\mathbf{x}} = 0$ if and only if each $x_i = \bar{x}$. Let's assume that we are not in such a situation, so $\sigma_{\mathbf{x}} > 0$,

$$\mathbf{u}_2 = \frac{1}{\sigma_{\mathbf{x}}} \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix},$$

and

$$A = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \sqrt{n} & \sqrt{n}\bar{x} \\ 0 & \sigma_{\mathbf{x}} \end{bmatrix}.$$

Thus we need to solve

$$\begin{bmatrix} \sqrt{n} & \sqrt{n}\bar{x} \\ 0 & \sigma_{\mathbf{x}} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix}^T \mathbf{y} = \begin{bmatrix} \mathbf{u}_1^T \mathbf{y} \\ \mathbf{u}_2^T \mathbf{y} \end{bmatrix}.$$

Note that

$$\begin{aligned} \mathbf{u}_1^T \mathbf{y} &= \frac{y_1 + \cdots + y_n}{\sqrt{n}} = \sqrt{n}\bar{y} \quad \text{with } \bar{y} = \frac{y_1 + \cdots + y_n}{n}, \\ \mathbf{u}_2^T \mathbf{y} &= \frac{(x_1 - \bar{x})y_1 + \cdots + (x_n - \bar{x})y_n}{\sigma_{\mathbf{x}}} \\ &= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sigma_{\mathbf{x}}} \\ &= \sigma_{\mathbf{y}} \mathrm{Cor}(\mathbf{x}, \mathbf{y}), \end{aligned}$$

using $(x_1 - \bar{x})\bar{y} + \cdots + (x_n - \bar{x})\bar{y} = 0$, and where $\sigma_{\mathbf{y}} = \sqrt{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}$ is the standard deviation of the data $y_1, \cdots, y_n$, and

$$\mathrm{Cor}(\mathbf{x}, \mathbf{y}) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}}$$

is the correlation coefficient between the data $\mathbf{x}$ and $\mathbf{y}$.
We can now conclude that

$$b = \frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}} \mathrm{Cor}(\mathbf{x}, \mathbf{y}) \text{ and } a = \bar{y} - b\bar{x}.$$

This results in the best fitting line to be

$$y = \bar{y} + b(x - \bar{x}), \text{ with } b = \frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}} \mathrm{Cor}(\mathbf{x}, \mathbf{y}).$$

Reading Quizzes/Questions:

(i) Find the least square line of fit to the set of five data points in the figure above: $\{(1,1),(2,2),(3,4),(4,4),(5,5)\}$.

(ii) If the columns of Q are orthonormal, and $Q\mathbf{y} = \mathbf{b}$, does it necessarily follow that $\mathbf{y} = Q^T\mathbf{b}$?

(iii) If the columns of Q are orthonormal, and $\mathbf{y} = Q^T\mathbf{b}$, does it necessarily follow that $Q\mathbf{y} = \mathbf{b}$?

(iv) If $A\mathbf{x}-\mathbf{b}$ is orthogonal to each column of A, why does it follow that $\|A\mathbf{x}-\mathbf{b}\| \leq \|A\mathbf{v}-\mathbf{b}\|$ for any $\mathbf{v}$?

(v) In the QR factorization, R has a pivot in each of its rows. If $R^T\mathbf{u} = R^T\mathbf{v}$, does it necessarily follow that $\mathbf{u} = \mathbf{v}$?

The QR factorization $A = QR$ can also be used to read off information about A from those of R or Q.

(a). $\text{Null}(A) = \text{Null}(R)$.

(b). $\text{Row}(A) = \text{Row}(R)$.

(c). $\text{Col}(A) = \text{Col}(Q)$.

(d). $\dim \text{Row}(A) = \dim \text{Row}(R) = r = \dim \text{Col}(A)$, namely, the dimensions of the row space and column space of A are equal, which is the fundamental theorem of linear algebra.

Proof. For (a), if $\mathbf{x} \in \text{Null}(R)$, then $\mathbf{x}$ satisfies $R\mathbf{x} = \mathbf{0}$, and it follows that $A\mathbf{x} = QR\mathbf{x} = \mathbf{0}$, so $\mathbf{x} \in \text{Null}(A)$. Conversely, if $\mathbf{x} \in \text{Null}(A)$, then $A\mathbf{x} = \mathbf{0}$, and it follows that $QR\mathbf{x} = \mathbf{0}$. Multiplying both sides by Q^T and using $Q^TQ = I_{r\times r}$, we obtain $R\mathbf{x} = Q^TQR\mathbf{x} = Q^T\mathbf{0} = \mathbf{0}$. Thus $\mathbf{x} \in \text{Null}(R)$.

For (b), if $\mathbf{v} \in \text{Row}(A)$, then $\mathbf{v} = x_1\text{Row}_1(A) + \ldots + x_m\text{Row}_m(A)$ for some coefficients $x_1, \ldots, x_m$. This is written as $\mathbf{v} = [x_1 \ldots x_m]A = [x_1 \ldots x_m]QR$. Setting $[x_1 \ldots x_m]Q = [y_1 \ldots y_m]$, then $\mathbf{v} = [y_1 \ldots y_m]R$, implying that $\mathbf{v}$ is a linear combination of rows of R. Conversely, if $\mathbf{v} = [y_1 \ldots y_m]R$ for some coefficients $y_1, \ldots, y_m$, we set $[x_1 \ldots x_m] = [y_1 \ldots y_m]Q^T$. Using $Q^TQ = I_{r\times r}$ again, we have

$[x_1 \dots x_m]A = [y_1 \dots y_m]Q^TQR = [y_1 \dots y_m]R = \mathbf{v}$, implying that $\mathbf{v}$ is in the two space of A.

(c) is really a re-statement of one part of the Gram-Schmidt algorithm.

(d) is a direct consequence of (a), (b), and (c), as $r = \dim \mathrm{Col}(A) = \#$ columns of Q by definition, and $\dim \mathrm{Row}(A) = \dim \mathrm{Row}(R) = \#$ of pivots in $R = \#$ of rows of $R = \#$ columns of Q, which is r. We also see that

$$n - \dim \mathrm{Col}(A) = n - r = \dim \mathrm{Null}(A) = \# \text{ of free variables in solving } A\mathbf{x} = \mathbf{0}.$$

□

The discussion on pivots also reveals a useful relation: The columns of A split into pivotal and non-pivotal ones, the number of pivotal ones corresponds to the rank of A (also the dimension of the column space of A), while the number of non-pivotal ones corresponds to the dimension of the null space of A. So

> Rank of A + dimension of the null space of A = the number of columns of A.

Reading Quizzes/Questions: Investigate the following True or False questions:

(i). If the columns of Q are orthonormal, then $QQ^T = I$.

(ii). If the rank of a matrix A is less than the number of its columns, then its null space is non-trivial.

(iii). Is it true that a 2×3 matrix must have a non-pivotal column? Why?

(iv). Can one find 4 pivotal columns in a 3×5 matrix? Why?

(v). Is it true that $\mathrm{Null}(A) = \mathrm{Null}(A^TA)$ always holds?

(vi). Is it true that $\mathrm{Rank}(A) = \mathrm{Rank}(A^TA)$ always holds?

(vii). Is it true that $\mathrm{Rank}(AA^T) = \mathrm{Rank}(A^TA)$ always holds?

(viii). Is it true that dimension of the null space of AA^T = dimension of the null space of A^TA always holds?

4.2.5 Matrix inverses

Professor Carlen uses the more abstract concept of linear transformations and their inverse to discuss the inverse of a matrix. He does not give an explicit definition of the inverse of a matrix; implicit in his discussion is the following

Definition 4.2.1

An $n \times n$ matrix A is invertible if there exists an $n \times n$ matrix B such that $BA = I_n$ and $AB = I_n$.

Professor Carlen exploits the first relation $BA = I_n$: if $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ denote the rows of B, and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ denote the columns of A, then $BA = I_n$ is equivalent to $\mathbf{w}_i \cdot \mathbf{v}_j = 1$ for $i = j$, and $= 0$ for $i \neq j$. When such a B exists, he uses an abstract argument to imply that $AB = I_n$ automatically holds.

If we exploit $AB = I_n$, then, with $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ denoting the jth column of B, $AB = I_n$ is equivalent to $x_{1j}\mathrm{Col}_1(A) + x_{2j}\mathrm{Col}_2(A) + \ldots + x_{nj}\mathrm{Col}_n(A) = \mathbf{e}_j$, namely, $A\mathbf{x}_j = \mathbf{e}_j$, for each $1 \leq j \leq n$. Thus constructing some matrix B satisfying $AB = I_n$ is equivalent to solving $A\mathbf{x}_j = \mathbf{e}_j$, for each $1 \leq j \leq n$. This gives an algorithm to find whether $\mathbf{x}_j$ exists and compute it when it exists. This approach still does not answer whether the B constructed this way, when it exists, also satisfies $BA = I_n$.

That requires a separate argument based on

(i). If there exist $n \times n$ matrices B and C such that $BA = I_n$, and $AC = I_n$, then $B = C$.

(ii). If there exists an $n \times n$ matrix B such that $BA = I_n$, then there exists an $n \times n$ matrix C such that $AC = I_n$.

(iii). If If there exists an $n \times n$ matrix C such that $AC = I_n$, then there exists an $n \times n$ matrix B such that $BA = I_n$.

In summary, if A has an inverse, then it has a unique inverse; and B is the inverse, if and only if it satisfies $AB = I_n$ or $BA = I_n$. Due to the uniqueness of inverse, when it exists, we denote the inverse of A by A^{-1}.

(i) follows easily: $B = BI_n = B(AC) = (BA)C = I_nC = C$. For (ii), if there exists an $n \times n$ matrix B such that $BA = I_n$, then the span of the rows of A is the full $\mathbb{R}^n$. By the fundamental theorem of linear algebra, the dimension of the column space of A is n, which means that we can find an $n \times n$ matrix C solving $AC = I_n$.

A direct argument for (ii) goes as follows. $BA = I_n$ implies that the only vector

$\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$, for, $A\mathbf{x} = \mathbf{0}$ implies $B(A\mathbf{x}) = \mathbf{0}$, but that means $\mathbf{0} = (BA)\mathbf{x} = I_n\mathbf{x} = \mathbf{x}$. We now claim that the rank of A is n, namely, each column of A is a pivot column (in the Gram-Schmidt Orthogonalization process). If A has a non-pivotal column, say, jth column, then $\text{Col}_j(A) = c_1\text{Col}_1(A) + \ldots + c_{j-1}\text{Col}_{j-1}(A)$ for some coefficients $c_1, \ldots, c_{j-1}$. But this would mean

$$A\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

making $A\mathbf{x} = \mathbf{0}$ having a non-zero solution! This contradiction shows that the rank of A is n, therefore we can solve for C such that $AC = I_n$.

For (iii), the condition implies that $A\mathbf{x} = \mathbf{b}$ always has a solution for any $\mathbf{b}$, for $A(C\mathbf{b}) = (AC)\mathbf{b} = I_n\mathbf{b} = \mathbf{b}$. It follows that the column space of A is $\mathbb{R}^n$. By the fundamental theorem of linear algebra (**Theorem 50**), the row space of A must also be $\mathbb{R}^n$. So any row vector e_i must be a linear combination of row vectors of A: $e_i = b_{i1}\text{Row}_1(A) + \ldots + b_{in}\text{Row}_n(A)$, for $i = 1, 2, \ldots, n$. But these equations precisely say $BA = I_n$, if we use $[b_{i1} \ldots b_{in}]$ as the ith row of B.

Example 4.2.7

Although computing packages can compute the inverse of a matrix easily, it is important to understand the underlying tasks involved. Based on our discussion, to compute the inverse of $\begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix}$, we need to find a 3×3 matrix B such that

$$\begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix} B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

If the three column vectors of B are $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, respectively, then we need to solve

$$\begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix}\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix}\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix}\mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

We can now use the QR factorization

$$\begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \\ 0 & 0 & 3 \end{bmatrix}$$

of the coefficient matrix to solve these three systems.
In most elementary linear algebra courses, these systems are solved simultaneously using the Gauss-Jordan elimination method by performing elementary row operations. Carlen takes a different approach, focusing on using the QR factorization of the coefficient matrix and reduce the problem to solving two simpler systems, one involving a linear system with an orthogonal matrix as the coefficient matrix, and another involving a linear system with an upper triangular matrix in echelon form as the coefficient matrix.
For those who know the Gauss-Jordan elimination method, the direct application of the method would require performing elementary row operations every time the right hand side vector is changed, which wastes a lot of computational resources. If we carry out the QR factorization of the coefficient matrix, and store these two factor matrices, then, when Q is full rank, for any given right hand side vector $\mathbf{b}$, we only need to carry out $\mathbf{y} = Q^T\mathbf{b}$, and solve for $\mathbf{x}$ from $R\mathbf{x} = \mathbf{y}$. A certain computational resources are still needed, but at some savings.
We also take this opportunity to remind the reader that

$$\begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}^{T} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix},$$

and

$$\begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \\ 0 & 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -3 & -2 \\ 2 & 0 & 5 \\ 2 & 6 & 5 \end{bmatrix}^{-1}.$$

When we discuss eigenvalues and eigenvectors later, we need a criterion to characterize when a square matrix A is not invertible. Based on our discussion, A is not invertible, if and only if A does not have full rank, i.e., it has at least one non-pivotal column. But that means there exists some coefficients $x_1, \ldots, x_n$, *not all zero*, such that $x_1\text{Col}_1(A) + \ldots + x_n\text{Col}_n(A) = \mathbf{0}$. In summary,

Matrix A is not invertible, if and only if there exists some $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{0}$.
Equivalently, A is invertible if and only if the only solution to $A\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.

When $n = 2$, this criterion means that one column of $A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$ is a multiple of the other column. Algebraically this condition can be characterized as $ad - bc = 0$.

When $n = 3$, this criterion means that one column of $A = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}$ is a linear combination of the other two columns, say, $\mathbf{v}_1 = a\mathbf{v}_2 + b\mathbf{v}_3$. A geometric way to describe this without involving the coefficients a and b (to be worked out) is that $\mathbf{v}_1 \perp \mathbf{v}_2 \times \mathbf{v}_3$. But an algebraic way to express this relation is $\mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3) = 0$. To summarize, $A = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}$ is not invertible, if and only if $\mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3) = 0$. A similar criterion can be formulated in terms of the rows $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ of A, namely, $\mathbf{r}_1 \cdot (\mathbf{r}_2 \times \mathbf{r}_3) = 0$. It turns out that $\mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3) = \mathbf{r}_1 \cdot (\mathbf{r}_2 \times \mathbf{r}_3)$ for all 3×3 matrices, and it is called the determinant of A. Note also that $\mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3) = \mathbf{v}_3 \cdot (\mathbf{v}_1 \times \mathbf{v}_2)$.

Let's confirm this for a 3×3 matrix A whose entries are labeled as a_{ij}, $1 \leq i, j \leq 3$. Note that

$$\mathbf{v}_2 \times \mathbf{v}_3 = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} \times \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix} = \begin{bmatrix} a_{22}a_{33} - a_{32}a_{23} \\ a_{32}a_{13} - a_{12}a_{33} \\ a_{12}a_{23} - a_{22}a_{13} \end{bmatrix},$$

so

$$\mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3) = a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{21}(a_{32}a_{13} - a_{12}a_{33}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}).$$

On the other hand,

$$\mathbf{r}_2 \times \mathbf{r}_3 = \begin{bmatrix} a_{21} & a_{22} & a_{23} \end{bmatrix} \times \begin{bmatrix} a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{22}a_{33} - a_{32}a_{23} & a_{23}a_{31} - a_{21}a_{33} & a_{21}a_{32} - a_{22}a_{31} \end{bmatrix},$$

so

$$\mathbf{r}_1 \cdot (\mathbf{r}_2 \times \mathbf{r}_3) = a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) = \mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3).$$

Example 4.2.8

The 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - t \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a - t & b \\ c & d - t \end{bmatrix}$$

is not invertible when

$$\det \begin{bmatrix} a - t & b \\ c & d - t \end{bmatrix} = (a - t)(d - t) - bc = t^2 - (a + d)t + ad - bc = 0.$$

The 3×3 matrix

$$\begin{bmatrix} 3 & 3 & 6 \\ 0 & 6 & 3 \\ 0 & 0 & 3 \end{bmatrix} - tI_3 = \begin{bmatrix} 3-t & 3 & 6 \\ 0 & 6-t & 3 \\ 0 & 0 & 3-t \end{bmatrix}$$

is not invertible when

$$\det \begin{bmatrix} 3-t & 3 & 6 \\ 0 & 6-t & 3 \\ 0 & 0 & 3-t \end{bmatrix} = (3-t)[(6-t)(3-t) - 0 * 3] = (3-t)^2(6-t) = 0.$$

In the above since the first column only has its first entry non-zero, in computing $\mathbf{r}_2 \times \mathbf{r}_3$, we only need to work out its first entry $a_{22}a_{33} - a_{32}a_{23}$.

Two other useful properties about the inverse of a matrix are

If A is invertible, then so is A^T, and $(A^T)^{-1} = (A^{-1})^T$.
If A_1 and A_2 are invertible $n \times n$ matrices, then so is A_1A_2, and $(A_1A_2)^{-1} = A_2^{-1}A_1^{-1}$.

Reading Quizzes/Questions: Investigate the following True or False questions about an $n \times n$ square matrix A:

(a). A is invertible if and only if $A\mathbf{x} = \mathbf{e}_j$ has a solution, for each $1 \leq j \leq n$.

(b). A is invertible if and only if $A\mathbf{x} = \mathbf{b}$ has a solution for any vector $\mathbf{b} \in \mathbb{R}^n$.

(c). If A is invertible, then the column space of A is $\mathbb{R}^n$.

(d). If the column space of A is $\mathbb{R}^n$, then A is invertible.

(e). If the null space of A is $\{\mathbf{0}\}$, then A is invertible.

(f). If an $n \times n$ matrix B satisfies $AB = I_n$, then $BA = I_n$.

(g). If A_1 and A_2 are invertible $n \times n$ matrices, then so is $A_1 + A_2$, and $(A_1 + A_2)^{-1} = A_2^{-1} + A_1^{-1}$.

(h). If A is an $n \times n$ matrix, then $A\mathbf{x} = \lambda\mathbf{x}$ has a non-zero solution $\mathbf{x}$ if and only if the matrix $A - \lambda I_n$ is not invertible.

4.2.6 Continuity of matrix inverses

Our focus here is to discuss the Frobenius norm of a matrix and its properties. A norm of a matrix is used to measure its size. There are many ways to define a norm of a matrix. E.g., for the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we could use any of the following three quantities to measure its size:

$$\begin{aligned} \|A\|_1 &:= |a| + |b| + |c| + |d|, \text{ or} \\ \|A\|_2 &:= \sqrt{|a|^2 + |b|^2 + |c|^2 + |d|^2}, \text{ or} \\ \|A\|_\infty &:= \max\{|a|, |b|, |c|, |d|\} \end{aligned}$$

They all share the following three properties required for the notion of a norm:

(a). $\|A\| \geq 0$ for any A, and $= 0$ only when $A = O$;

(b). $\|cA\| = |c|\|A\|$ for A and any scalar c;

(c). $\|A + B\| \leq \|A\| + \|B\|$ for any A, B.

But for matrices, we often would like an additional property:

$$\|A\mathbf{v}\| \leq \|A\|\|\mathbf{v}\| \text{ for any matrix } A \text{ and any vector } \mathbf{v} \text{ such that } A\mathbf{v} \text{ is defined.} \quad (4.3)$$

This then implies that whenever $AB\mathbf{v}$ is defined, we have

$$\|AB\mathbf{v}\| = \|A(B\mathbf{v})\| \leq \|A\|\|B\mathbf{v}\| \leq \|A\|\|B\|\|\mathbf{v}\|.$$

In this course we are using the Euclidean norm for vectors. Then $\|A\|_2$ has this property. This norm is called the Frobenius norm of matrix A, and is also denoted as $\|A\|_{\mathcal{F}}$. This is seen by noting that the i-th component of $A\mathbf{v}$ is $\text{Row}_i(A) \cdot \mathbf{v}$, and by the Cauchy-Schwarz inequality, $|\text{Row}_i(A) \cdot \mathbf{v}|^2 \leq \|\text{Row}_i(A)\|^2\|\mathbf{v}\|^2$, so

$$\|A\mathbf{v}\|^2 = \sum_{i=1}^{m} |\text{Row}_i(A) \cdot \mathbf{v}|^2 \leq \sum_{i=1}^{m} \|\text{Row}_i(A)\|^2\|\mathbf{v}\|^2 \leq \|A\|_{\mathcal{F}}^2\|\mathbf{v}\|^2,$$

as $\|\mathbf{v}\|^2 = \sum_{i=1}^{m} \|\text{Row}_i(A)\|^2$.

It turns out there is a way to define a norm for a matrix using (4.3). For any $m \times n$ matrix A, we define its operator norm, denoted as $\|A\|_{\text{op}}$, as

$$\|A\|_{\text{op}} := \sup_{\mathbf{u}:\|\mathbf{u}\|=1} \|A\mathbf{u}\|.$$

Then for any $\mathbf{v} \neq \mathbf{0}$, we take $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|$, then $\|\mathbf{u}\| = 1$, so $\|A\mathbf{u}\| \leq \|A\|_{\text{op}}$. But $\|A\mathbf{u}\| = \|A(\frac{\mathbf{v}}{\|\mathbf{v}\|})\| = \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|}$, so we see that $\|A\mathbf{v}\| \leq \|A\|_{\text{op}}\|\mathbf{v}\|$. From this discussion we see that $\|A\|_{\text{op}}$ satisfies (4.3), and it is $\leq \|A\|_2$. We haven't checked (a)–(c) for $\|A\|_{\text{op}}$, but that is fairly routine.

By its definition $\|A\|_{\text{op}}$ measures the maximum stretching factor when A multiplies to a vector $\mathbf{v}$; it is the optimal number for (4.3) to hold for all $\mathbf{v}$, but we often use $\|A\|_2$ as the latter can be computed directly, while $\|A\|_{\text{op}}$ is often not easy to compute, and using $\|A\|_2$ in the relation $\|A\|_{\text{op}} \leq \|A\|_2$ to estimate $\|A\|_{\text{op}}$ often suffices for our purposes.

Using **Theorem 39**, we can prove that $\|A\|_{\text{op}}$ is attained by some vector $\mathbf{w}$: $\|A\mathbf{w}\| = \|A\|_{\text{op}}$, and $\|\mathbf{w}\| = 1$, so the $\sup_{\mathbf{u}:\|\mathbf{u}\|=1} \|A\mathbf{u}\|$ is really $\max_{\mathbf{u}:\|\mathbf{u}\|=1} \|A\mathbf{u}\|$. In a later section, we will see that $\|A\|_{\text{op}}$ is the eigenvalue of a matrix derived from A.

Example 4.2.9

For $A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $A_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, we see that $\|A_1\|_2 = \|A_2\|_2 = \sqrt{2}$, $\|A_1\|_{\text{op}} = 1$, but $\|A_2\|_{\text{op}} = \sqrt{2}$, since $\|A_2\mathbf{v}\| = \sqrt{2}|v_1|$ for any $\mathbf{v} = (v_1, v_2)$.

Lastly, using any norm for matrices, we can measure the distance between two matrices A and B by $\|A - B\|$. Since the addition of two $m \times n$ matrices is well defined, so is the scalar multiplication of a matrix, so we can treat the set of all $m \times n$ matrices as the set of vectors with mn components. In fact, we can simply treat it as $\mathbb{R}^{mn}$, as the Frobenius norm of an $m \times n$ matrix is simply the Euclidean norm when it is treated as a vector of $\mathbb{R}^{mn}$. Furthermore, we can discuss continuity, or even differentiability, of functions defined on the set of matrices.

4.3 Differentiability of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$

4.3.1 Differentiability and best linear approximation in several variables

The central questions of this subsection are

(i). What does differentiability of $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ at $\mathbf{x}_0$ mean?

(ii). What is the derivative or Jacobian matrix of $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ at $\mathbf{x}_0$? What is it used for?

(iii). What is an easy-to-use criterion to check the differentiability of $f : \mathbb{R}^n \mapsto$

$\mathbb{R}^m$ at $\mathbf{x}_0$?

The differentiability of $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ at $\mathbf{x}_0$ is defined in terms of (4.58), or equivalently, (4.59). We work it out in the case that $m = 2$, $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$.

If (4.59) holds for some matrix A, A would be a $2 \times n$ matrix. Let $\mathbf{a}_1$ and $\mathbf{a}_2$ be its two rows, then $A(\mathbf{x} - \mathbf{x}_0) = \begin{bmatrix} \mathbf{a}_1 \cdot (\mathbf{x} - \mathbf{x}_0) \\ \mathbf{a}_2 \cdot (\mathbf{x} - \mathbf{x}_0) \end{bmatrix}$, and

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0) = \begin{bmatrix} f_1(\mathbf{x}) - f_1(\mathbf{x}_0) - \mathbf{a}_1 \cdot (\mathbf{x} - \mathbf{x}_0) \\ f_2(\mathbf{x}) - f_2(\mathbf{x}_0) - \mathbf{a}_2 \cdot (\mathbf{x} - \mathbf{x}_0) \end{bmatrix},$$

so

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0,$$

implies that

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{|f_i(\mathbf{x}) - f_i(\mathbf{x}_0) - \mathbf{a}_i \cdot (\mathbf{x} - \mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0,$$

for $i = 1, 2$. But this is simply the differentiability of $f_i(\mathbf{x})$ at $\mathbf{x}_0$.

This computation works for any m. In summary,

$f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, then each of its component f_i is differentiable at $\mathbf{x}_0$, and $\nabla f_i(\mathbf{x}_0)$ should be the ith row of the matrix A in (4.58), namely, the derivative or Jacobian matrix $[D\mathbf{f}(\mathbf{x}_0)]$ of $\mathbf{f}$ at $\mathbf{x}_0$ is the $m \times n$ matrix whose ith row is simply the gradient vector $\nabla f_i(\mathbf{x}_0)$ of f_i at $\mathbf{x}_0$—review (4.62), and this discussion also shows how to compute $[D\mathbf{f}(\mathbf{x}_0)]$. The converse also holds, namely, if each of its component f_i is differentiable at $\mathbf{x}_0$, then $\mathbf{f}$ is differentiable at $\mathbf{x}_0$.

How does one interpret the columns of the Jacobian matrix $[D\mathbf{f}(\mathbf{x}_0)]$? The j-column is simply $\dfrac{\partial \mathbf{f}}{\partial x_j}$, which is the tangent vector to the curve in $\mathbb{R}^m$: $t \mapsto \mathbf{f}(\mathbf{x} + t\mathbf{e}_j)$ at $t = 0$, namely, holding all the variables except for the j-th one as constant, and treating $\mathbf{f}$ as a function of its j-th variable alone. The relations among the n column vectors $\{\dfrac{\partial \mathbf{f}}{\partial x_1}, \ldots, \dfrac{\partial \mathbf{f}}{\partial x_n}\}$ would describe how $\mathbf{f}$ behaves near $\mathbf{x}$.

The significance of $[D\mathbf{f}(\mathbf{x}_0)]$ is that, when it exists, then $\mathbf{L}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)$ is a linear approximation to $\mathbf{f}(\mathbf{x})$ for $\mathbf{x}$ near $\mathbf{x}_0$, in the sense that, if we set $\mathbf{R}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{L}(\mathbf{x})$, then

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\|\mathbf{R}(\mathbf{x})\|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0.$$

A typical situation to use such a linear approximation is when $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$, and for $\mathbf{y}$ near $\mathbf{y}_0$, one would like to solve for $\mathbf{x}$ near $\mathbf{x}_0$ such that $\mathbf{f}(x) = \mathbf{y}$. If we use the linear approximation $\mathbf{f}(\mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)$ to replace $\mathbf{f}(\mathbf{x})$, then this amounts to solving $[D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) = \mathbf{y} - \mathbf{y}_0$, which is a linear equation for $\mathbf{x} - \mathbf{x}_0$ with $[D\mathbf{f}(\mathbf{x}_0)]$ as the coefficient matrix!—This also shows why we need to understand the solvability of a general linear system of the form $A\mathbf{x} = \mathbf{y}$.

Example 4.3.1

Suppose that $\mathbf{x} = (x, y, z) = \mathbf{f}(r, \theta, \phi) = (r\sin\phi\cos\theta, r\sin\phi\sin\theta, r\cos\phi)$ for $(r, \theta, \phi) \in SPC := \{0 \le r < \infty, 0 \le \theta \le 2\pi, 0 \le \phi \le \pi\} \subset \mathbb{R}^3$ — this $\mathbf{f}$ is really the map sending the spherical polar coordinates (r, θ, ϕ) of a point in $\mathbb{R}^3$ to its rectangular coordinates (x, y, z). Each component of $\mathbf{f}$ has continuous partial derivatives with respect to (r, θ, ϕ), so $\mathbf{f}$ is differentiable, and we

$$\nabla f_1 = \begin{bmatrix} \frac{\partial f_1}{\partial r} & \frac{\partial f_1}{\partial \theta} & \frac{\partial f_1}{\partial \phi} \end{bmatrix} = \begin{bmatrix} \sin\phi\cos\theta & -r\sin\phi\sin\theta & r\cos\phi\cos\theta \end{bmatrix}$$
$$\nabla f_2 = \begin{bmatrix} \frac{\partial f_2}{\partial r} & \frac{\partial f_2}{\partial \theta} & \frac{\partial f_2}{\partial \phi} \end{bmatrix} = \begin{bmatrix} \sin\phi\sin\theta & r\sin\phi\cos\theta & r\cos\phi\sin\theta \end{bmatrix}$$
$$\nabla f_3 = \begin{bmatrix} \frac{\partial f_3}{\partial r} & \frac{\partial f_3}{\partial \theta} & \frac{\partial f_3}{\partial \phi} \end{bmatrix} = \begin{bmatrix} \cos\phi & 0 & -r\sin\phi \end{bmatrix}$$

so the Jacobian matrix of $\mathbf{f}$ is

$$\begin{bmatrix} \sin\phi\cos\theta & -r\sin\phi\sin\theta & r\cos\phi\cos\theta \\ \sin\phi\sin\theta & r\sin\phi\cos\theta & r\cos\phi\sin\theta \\ \cos\phi & 0 & -r\sin\phi \end{bmatrix}.$$

Note that the three columns

$$\begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}, \begin{bmatrix} -r\sin\phi\sin\theta \\ r\sin\phi\cos\theta \\ 0 \end{bmatrix}, \begin{bmatrix} r\cos\phi\cos\theta \\ r\cos\phi\sin\theta \\ -r\sin\phi \end{bmatrix},$$

of the Jacobian matrix are $\frac{\partial \mathbf{x}}{\partial r}, \frac{\partial \mathbf{x}}{\partial \theta}, \frac{\partial \mathbf{x}}{\partial \phi}$ respectively. In the case here, they (the three tangent vectors induced by the motions of r, θ, and ϕ respectively) are orthogonal to each other; $\|\frac{\partial \mathbf{x}}{\partial r}\| = 1$, $\|\frac{\partial \mathbf{x}}{\partial \theta}\| = r\sin\phi$, but $\|\frac{\partial \mathbf{x}}{\partial \phi}\| = r$ (Why?—HINT: Think about how $\mathbf{x}$ depends r, θ, and ϕ).

As a consequence, if B is a rectangular box whose edges are along the r, θ, and ϕ axes, respectively, with edge lengths Δr, $\Delta\theta$, and $\Delta\phi$, respectively, then

$D\mathbf{f}(r,\theta,\phi)$ maps B into a box whose three edges are still orthogonal to each other, but with edge lengths equal to $1\cdot\Delta r$, $r\sin\phi\cdot\Delta\theta$, and $r\cdot\Delta\phi$, respectively, so its volume would be $r^2\sin\theta\Delta r\Delta\theta\Delta\phi$—we will meet this relation again when discussing integrals in multi-variables.

Example 4.3.2

Suppose that $\mathbf{f}(\mathbf{x}) = (xyz, x^2+y^2-z^2)$ for $\mathbf{x} = (x,y,z) \in \mathbb{R}^3$. Then

$$[D\mathbf{f}(\mathbf{x})] = \begin{bmatrix} \nabla(xyz) \\ \nabla(x^2+y^2-z^2) \end{bmatrix} = \begin{bmatrix} yz & xz & xy \\ 2x & 2y & -2z \end{bmatrix},$$

so at $\mathbf{x}_0 = (1,-1,-1)$, the linear approximation to $\mathbf{f}(\mathbf{x})$ is

$$\begin{aligned} &\mathbf{f}(\mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0) \\ &= \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & -1 \\ 2 & -2 & 2 \end{bmatrix} \begin{bmatrix} x-1 \\ y+1 \\ z+1 \end{bmatrix} \\ &= \begin{bmatrix} -1+2(x-1)-1(y+1)-1(z+1) \\ 1+2(x-1)-2(y+1)+2(z+1) \end{bmatrix} \\ &= \begin{bmatrix} 2x-y-z-5 \\ 2x-2y+2z-1 \end{bmatrix}. \end{aligned}$$

Theorem 63 provides an easy-to-use criterion to check the differentiability of $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ at $\mathbf{x}_0$ in terms of the continuity of the partial derivatives $\dfrac{\partial f_i(\mathbf{x})}{\partial x_j}$.

4.3.2 The general chain rule

This is used when $\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, with $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0) \in \mathbb{R}^m$, and $\mathbf{g} : \mathbb{R}^m \mapsto \mathbb{R}^l$ is differentiable at $\mathbf{y}_0$, then we can conclude that $\mathbf{g}\circ\mathbf{f}$ is differentiable at $\mathbf{x}_0$, with its derivative $[D(\mathbf{g}\circ\mathbf{f})(\mathbf{x}_0)]$ given by $[D\mathbf{g}(\mathbf{y}_0)][D\mathbf{f}(\mathbf{x}_0)]$.

Note the following two points in the proof of **Theorem 64**

(i). In the middle of p.171, Professor Carlen used the inequality $\|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}\| \le \|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\|_F\|\mathbf{w}\|$. This is simply a short-handed way of saying that, in the case $\mathbf{g} \in \mathbb{R}^2$, so $[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]$ is some $2\times n$

matrix with two rows $\mathbf{r}_1$ and $\mathbf{r}_2$ in $\mathbb{R}^n$,

$$\begin{aligned}
&\| \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \mathbf{w} \| \\
=&\| \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{w} \\ \mathbf{r}_2 \cdot \mathbf{w} \end{bmatrix} \| \\
=&\sqrt{(\mathbf{r}_1 \cdot \mathbf{w})^2 + (\mathbf{r}_2 \cdot \mathbf{w})^2} \\
\leq&\sqrt{\|\mathbf{r}_1\|^2\|\mathbf{w}\|^2 + \|\mathbf{r}_2\|^2\|\mathbf{w}\|^2} \quad \text{using Cauchy-Schwarz inequality} \\
=&\sqrt{\|\mathbf{r}_1\|^2 + \|\mathbf{r}_2\|^2}\|\mathbf{w}\|,
\end{aligned}$$

where $\sqrt{\|\mathbf{r}_1\|^2 + \|\mathbf{r}_2\|^2}\|$ is called the Frobenius norm of the matrix $\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}$.

(ii). It is natural to work out $\mathbf{g} \circ \mathbf{f}(\mathbf{x}) - \mathbf{g} \circ \mathbf{f}(\mathbf{x}_0)$ by using the differentiability of $\mathbf{f}$ at $\mathbf{x}_0$ and of $\mathbf{g}$ at $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$: first,

$$\begin{aligned}
&\mathbf{g} \circ \mathbf{f}(\mathbf{x}) - \mathbf{g} \circ \mathbf{f}(\mathbf{x}_0) \\
=&\mathbf{g}(\mathbf{f}(\mathbf{x})) - \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) \\
=&D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))\,[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)] + z(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0))
\end{aligned}$$

where $\|z(\mathbf{y}, \mathbf{y}_0)\|/\|\mathbf{y} - \mathbf{y}_0\| \to 0$ as $\mathbf{y} \to \mathbf{y}_0$; next

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) = [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) + \mathbf{w}(\mathbf{x}, \mathbf{x}_0),$$

where $\|\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\|/\|\mathbf{x} - \mathbf{x}_0\| \to 0$ as $\mathbf{x} \to \mathbf{x}_0$, so we have

$$\begin{aligned}
\mathbf{g} \circ \mathbf{f}(\mathbf{x}) - \mathbf{g} \circ \mathbf{f}(\mathbf{x}_0) =& [D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))][D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) \\
&+ [D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0) + \mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0)).
\end{aligned}$$

The differentiability of $\mathbf{g} \circ \mathbf{f}(\mathbf{x})$ at $\mathbf{x}_0$ is equivalent to

$$\|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0) + \mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0))\|/\|\mathbf{x} - \mathbf{x}_0\| \to 0 \text{ as } \mathbf{x} \to \mathbf{x}_0.$$

Professor Carlen then treated $\mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0))$ by

$$\frac{\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} = \frac{\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|} \frac{\|\mathbf{y} - \mathbf{y}_0\|}{\|\mathbf{x} - \mathbf{x}_0\|},$$

where $\mathbf{y} = \mathbf{f}(\mathbf{x})$ and $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$, and using

$$\frac{\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|} \to 0 \text{ as } \mathbf{y} \to \mathbf{y}_0, \text{ and}$$

$$\begin{aligned}&\frac{\|\mathbf{y}-\mathbf{y}_0\|}{\|\mathbf{x}-\mathbf{x}_0\|}\\ =&\frac{\|\mathbf{w}(\mathbf{x},\mathbf{x}_0)+[D\mathbf{f}(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)\|}{\|\mathbf{x}-\mathbf{x}_0\|}\\ \leq&\frac{\|\mathbf{w}(\mathbf{x},\mathbf{x}_0)\|}{\|\mathbf{x}-\mathbf{x}_0\|}+\frac{\|[D\mathbf{f}(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)\|}{\|\mathbf{x}-\mathbf{x}_0\|}.\end{aligned}$$

The first term $\to 0$ as $\mathbf{x}\to\mathbf{x}_0$, while the second term is $\leq \|D\mathbf{f}(\mathbf{x}_0)\|_F$ as in (i). Thus $\dfrac{\|\mathbf{y}-\mathbf{y}_0\|}{\|\mathbf{x}-\mathbf{x}_0\|}\leq 1+\|D\mathbf{f}(\mathbf{x}_0)\|_F$ for $\mathbf{x}$ sufficiently close to $\mathbf{x}_0$. Using this and the Squeeze Theorem, we see that $\lim\limits_{\mathbf{x}\to\mathbf{x}_0}\dfrac{\|\mathbf{z}(\mathbf{y},\mathbf{y}_0)\|}{\|\mathbf{x}-\mathbf{x}_0\|}=0$.

But the argument has a small flaw: when $\mathbf{x}\to\mathbf{x}_0$, it may happen that $\mathbf{y}-\mathbf{y}_0=\mathbf{0}$ for certain $\mathbf{x}$, which would void the argument of placing $\|\mathbf{y}-\mathbf{y}_0\|$ in the denominator. This can be handled in the same way that we handled the scalar valued function case on pp.26-27.

When is the Chain Rule most useful? When $\mathbf{z}=\mathbf{g}(\mathbf{y})$ for $\mathbf{y}\in\mathbb{R}^m$, and $\mathbf{y}=\mathbf{f}(\mathbf{x})$ for $\mathbf{x}\in\mathbb{R}^n$ are both differentiable functions, and we need to produce a linear approximation for $\mathbf{g}\circ\mathbf{f}(\mathbf{x})$ at $\mathbf{x}_0$, instead of substituting $\mathbf{y}=\mathbf{f}(\mathbf{x})$ into $\mathbf{z}=\mathbf{g}(\mathbf{y})$ to obtain $\mathbf{z}$ as an explicit function of $\mathbf{x}$ and computing the Jacobian matrix of that function, the chain rule allows us to compute $D\mathbf{g}(\mathbf{x}_0)$ and $D\mathbf{f}(\mathbf{x}_0)$ separately, and multiply these two matrices to obtain the desired Jacobian matrix.

Example 4.3.3

Suppose that $\mathbf{x}=\mathbf{f}(r,\theta,\phi)=(r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)$ for $(r,\theta,\phi)\in SPC:=\{0\leq r<\infty, 0\leq\theta\leq\pi, 0\leq\phi\leq 2\pi\}\subset\mathbb{R}^3$ is the map sending the spherical polar coordinates (r,θ,ϕ) of a point in $\mathbb{R}^3$ to its rectangular coordinates (x,y,z). Suppose that $\mathbf{g}(\mathbf{x})=A\mathbf{x}$, where $\mathbf{x}\in\mathbb{R}^3$ and A is a 3×3 matrix. Then, instead of substituting $\mathbf{x}$ in terms of (r,θ,ϕ) into $\mathbf{g}(\mathbf{x})$ and finding its Jacobian matrix, we use our knowledge that the Jacobian matrix of $\mathbf{f}$ is

$$\begin{bmatrix}\sin\phi\cos\theta & -r\sin\phi\sin\theta & r\cos\phi\cos\theta\\ \sin\phi\sin\theta & r\sin\phi\cos\theta & r\cos\phi\sin\theta\\ \cos\phi & 0 & -r\sin\phi\end{bmatrix}.$$

while the Jacobian matrix of $\mathbf{g}$ is simply A. So the linear approximation to

$\mathbf{g} \circ \mathbf{f} = A\mathbf{f}$ at (r_0, θ_0, ϕ_0) is then given by

$$\mathbf{g} \circ \mathbf{f}(r_0, \theta_0, \phi_0) + [Dg(f(r_0, \theta_0, \phi_0))][Df(r_0, \theta_0, \phi_0)] \begin{bmatrix} r - r_0 \\ \theta - \theta_0 \\ \phi - \phi_0 \end{bmatrix}$$

$$= A \begin{bmatrix} r_0 \sin\theta_0 \cos\phi_0 \\ r_0 \sin\theta_0 \sin\phi_0 \\ r\cos\theta_0 \end{bmatrix} + A \begin{bmatrix} \sin\phi\cos\theta & -r\sin\phi\sin\theta & r\cos\phi\cos\theta \\ \sin\phi\sin\theta & r\sin\phi\cos\theta & r\cos\phi\sin\theta \\ \cos\phi & 0 & -r\sin\phi \end{bmatrix} \begin{bmatrix} r - r_0 \\ \theta - \theta_0 \\ \phi - \phi_0 \end{bmatrix}.$$

Exercise 4.3.1. *Suppose that $f(x, y)$ is differentiable for $(x, y) \in \mathbb{R}^2$. Let (r, θ) be the polar coordinates of (x, y), namely $(x, y) = P(r, \theta) = (r\cos\theta, r\sin\theta)$. Compute the Jacobian matrix of P and verify that*

$$\begin{cases} \dfrac{\partial f}{\partial r} = \cos\theta \dfrac{\partial f}{\partial x} + \sin\theta \dfrac{\partial f}{\partial y}, \\ \dfrac{\partial f}{\partial \theta} = -r\sin\theta \dfrac{\partial f}{\partial x} + r\cos\theta \dfrac{\partial f}{\partial y}, \end{cases}$$

In Matrix form, this is written as

$$\begin{bmatrix} \frac{\partial f}{\partial r} & \frac{\partial f}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

Note that we have abused the notation on the left hand side, as the function on the left hand side really represents the composition $f \circ P$ of f with P.

Exercise 4.3.2. *Suppose that $g(u, v, w) = (u/w, v/w, 2/w - 1)$ and $f(x, y) = (2x, 2y, 1 + x^2 + y^2)$. Determine the Jacobian matrices $[Df(x, y)]$, $[Dg(u, v, w)]$, and $[D(g \circ f)(x, y)]$.*

Chapter 5

The Implicit Function Theorem and Its Consequences

5.1 Horizontal slices and contour curves

Given a function of multi-variables, e.g., $z = f(x, y)$, the usual geometric way to study such a function is to plot its graph over the x-y planar domain. Another approach is to study the geometry of its level curves, namely, for a given scalar $c \in \mathbb{R}$, the set $f_c := \{(x, y) : f(x, y) = c\}$. f_c is also called a contour curve of f; along the contour f_c, f takes on the constant value c.

When f is a linear or quadratic polynomial in x and y, one can see that such a set is often a "smooth curve", but for certain choices of c, the set f_c could have a branch point, or consist of isolated points, or be empty (Play with the case $f(x, y) = x^2 - y^2$ or $f(x, y) = x^2 + y^2$).

When we choose a range of values of c, and plot the sets f_c in the same x-y plane, we obtain a contour plot for the function f.

5.1.1 Implicit and explicit descriptions of planar curves

The easiest descriptions of (planar) curves are given in parametric form or as a graph: $t \in (a, b) \mapsto (\phi(t), \psi(t)) \in \mathbb{R}^2$ for some (differentiable) $\phi(t)$ and $\psi(t)$; or $y = \psi(x)$, or $x = \phi(y)$. Note that a graph is a special case of a parametric curve: $x \mapsto (x, \psi(x)$ or $y \mapsto (\phi(y), y)$.

But often times a planar curve arises from a constrained equation implicitly, such as given by $x^2 + y^2 = 1$ or $x^2 - y^2 = c$. We will answer the following questions in the following discussions.

(i). Given a (continuously differentiable) function $f(x, y)$. What is the criterion for $f_c := \{(x, y) : f(x, y) = c\}$ to be a smooth curve?

(ii). When the answer to the above question is positive, how to compute an equation of the tangent line to the curve f_c at a point (x_0, y_0) on it?

5.1.2 When is the contour curve actually a curve?

Theorem 65 gives an answer to (i) above. In fact, one can give a more specific answer: since it is assumed that $\nabla f(x_0, y_0) \neq (0, 0)$, at least one of

$$f_x(x_0, y_0) := \frac{\partial f}{\partial x}(x_0, y_0), \quad f_y(x_0, y_0) := \frac{\partial f}{\partial y}(x_0, y_0)$$

is not 0; if $f_x(x_0, y_0) \neq 0$, then near (x_0, y_0) the set $f(x, y) = f(x_0, y_0)$ can be given in the form of $x = \psi(y)$, namely,

there exist $r > 0$ and a function $\psi(y)$ defined on some interval (y_1, y_2), with $y_1 < y_0 < y_2$, such that

(a). $f(\psi(y), y) = f(x_0, y_0)$ for all $y \in (y_1, y_2)$,

(b). $\psi(y_0) = x_0$,

(c). $|y - y_0|^2 + |\psi(y) - \psi(y_0)|^2 < r^2$ for all $y \in (y_1, y_2)$, and

(d). if $f(x, y) = f(x_0, y_0)$ and $|x - x_0|^2 + |y - y_0|^2 < r^2$, then $y \in (y_1, y_2)$ and $x = \psi(y)$;

and if $f_y(x_0, y_0) \neq 0$, then a similar statement holds with x and y interchanged.

Here is a heuristic argument for the above statement. $L_f(x, y) := f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$ is the linear approximation to $f(x, y)$ near (x_0, y_0) in the sense that $R(x, y) = f(x, y) - L_f(x, y)$ satisfies

$$\lim_{(x,y)\to(x_0,y_0)} \frac{|R(x, y)|}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0.$$

So solving $f(x, y) = f(x_0, y_0)$ is the same as solving $f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + R(x, y) = 0$, which is approximated by the tangent line $f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) = 0$. When $f_y(x_0, y_0) \neq 0$, this tangent line can be given in the form

of y as a linear function of x, so we expect the solution to $f(x,y) = f(x_0,y_0)$ in such a situation can also be given in the form of $y = \phi(x)$ for some $\phi(x)$; likewise when $f_x(x_0,y_0) \neq 0$.

Here is an answer for (ii) above. $\nabla f(x_0,y_0)$ is orthogonal to the curve $f(x,y) = f(x_0,y_0)$ at (x_0,y_0), thus $\nabla f(x_0,y_0)^\perp := (-f_y(x_0,y_0), f_x(x_0,y_0))$, which is perpendicular to $\nabla f(x_0,y_0)$, is a tangent vector to the curve $f(x,y) = f(x_0,y_0)$ at (x_0,y_0). So an equation for the tangent line can be given either as

$$f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0) = 0 \quad \text{which says } (x-x_0, y-y_0) \perp \nabla f(x_0,y_0);$$

or as in point-slope form when $f_y(x_0,y_0) \neq 0$.

$$y - y_0 = -\frac{f_x(x_0,y_0)}{f_y(x_0,y_0)}(x-x_0).$$

Note that the discussion here also reveals that the tangent is horizontal if and only if $f_x(x_0,y_0) = 0$, and is vertical if and only if $f_y(x_0,y_0) = 0$. This piece of information is useful in identifying or eliminating contour curves of a given function.

Example 5.1.1

In **Exercise 5.2**, it is asked whether either of the curves in the following figure could be a contour plot of $f(x,y) = x^2y + xy - xy^2$. In the first plot, the two branches meet at $(1,1)$. If it is a contour plot of f, then by the Implicit Function Theorem, we must have $\nabla f(1,1) = (0,0)$. We compute $\nabla f(x,y) = (2xy + y - y^2, x^2 + x - 2xy)$, from which we get $\nabla f(1,1) = (2,0)$. So we can conclude that the first plot can't be a contour plot of f. The second plot shows that the tangent line at $(1,1)$ is vertical. This can happen only if $\nabla f(1,1) = (*,0)$, as the tangent line would be of the form $f_x(1,1)(x-1) + f_y(1,1)(y-1) = 0$. Since $\nabla f(1,1) = (2,0)$, this is consistent with the analysis above, and the second plot could be a contour plot of this f. If one is given a figure of contours such as those in **Exercises 5.1** or **5.3**, then one needs to identify the critical points and examine whether their positions are consistent with the given contours (E.g., each closed contour curve of a differentiable function should enclose a critical point, and any point through which there is not a unique differentiable curve passing should be a critical point, as implied by the Implicit Function Theorem.).

Reading Quizzes/Questions:

(i) If f is differentiable at (x_0,y_0), and $\nabla f(x_0,y_0) \neq (0,0)$, which vector $\nabla f(x_0,y_0)$ or $\nabla^\perp f(x_0,y_0)$, is tangent to the level curve $\{(x,y) : f(x,y) = f(x_0,y_0)\}$?

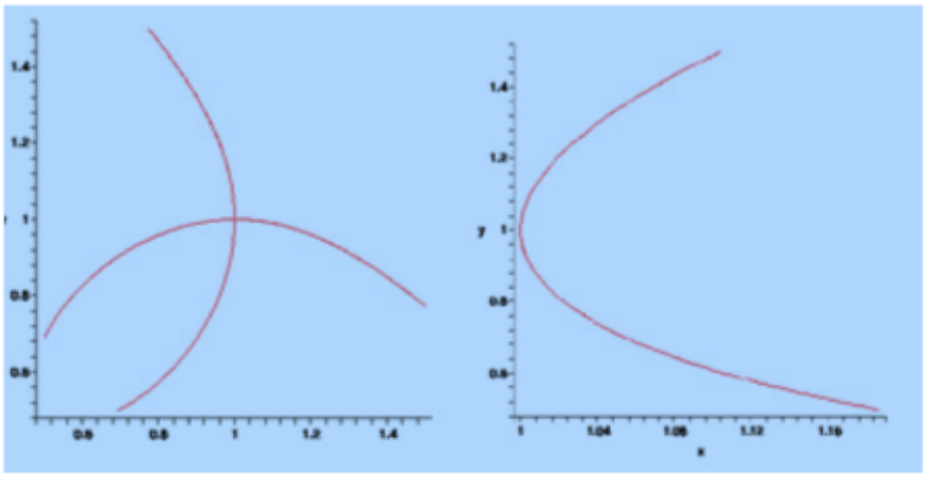


Figure 5.1: Could either plot possibly be a contour curve of $x^2y + xy - y^2x$?

(ii) If the contour curve $\{(x, y) : f(x, y) = f(x_0, y_0)\}$ has a vertical tangent at (x_0, y_0), what can be inferred about $f_x(x_0, y_0)$ or $f_y(x_0, y_0)$?

5.2 Constrained Optimization in Two Variables

This section gives methods to find maximum and minimum values of a continuously differentiable function defined on a closed bounded domain with boundary. The key new ingredient is how to find the maximum and minimum values of a continuously differentiable function defined along the boundary curve, which is often given implicitly in the form of $g(x, y) = c$ for some continuously differentiable g. And that criterion is Lagrange's criterion as given in **Theorem 67**.

5.2.1 Lagrange's criterion for optimizers on the boundary

The key ideas in **Theorem 67** consist of the following

(i). If $f(\mathbf{x})$ attains its maximum or minimum at $\mathbf{x}_0 = (x_0, y_0)$ on $\{\mathbf{x} : g(\mathbf{x}) = 0\}$, then $\nabla f(\mathbf{x}_0) \cdot \mathbf{T}(\mathbf{x}_0) = 0$, where $\mathbf{T}(\mathbf{x}_0)$ is any tangent vector to the curve $\{\mathbf{x} : g(\mathbf{x}) = 0\}$ at $\mathbf{x}_0$. This is because, if $(x(t), y(t))$ is any differentiable curve lying on the level set $\{\mathbf{x} : g(\mathbf{x}) = 0\}$, with $(x(t_0), y(t_0)) = \mathbf{x}_0$, then $(x'(t_0), y'(t_0))$ is a tangent to the level set $\{\mathbf{x} : g(\mathbf{x}) = 0\}$ at $\mathbf{x}_0$, and the one variable function $h(t) := f(x(t), y(t))$ of t attains is maximum or minimum at t_0, so $h'(t_0) = 0$. But $h'(t_0) = \nabla f(\mathbf{x}_0) \cdot (x'(t_0), y'(t_0))$, so it follows that $\nabla f(\mathbf{x}_0) \perp (x'(t_0), y'(t_0))$.

(ii). At any point $\mathbf{x}$ on $\{\mathbf{x} : g(\mathbf{x}) = 0\}$, $\nabla g(\mathbf{x})$ is perpendicular to any tangent at $\mathbf{x}$, as for any differentiable curve $(x(t), y(t))$ lying on the level set $\{\mathbf{x} : g(\mathbf{x}) = 0\}$, $g(x(t), y(t)) \equiv 0$, so taking its derivative in t gives $\nabla g(x(t), y(t)) \cdot (x'(t), y'(t)) \equiv 0$.

Combing (i) and (ii) above, we see that if $f(\mathbf{x})$ attains its maximum or minimum at $\mathbf{x}_0 = (x_0, y_0)$ on $\{\mathbf{x} : g(\mathbf{x}) = 0\}$, then both $\nabla f(\mathbf{x}_0)$ and $\nabla g(\mathbf{x}_0)$ are perpendicular to any tangent to $\{\mathbf{x} : g(\mathbf{x}) = 0\}$ at $\mathbf{x}_0 = (x_0, y_0)$ (one or both could be the zero vector): when $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$, all the tangents to $\{\mathbf{x} : g(\mathbf{x}) = 0\}$ at $\mathbf{x}_0 = (x_0, y_0)$ lie in a line through $\mathbf{x}_0 = (x_0, y_0)$, called the tangent line to $\{\mathbf{x} : g(\mathbf{x}) = 0\}$ at $\mathbf{x}_0 = (x_0, y_0)$, and $\nabla f(\mathbf{x}_0) \parallel \nabla g(\mathbf{x}_0)$. This relation can be expressed as $\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0)$ for some multiplier λ, called Lagrange's multiplier; when $\nabla g(\mathbf{x}_0) = \mathbf{0}$, we may not have $\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0)$, but such an $\mathbf{x}_0$ is still a candidate for a maximum or minimum of f on the level set $\{\mathbf{x} : g(\mathbf{x}) = 0\}$.

This argument generalizes readily to higher dimensions. **Theorem 67** in Carlen's notes formulates the criterion in terms of vanishing of the determinant of the 2×2 matrix whose two rows are the gradient vectors of $f(\mathbf{x})$ and $g(\mathbf{x})$ at $\mathbf{x}_0$ respectively. That criterion would not generalize if we need to find a maximizer/minimizer of a function $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x}) = 0$ when $\mathbf{x} \in \mathbb{R}^n$ and $n \geq 3$.

Example 5.2.1

Find the maximizer and minimizer of $f(x, y) = x^4 + y^4 + 4xy$ on the circle $x^2 + y^2 = 16$.
The constraint $x^2 + y^2 = 16$ is given by $g(x, y) = x^2 + y^2 - 16 = 0$. $\nabla g(x, y) = (2x, 2y)$, which equals $(0, 0)$ only when $(x, y) = (0, 0)$. So when $g(x, y) = 0$, we know $\nabla g(x, y) \neq (0, 0)$.

We also know that the set $\{(x,y) : g(x,y) = 0\}$ is a bounded and closed set in $\mathbb{R}^2$, and that $f(x,y)$ is a continuous function on $\mathbb{R}^2$, so it attains its maximum and minimum values on the set $\{(x,y) : g(x,y) = 0\}$.
If (x,y) attains the maximum or minimum value of $f(x,y)$ on the constraint $\{(x,y) : g(x,y) = 0\}$, then the Lagrange multiplier method implies the existence of some multiplier λ such that $\nabla f(x,y) = \lambda \nabla g(x,y)$ and $g(x,y) = 0$. Written out in detail, we have

$$\begin{cases} 4x^3 + 4y = \lambda(2x) \\ 4y^3 + 4x = \lambda(2y) \\ x^2 + y^2 = 16 \end{cases}$$

This is a system of three nonlinear equations for the three unknowns x, y, λ. There is no good general method for solving such a system; the basic guideline is to try to eliminate some variables to reduce to a single equation for a single variable. Here, one could divide the first two equations to eliminate λ (this process may involve dividing by 0, so one needs to rule out such a possibility) to obtain

$$2y(4x^3 + 4y) = 2x(4y^3 + 4x),$$

from which one obtains $x^3y - xy^3 + y^2 - x^2 = xy(x^2 - y^2) + (y^2 - x^2) = 0$. Thus, either $x^2 - y^2 = 0$, or $xy - 1 = 0$. In the former case, combining with $x^2 + y^2 = 16$, one obtains $x^2 = y^2 = 8$, so $x = \pm\sqrt{8}$ and $y = \pm\sqrt{8}$; conversely, one checks that $(\pm\sqrt{8}, \pm\sqrt{8})$ satisfy the above two equations for (x,y). In the latter case, combining with $x^2 + y^2 = 16$, one obtains $x^2 + x^{-2} = 16$, which is a quadratic equation for $u = x^2$: $u + u^{-1} = 16$ so $u^2 - 16u + 1 = 0$. Its roots are given by $u = (16 \pm \sqrt{16^2 - 4})/2 = 8 \pm \sqrt{63}$, both of which are > 0. We then solve for $x = \pm\sqrt{u} = \pm\sqrt{8 \pm \sqrt{63}}$ and $y = x^{-1}$.
In conclusion, we have found eight candidates for the maximizer/minimizer of f on the constraint $x^2 + y^2 = 16$. What remains is to evaluate f at these eight points, and identify those which give the maximum value and minimum value respectively.
Note that the multiplier λ is not of interest in the final solution; its role is to set up the criterion for a maximizer/minimizer.
It turns out that the maximum of this f on the circle $x^2 + y^2 = 16$ is attained at $(x, 1/x)$ when $x = \pm\sqrt{8 \pm \sqrt{63}}$, with its value equal to 258, while its minimum on this circle is attained at $(-2\sqrt{2}, 2\sqrt{2})$ and $(2\sqrt{2}, -2\sqrt{2})$, with its value equal to 96.

Exercise 5.2.1. *Find the maximum and minimum distance from the origin to points*

on the curve $x^2 + xy + y^2 = 16$.

Example 5.2.2

Use Lagrange multiplier method to find the distance from $\mathbf{x}_0 \in \mathbb{R}^n$ to the (hyper-) plane $\mathbf{n} \cdot \mathbf{x} = d$.
The question is to find the minimum of $f(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}_0||$ subject to $g(\mathbf{x}) = \mathbf{n}\cdot\mathbf{x} - d = 0$. We will choose to work with a different function, $h(\mathbf{x}) = ||\mathbf{x}-\mathbf{x}_0||^2$, as the gradient of $||\mathbf{x}-\mathbf{x}_0||^2$ is easier to work with, and a minimizer for $||\mathbf{x}-\mathbf{x}_0||^2$ is also a minimizer for $||\mathbf{x} - \mathbf{x}_0||$, and vice versa.
We will write out the details for the $n = 3$ case to make the computations more concrete. So we may set up $\mathbf{n} = (a, b, c)$, $\mathbf{x} = (x, y, z)$, and $\mathbf{x}_0 = (x_0, y_0, z_0)$. Then $h(\mathbf{x}) = h(x, y, z) = (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2$, $\nabla h(x, y, z) = 2(x - x_0, y - y_0, z - z_0)$, and $\nabla g(x, y, z) = (a, b, c)$.
Note that the set $\{(x, y, z) : g(x, y, z) = 0\}$ is a closed but unbounded set in $\mathbb{R}^3$, so we can't directly quote Bolzano-Weierstrass Theorem to say that $h(x, y, z)$ attains its minimum on this set. But we note that $h(x, y, z) \to \infty$ as $(x, y, z) \to \infty$. Technically, we use the triangle inequality in the form of $||\mathbf{x} - \mathbf{x}_0|| \geq ||\mathbf{x}|| - ||\mathbf{x}_0||$, so when $||\mathbf{x}|| \geq ||\mathbf{x}_0||$, we have

$$\begin{aligned} h(\mathbf{x}) \geq & (||\mathbf{x}|| - ||\mathbf{x}_0||)^2 \\ = & ||\mathbf{x}||^2 - 2||\mathbf{x}||||\mathbf{x}_0|| + ||\mathbf{x}_0||^2 \\ \geq & ||\mathbf{x}||^2 - \frac{1}{2}||\mathbf{x}||^2 - 2||\mathbf{x}_0||^2 + ||\mathbf{x}_0||^2 \\ \geq & \frac{1}{2}||\mathbf{x}||^2 - ||\mathbf{x}_0||^2 \to \infty \text{ as } ||\mathbf{x}|| \to \infty. \end{aligned}$$

In the above, we used $2||\mathbf{x}||||\mathbf{x}_0|| \leq \frac{1}{2}||\mathbf{x}||^2 + 2||\mathbf{x}_0||^2$, which follows from $2AB \leq A^2 + B^2$ by identifying $A = \frac{1}{\sqrt{2}}||\mathbf{x}||$ and $B = \sqrt{2}||\mathbf{x}_0||$.
More specifically, we only need to look for a minimizer of $h(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$ and $||\mathbf{x}|| \leq R$ for some sufficiently large $R > 0$. At this point we can assert that a minimizer $\mathbf{x}$ exists, and $||\mathbf{x}|| < R$. By the criterion for the Lagrange multiplier, the equations $\nabla h(\mathbf{x}) = \lambda g(\mathbf{x})$ and $g(\mathbf{x}) = 0$ in the $n = 3$ case turn into

$$\begin{cases} 2(x - x_0) = \lambda a, \\ 2(y - y_0) = \lambda b, \\ 2(z - z_0) = \lambda c, \\ ax + by + cz = d. \end{cases}$$

From the first three equations, we obtain $x = x_0 + \lambda a/2, y = y_0 + \lambda b/2, z =$

$z_0 + \lambda c/2$. Substituting these into the last equation, we see that

$$a(x_0 + \lambda a/2) + b(y_0 + \lambda b/2) + c(z_0 + \lambda c/2) = d,$$

from which we obtain $\lambda = -2[ax_0 + by_0 + cz_0 - d]/[a^2 + b^2 + c^2] = -2[\mathbf{n} \cdot \mathbf{x}_0 - d]/||\mathbf{n}||^2$. Therefore

$$\mathbf{x} = \mathbf{x}_0 + \frac{d - \mathbf{n} \cdot \mathbf{x}_0}{||\mathbf{n}||^2}\mathbf{n}.$$

We evaluate h at this $\mathbf{x}$ to obtain

$$||\mathbf{x} - \mathbf{x}_0||^2 = \frac{|d - \mathbf{n} \cdot \mathbf{x}_0|^2}{||\mathbf{n}||^2}.$$

So in the $n = 3$ case the distance from $\mathbf{x}_0 = (x_0, y_0, z_0)$ to the plane $ax+by+cz = d$ is $\frac{|ax_0+by_0+cz_0-d|}{\sqrt{a^2+b^2+c^2}}$, which is consistent with the result we obtained using vector dot product.

Exercise 5.2.2. *Carry out the details of analysis for the above example for the general n cases.*

Exercise 5.2.3. *Find the longest and shortest distances, respectively, from the origin to points on the ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$.*

Example 5.2.3

Prove that, for any $m \times n$ matrix A, $\max ||A\mathbf{x}||$ and $\min ||A\mathbf{x}||$ subject to the constraint $||\mathbf{x}|| = 1$ are attained, and identify the conditions satisfied by a maximizer/minimizer.

The set $\{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}|| = 1\}$ is a closed and bounded set of $\mathbb{R}^n$, and the function $||A\mathbf{x}||$ is a continuous function of $\mathbf{x}$, so it attains its maximum and minimum values on the constraint set $\{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}|| = 1\}$.

To identify the conditions satisfied by a maximizer/minimizer, we work with $f(\mathbf{x}) = ||A\mathbf{x}||^2$ subject to the constraint $g(\mathbf{x}) = ||\mathbf{x}||^2 - 1 = 0$, since $||\mathbf{x}|| = 1$ iff $g(\mathbf{x}) = 0$, and $\mathbf{x}$ is a maximizer/minimizer of $||A\mathbf{x}||$ iff it is a maximizer/minimizer of $||A\mathbf{x}||^2$. The reason we make this choice is that $\nabla||A\mathbf{x}||^2$ and $\nabla||\mathbf{x}||^2$ are easier to work with than $\nabla||A\mathbf{x}||$ and $\nabla||\mathbf{x}||$: using $||A\mathbf{x}||^2 = (A\mathbf{x}) \cdot (A\mathbf{x}) = (A\mathbf{x})^T(A\mathbf{x}) = \mathbf{x}^T A^T A\mathbf{x}$, we see that $f(\mathbf{x}) = \sum_{i,j=1}^n x_i b_{ij} x_j$, where b_{ij} is the (i, j) entry of $B = A^T A$. So $\frac{\partial f(\mathbf{x})}{\partial x_i} = \sum_{j=1}^n b_{ij}x_j + \sum_{k=1}^n b_{ki}x_k$. But

$b_{ki} = b_{ik}$, which follows from $B^T = (A^TA)^T = A^T(A^T)^T = A^TA = B$, so $\frac{\partial f(\mathbf{x})}{\partial x_i} = 2\sum_{j=1}^n b_{ij}x_j$. Similarly, $\frac{\partial g(\mathbf{x})}{\partial x_i} = 2x_i$, so the Lagrange multiplier equations are

$$\begin{cases} 2\sum_{j=1}^n b_{ij}x_j = \lambda(2x_i), i = 1, \ldots, n, \\ \quad ||\mathbf{x}||^2 = 1. \end{cases}$$

The first set of equations can be written in a matrix equation form: $B\mathbf{x} = \lambda\mathbf{x}$. A non-zero vector $\mathbf{x}$ satisfying $B\mathbf{x} = \lambda\mathbf{x}$ is called an eigenvector of the matrix B, and the corresponding λ is called an eigenvalue of B. Note that for a maximizer/minimizer $\mathbf{x}$, we have

$$f(\mathbf{x}) = \mathbf{x}^TB\mathbf{x} = \mathbf{x}^T(\lambda\mathbf{x}) = \lambda\mathbf{x}^T\mathbf{x} = \lambda.$$

So our task has been reduced to finding all eigenvectors and eigenvalues of $B = A^TA$.

Chapter 6

CURVATURE AND QUADRATIC APPROXIMATION

6.1 Quadratic functions

6.1.1 The matrix form of a purely quadratic function

Beyond linear functions, the next simplest functions are purely quadratic functions. A purely quadratic function of two variables has the form $ax^2 + 2bxy + cy^2$ for some coefficients a, b, c, which are not all 0. When a linear part is included, such as in $ax^2 + 2bxy + cy^2 + dx + ey + f$ we call it a quadratic function of (x, y). One of the basic questions that we are interested in is:

> How does such a function behave (whether it has a minimum or a maximum or saddle)? And how is the behavior affected by the coefficients?

Below are the plots of the purely quadratic functions, $4x^2 + 2xy + 16y^2$, $4x^2 + 16xy + 16y^2$, $4x^2 + 32xy + 16y^2$, respectively. One can see that these functions have very distinct behavior.

For a quadratic function of two variables, one can study its behavior by various elementary means, including completion of squares. One natural question is how to study a quadratic function of more than two variables?

The key here is that a purely quadratic function of n variables takes the form $\sum_{i,j=1}^{n} a_{ij}x_ix_j$, and can always be arranged such that

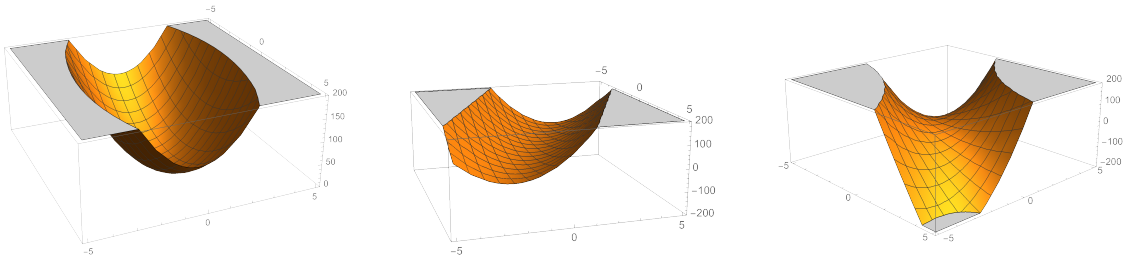


Figure 6.1: Graphs of $4x^2 + 2bxy + 16y^2$ for $b = 1, 8, 16$ respectively

(i). $a_{ij} = a_{ji}$ for all i, j, and

(ii). $\sum_{i,j=1}^{n} a_{ij}x_ix_j = \mathbf{x}\cdot A\mathbf{x} = (\mathbf{x})^T A\mathbf{x}$, where $A = (a_{ij})$ is a symmetric matrix.

Example 6.1.1

The quadratic function $x^2 + 4xy + 3y^2$ can be represented in the form of $\begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix}$ for a 2×2 matrix for many different choices of A, but if we require A to be symmetric, then there is only one such A.

$$x^2+4xy+3y^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 5 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

We are going to develop tools of linear algebra and matrix algebra, and use them to study the behavior of such functions.

Reading Quizzes/Questions:

1. Write the quadratic function $41x^2 - 24xy + 34y^2$ in the form of $\begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix}$, where A is a 2×2 symmetric matrix.

2. Write the quadratic function $21x^2 + 36y^2 - 3z^2 - 84xy - 12yz + 72xz$ in the form of $\begin{bmatrix} x & y & z \end{bmatrix} A \begin{bmatrix} x \\ y \\ z \end{bmatrix}$, where A is a 3×3 symmetric matrix.

6.1.2 Purely quadratic functions as sums of squares

The behavior of a purely quadratic function $\mathbf{x} \cdot A\mathbf{x}$ simplifies tremendously along an eigenvector of A: if $\mathbf{u}$ is such that $A\mathbf{u} = \lambda\mathbf{u}$, then for $\mathbf{x} = s\mathbf{u}$, we have

$$\mathbf{x} \cdot A\mathbf{x} = (s\mathbf{u}) \cdot A(s\mathbf{u}) = (s\mathbf{u}) \cdot (s\lambda\mathbf{u}) = \lambda s^2 \|\mathbf{u}\|^2.$$

If $\mathbf{v}$ is another eigenvector of A: $A\mathbf{v} = \mu\mathbf{v}$ for some μ, then for $\mathbf{x} = s\mathbf{u} + t\mathbf{v}$,

$$\mathbf{x} \cdot A\mathbf{x} = (s\mathbf{u} + t\mathbf{v}) \cdot A(s\mathbf{u} + t\mathbf{v}) = s^2\mathbf{u} \cdot A\mathbf{u} + st\,(\mathbf{u} \cdot A\mathbf{v} + \mathbf{v} \cdot A\mathbf{u}) + t^2\mathbf{v} \cdot A\mathbf{v}.$$

If we are in a situation such that

$$\mathbf{u} \cdot A\mathbf{v} = \mathbf{u} \cdot (\mu\mathbf{v}) = 0 \quad \text{ and } \mathbf{v} \cdot A\mathbf{u} = \mathbf{v} \cdot (\lambda\mathbf{u}) = 0,$$

then

$$\mathbf{x} \cdot A\mathbf{x} = (s\mathbf{u} + t\mathbf{v}) \cdot A(s\mathbf{u} + t\mathbf{v}) = \lambda s^2 \|\mathbf{u}\|^2 + \mu t^2 \|\mathbf{v}\|^2,$$

which is, up to the scalars λ and μ, an algebraic sum of squares. We will show that when A is a symmetric matrix, $A\mathbf{u} = \lambda\mathbf{u}$, $A\mathbf{v} = \mu\mathbf{v}$ for some $\lambda \neq \mu$, then indeed we have $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u} = 0$.

Professor Carlen's notes use the function $f(x, y) = x^2 - xy + y^2$ to illustrate how to represent it as a sum of squares. But we can also use completion of squares to write it as $f(x, y) = (x - \frac{1}{2}y)^2 + \frac{3}{4}y^2$. If we set $\xi = x - \frac{1}{2}y$ and $\eta = y$, then $f = \xi^2 + \frac{3}{4}\eta^2$. What's the difference between this change of variables and the one in Professor Carlen's notes? Under this change of variables, the x-axis, represented by the equation $y = 0$, corresponds to $\eta = 0$, which is the ξ-axis; but the y-axis, represented by the equation $x = 0$, corresponds to $\xi = -\frac{1}{2}\eta$, which is a line not orthogonal to the ξ-axis. Conversely, the η-axis, represented by $\xi = 0$, corresponds to $x - \frac{1}{2}y = 0$. Note, however, that the change of variables $u = (x + y)/\sqrt{2}$, $v = (x - y)/\sqrt{2}$, preserves angels and lengths.

One major advantage of linear change of variables which preserve angles and lengths is that the geometry of the level curves can be easily related from the u-v coordinates to the original x-y coordinates. For instance, with the u-v coordinate above, $f = \frac{1}{2}u^2 + \frac{3}{2}v^2$, so the level curve $f = c$, for $c > 0$, given by $\frac{1}{2}u^2 + \frac{3}{2}v^2 = c$ is an ellipse whose major axis is along the u-axis. The geometry of this ellipse in the x-y coordinates, other than its orientation, looks the same. On the other hand, if we used the coordinates ξ-η, then the ellipse $\xi^2 + \frac{3}{4}\eta^2 = c$ would have its major axis along the η-axis, and its shape in the coordinates ξ-η would look different from its shape in the x-y axis.

Thus we prefer to work with linear change of variables of the form $\mathbf{x} = Q\mathbf{u}$ such that:

(a). It has an inverse in a similar form $\mathbf{u} = P\mathbf{x}$ for some $n \times n$ matrix in the sense that

$$\mathbf{x} = QP\mathbf{x} \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \text{ and } \mathbf{u} = PQ\mathbf{u} \quad \text{for all } \mathbf{u} \in \mathbb{R}^n. \tag{6.1}$$

(b). The transformation $\mathbf{u} \mapsto \mathbf{x} = Q\mathbf{u}$ preserves angels and lengths so that the u_i-axes are mapped to axes which are still orthogonal to each other.

(6.1) implies that $e_j = QPe_j$ for each j, but QPe_j is the jth column of QP. Thus $QP = [e_1\, e_2 \cdots e_n] = I_n$. Likewise, $PQ = I_n$. Thus the requirement of (a) is equivalent to the existence of matrix P such that

$$QP = PQ = I_n. \tag{6.2}$$

In other words, the requirement of (a) is equivalent to the invertibility of the matrix Q.

(b) requires that Qe_j is still a unit vector and $Qe_j \perp Qe_k$ for $j \neq k$. But Qe_j is the jth column of Q. Thus the requirement of (b) implies that the columns of Q form an orthonormal set of vectors. This condition can be expressed in a compact matrix form: $Q^TQ = I_n$. In this context Q is an $n \times n$ square matrix. An $n \times n$ matrix Q satisfying $Q^TQ = I_n$ is called an orthogonal matrix*; it turns out that in such a case we automatically have $QQ^T = I_n$. Also, the condition $Q^TQ = I_n$ implies that $\mathbf{x} = Q\mathbf{u}$ preserves the dot product between two vectors in $\mathbb{R}^n$, therefore preserves the length of vectors and angle between vectors. This is seen by noting that $\mathbf{u}\cdot\mathbf{v} = (\mathbf{u})^T\mathbf{v}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, and treating them as column vectors, so it follows that

$$Q\mathbf{u} \cdot Q\mathbf{v} = (Q\mathbf{u})^TQ\mathbf{v} = (\mathbf{u})^TQ^TQ\mathbf{v} = (\mathbf{u})^TI_n\mathbf{v} = (\mathbf{u})^T\mathbf{v} = \mathbf{u}\cdot\mathbf{v}.$$

Reading Quizzes/Questions:

(i) Is the matrix $\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$ an orthogonal matrix?

(ii) If Q is an $n \times r$ such that $Q^TQ = I_r$, and $n \neq r$, do we still have $QQ^T = I_n$? Can $r > n$ under the condition here? Why?

*Note that the columns of an orthogonal matrix are orthonormal, not just orthogonal to each other. If Q is an $n \times r$ matrix whose columns form an orthonormal set of vectors, but $r \neq n$ (which necessarily implies that $r < n$), then we will still have $Q^TQ = I_r$. But such a Q is not called an orthogonal matrix, and $QQ^T \neq I_n$ in such a case.

(iii) Construct a 3×2 matrix Q whose columns consist of orthonormal vectors. Then compute QQ^T.

(iv) If Q is an orthogonal matrix, does it follow that its rows are also orthonormal?

The remaining question is:

Given a purely quadratic function of the form $\mathbf{x}\cdot A\mathbf{x}$, where A is a symmetric matrix. Is there an orthogonal matrix Q such that after a change of variables $\mathbf{x}=Q\mathbf{u}$, the function $\mathbf{x}\cdot A\mathbf{x}$ in terms of $\mathbf{u}$ becomes a purely quadratic function of the simplest form: $\lambda_1 u_1^2+\lambda_2 u_2^2+\ldots+\lambda_n u_n^2$ for some scalars $\lambda_1,\ldots\lambda_n$?

Reading Quizzes/Questions: For the quadratic function defined by

$$\begin{bmatrix} x & y\end{bmatrix}\begin{bmatrix} 41 & -12\\ -12 & 34\end{bmatrix}\begin{bmatrix} x\\ y\end{bmatrix},$$

perform the change of variable

$$\begin{bmatrix} x\\ y\end{bmatrix}=\begin{bmatrix} \frac{3}{5} & -\frac{4}{5}\\ \frac{4}{5} & \frac{3}{5}\end{bmatrix}\begin{bmatrix} u\\ v\end{bmatrix}$$

and find the function in terms of u,v.

If Q is an orthogonal $n\times n$ matrix, using the dot product preserving property $Q\mathbf{x}\cdot Q\mathbf{y}=\mathbf{x}\cdot\mathbf{y}$ for any vectors $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, we note the following

(i). $\mathbf{x}\cdot A\mathbf{x}=Q\mathbf{u}\cdot AQ\mathbf{u}=(Q\mathbf{u})^TAQ\mathbf{u}=(\mathbf{u})^TQ^TAQ\mathbf{u}=\mathbf{u}\cdot(Q^TAQ)\mathbf{u}$.

(ii). $\lambda_1 u_1^2+\lambda_2 u_2^2+\ldots+\lambda_n u_n^2=\mathbf{u}\cdot D\mathbf{u}$, where D is the diagonal matrix whose ith diagonal entry is λ_i.

Thus our requirement is $\mathbf{u}\cdot(Q^TAQ)\mathbf{u}=\mathbf{u}\cdot D\mathbf{u}$ for all $\mathbf{u}$, which is equivalent to $Q^TAQ=D$.

Since $Q^TQ=QQ^T=I_n$, the condition $Q^TAQ=D$ is also equivalent to $AQ=QD$. But D is a diagonal matrix, so the jth column of QD is simply λ_j times the jth column of Q, while the jth column of AQ is A times the jth column of Q. Thus the equation $AQ=QD$ is saying that $A\text{Col}_j(Q)=\lambda_j\text{Col}_j(Q)$, for $j=1,\ldots,n$, which

is the condition that $\text{Col}_j(Q)$ is an eigenvector of the matrix A, with λ_j being the corresponding eigenvalue.

Reading Quizzes/Questions: Verify that $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}2\\-1\end{bmatrix}$ are eigenvectors of the matrix

$$\begin{bmatrix}2 & 2\\2 & 5\end{bmatrix}.$$

Identify the corresponding eigenvalues.

Recall the additional requirement that the columns of Q need to form an orthonormal set of vectors, thus the question we face is

> Given a symmetric matrix A, whether we can find a set of n eigenvectors of A such that they form an orthonormal set of vectors?

This is answered affirmatively in **6.1.3-4** by the Spectral Theorem, also called diagonalization of real symmetric matrices.

The key ingredients and steps are

(i). An eigenvalue λ of A is characterized by the condition that $(A-\lambda I_n)\mathbf{x} = \mathbf{0}$ has non-zero solutions; i.e., $A-\lambda I_n$ does not have an inverse. An algebraic condition is that the determinant of $A - \lambda I_n$ must be 0.

(ii). For each λ such that the determinant of $A - \lambda I_n$ is 0, find the general solutions to $(A - \lambda I_n)\mathbf{x} = \mathbf{0}$. It will be spanned by a set of solution vectors. Apply the Gram-Schmidt Algorithm to produce a set of orthonormal spanning set for $\text{Null}(A - \lambda I_n)$.

(iii). If $\lambda_1 \neq \lambda_2$ are two distinct eigenvalues of A, with $\mathbf{x}_1$ and $\mathbf{x}_2$ being corresponding eigenvectors, then $\mathbf{x}_1 \perp \mathbf{x}_2$.

(iv). If we collect all the eigenvectors constructed above, it forms an orthonormal basis of $\mathbb{R}^n$: $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$, with $A\mathbf{q}_j = \lambda_j \mathbf{q}_j$, $1 \leq j \leq n$. Setting $Q = [\mathbf{q}_1 \ldots \mathbf{q}_n]$, then $Q^TQ = I_n$, $QQ^T = I_n$, and

$$AQ = Q\begin{bmatrix}\lambda_1 & 0 & \ldots & 0\\0 & \lambda_2 & \ldots & 0\\\vdots & \vdots & \vdots & \vdots\\0 & 0 & \ldots & \lambda_n\end{bmatrix} = QD; \qquad \text{thus } Q^{-1}AQ = Q^TAQ = D.$$

In implementing the above algorithm, the first step is to identify all eigenvalues of the matrix A, and for each eigenvalue λ of A to find an orthonormal basis for the space of solutions of $(A - \lambda I)\mathbf{x} = \mathbf{0}$. The second step is usually accomplished by first finding a basis for the space of solutions of $(A - \lambda I)\mathbf{x} = \mathbf{0}$, then applying the Gram-Schmidt Algorithm to obtain an orthonormal basis.

For a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, its eigenvalues are roots of the quadratic polynomial of $\det(A - tI) = t^2 - (a + d)t + ad - bc = 0$. For a 3×3 matrix A, $\det(A - tI)$ is a cubic polynomial in t. Theoretically there is a formula for the roots of a cubic polynomial, but it is not used in practice, as it is algebraically too complicated.

Another potential complication in practice is that if one uses an approximate value for an eigenvalue, then $(A-\lambda I)$ becomes invertible, so theoretically, the only solution of $(A - \lambda I)\mathbf{x} = \mathbf{0}$ for an approximate eigenvalue λ is the zero vector $\mathbf{0}$. Such issues are analyzed in numerical linear algebra.

It is a simpler task to verify whether a given scalar λ is an eigenvalue of a matrix A: one simply checks whether one can find non-zero solutions of $(A - \lambda I)\mathbf{x} = \mathbf{0}$. It is even simpler to check whether a given vector $\mathbf{x}$ is an eigenvector of a matrix: one simply checks whether $A\mathbf{x}$ is a multiple of $\mathbf{x}$.

Example 6.1.2

For

$$A = \begin{bmatrix} 41 & -12 \\ -12 & 34 \end{bmatrix},$$

one can verify that

$$A \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 41 & -12 \\ -12 & 34 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 75 \\ 100 \end{bmatrix} = 25 \begin{bmatrix} 3 \\ 4 \end{bmatrix},$$

so $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ is an eigenvector corresponding to eigenvalue $\lambda = 25$.
To find the full set of eigenvalues, one solves

$$\det(A - tI) = \det \begin{bmatrix} 41 - t & -12 \\ -12 & 34 - t \end{bmatrix} = t^2 - 75t + 1250 = 0$$

to find $t = 25$ and $t = 50$.
Next we find all solution of

$$(A - 50I)\mathbf{x} = \begin{bmatrix} -9 & -12 \\ -12 & -16 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which are given by solving $-3x - 4y = 0$. Thus all solutions are multiples of $\begin{bmatrix} 4 \\ -3 \end{bmatrix}$.

Note that the eigenvectors $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ -3 \end{bmatrix}$ are orthogonal to each other. To produce an orthogonal matrix we need to make each a unit vector. Thus the two columns of the following matrix are orthonormal eigenvectors of the given matrix A

$$Q = \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix},$$

and the eigenvector relations are encoded in

$$AQ = Q \begin{bmatrix} 25 & 0 \\ 0 & 50 \end{bmatrix},$$

or equivalently, $Q^T AQ = \begin{bmatrix} 25 & 0 \\ 0 & 50 \end{bmatrix}$.

Note that in solving for $(A-50I)\mathbf{x} = \mathbf{0}$, we could have said that all solutions are multiples of $\begin{bmatrix} -4 \\ 3 \end{bmatrix}$. This would have resulted in a different orthogonal matrix

$$\begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}.$$

This matrix has determinant equal to 1, and represents a rotation matrix, while the previous choice has its determinant equal to -1 and represents a reflection matrix.

Reading Quizzes/Questions: For the matrix

$$A = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix},$$

find an orthogonal matrix Q such that $AQ = QD$ for some diagonal matrix. Identify D.

Reading Quizzes/Questions: Find the eigenvalues of the matrices

$$Q_1 = \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix} \text{ and } Q_2 = \begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}.$$

Remark 6.1.1

Our discussion shows that as long as an $n \times n$ matrix A has a set of n eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ which is linearly independent, namely, the $n \times n$ matrix V with $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ as its columns has rank n, and with $\{\lambda_1, \ldots, \lambda_n\}$ as its corresponding eigennvalues, then $AV = VD$, where D is the diagonal matrix with $\{\lambda_1, \ldots, \lambda_n\}$ as entries on its diagonal. Since V is invertible, it follows that $V^{-1}AV = D$, or equivalently, $A = VDV^{-1}$. We see that in such a situation, A is diagonalizable.
This kind of diagonalization has important application. For instance, it would allow us to write $A = VDV^{-1}$, so $A^k = VD^kV^{-1}$ can be computed easily. Another application is in solving a system of ODEs given by $\mathbf{x}'(t) = A\mathbf{x}(t)$. Making a change of variables $\mathbf{x} = V\mathbf{y}$, then the system reduces to $V\mathbf{y}'(t) = AV\mathbf{y}(t)$, so $\mathbf{y}'(t) = D\mathbf{y}(t)$, which is completely decoupled, and can be solved easily.

Example 6.1.3

The matrix $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ as 1 as its only eigenvalue, and when solving for the relevant eigenvectors, we find that they are given by

$$\begin{bmatrix} 1-1 & 2 \\ 0 & 1-1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which is just $2y = 0$. Thus the solutions are $\begin{bmatrix} x \\ 0 \end{bmatrix}$. This means that there is no way to find a set of two orthonormal eigenvectors.
In fact one can see that this matrix is not diagonalizable, for, it were to possible to write it as VDV^{-1} for some invertible V and a diagonal D, then the entries on the diagonal of D must be eigenvalues, therefore must be both 1's, which would force D to be I_2, which would then force $VDV^{-1} = VI_2V^{-1} = VV^{-1} = I_2$, which is not the case.

Reading Quizzes/Questions: Verify that the matrix $\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ -3 & 2 \end{bmatrix}$ has $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 6 \end{bmatrix}$ as eigenvectors, and use this information to determine A^{10} and $\lim_{n\to\infty} A^n$. Furthermore,

find the general solution to

$$\begin{cases} x_1'(t) = -\frac{1}{2}x_1(t) + \frac{1}{2}x_2(t), \\ x_2'(t) = -3x_1(t) + 2x_2(t). \end{cases}$$

Example 6.1.4

Consider the quadratic function $x^2 - xy + y^2$, which can be written as $[x\ y]A[x\ y]^T$, where $A = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$. We apply our algorithm above.

(i). An eigenvalue λ of A makes $A - \lambda I_2 = \begin{bmatrix} 1-\lambda & -\frac{1}{2} \\ -\frac{1}{2} & 1-\lambda \end{bmatrix}$ to have no inverse. This is equivalent to $\begin{bmatrix} 1-\lambda \\ -\frac{1}{2} \end{bmatrix} \parallel \begin{bmatrix} -\frac{1}{2} \\ 1-\lambda \end{bmatrix}$. But this is equivalent to $(1-\lambda)^2 - (-\frac{1}{2})^2 = 0$. This gives $\lambda = \frac{1}{2}$ or $\lambda = \frac{3}{2}$.

(ii). For $\lambda = \frac{1}{2}$, we need to find all solutions to

$$(A - \frac{1}{2}I_2)\mathbf{x} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The solutions $\mathbf{x} = (x\ \ y)^T$ are given by $x - y = 0$. Choosing y as a free variable, we see that the solutions are

$$\begin{bmatrix} y \\ y \end{bmatrix} = y \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We choose a normalized vector, which in this case is $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$.

Then for $\lambda = \frac{3}{2}$, we solve

$$(A - \frac{3}{2}I_2)\mathbf{x} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The solutions $\mathbf{x} = (x\ \ y)^T$ are given by $x + y = 0$. Choosing y as a free variable, we see that the solutions are

$$\begin{bmatrix} -y \\ y \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

We choose a normalized vector, which in this case is $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$.

(iii). We note that $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ and $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ are orthonormal. If we set

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},$$

then $Q^TQ = QQ^T = I_2$, and $AQ = Q \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}$. If we make the change of variables

$$\begin{bmatrix} x \\ y \end{bmatrix} = Q \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix},$$

then $\begin{bmatrix} x & y \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} Q^T$, and

$$\begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} Q^T A Q \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{1}{2}(x')^2 + \frac{3}{2}(y')^2.$$

So under this change of variables, $x^2 - xy + y^2 = \frac{1}{2}(x')^2 + \frac{3}{2}(y')^2$.
For any $c > 0$, the level curve $x^2 - xy + y^2 = c$ becomes $\frac{1}{2}(x')^2 + \frac{3}{2}(y')^2 = c$ in the x'-y' coordinates. Since the relation between (x, y) and (x', y') coordinates here are of a rotation ($(x', y') = (1, 0)$ gets mapped to $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$, which means that the x'-axis gets mapped to the line $y = x$; and $(x', y') = (0, 1)$ gets mapped to $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$, which means that the x'-axis gets mapped to the line $y = -x$). This change of variable allows us to see (a). the level curves of $x^2 - xy + y^2$ are ellipses, and (b). The function $x^2 - xy + y^2 = \frac{1}{2}(x')^2 + \frac{3}{2}(y')^2$ is non-negative, and can grow unbounded.

Remark 6.1.2

In step (ii) above for computing the eigenvectors associated with $\lambda = \frac{3}{2}$, if we had chosen x as a free variable, we would have gotten the eigenvectors as

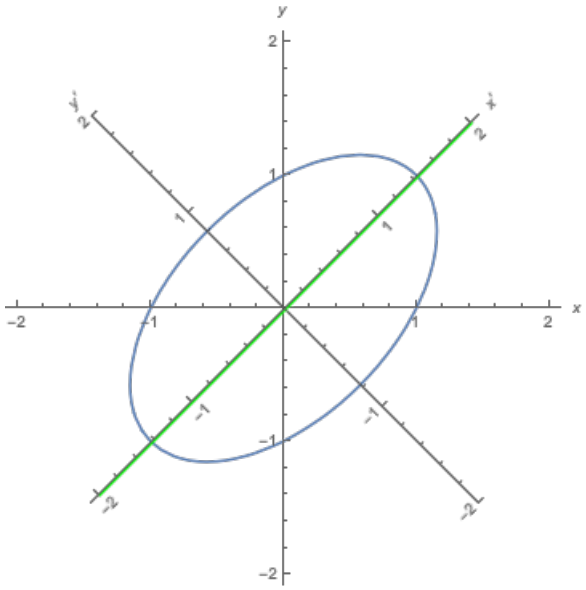


Figure 6.2: Contour plot of $x^2 - xy + y^2 = 1$ in the original x-y axes and in the rotated x'-y' axes.

$x \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, *and with* $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$ *as its normalized eigenvector;*

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix}$$

would still have been an legitimate orthogonal change of variables, but this time, the y'-axis would have been mapped to be along the direction of the vector $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$, *making the x'-y' coordinate system left handed. This change of variables is through a reflection, instead of a rotation.*
Algebraically a rotation is represented by an orthogonal matrix with a positive determinant, while a reflection is represented by an orthogonal matrix with a negative determinant.

Reading Quizzes/Questions:

(i) For $2x^2 + 4xy + 5y^2$, perform a change of variables of the form

$$\begin{bmatrix} x \\ y \end{bmatrix} = Q \begin{bmatrix} u \\ v \end{bmatrix}$$

where Q is some orthogonal matrix, so that $2x^2 + 4xy + 5y^2$ transforms into $\lambda_1 u^2 + \lambda_2 v^2$. Identify Q, λ_1 and λ_2.

(ii) Given that $9, -36, 81$ are eigenvalues of the matrix

$$A = \begin{bmatrix} 21 & -42 & 36 \\ -42 & 36 & -6 \\ 36 & -6 & -3 \end{bmatrix}.$$

Find their corresponding eigenvectors. Then construct an orthonormal matrix Q such that $Q^T AQ$ is a diagonal matrix.

(iii) Use the result of the previous problem to perform a change of variables of the form

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = Q \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

to transform the quadratic function $21x^2 + 36y^2 - 3z^2 - 84xy - 12yz + 72xz$ into a quadratic function in u, v, w. Then describe the geometry of the level set $21x^2 + 36y^2 - 3z^2 - 84xy - 12yz + 72xz = 81$.

The behavior of quadratic functions will be used in the second derivative test for mimima/maxima at a critical point.

6.2 The best quadratic approximation

6.2.1 Higher order directional derivatives and repeated partial differentiation

The key of this subsection is (6.1.2-4). The idea is to examine the behavior of the function $f(\mathbf{x})$ along the one-dimensional line $\mathbf{x}_0 + t\mathbf{v}$: $g(t) = f(\mathbf{x}_0 + t\mathbf{v})$; specifically, computing its first order derivative in t gives $g'(t) = \sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + t\mathbf{v})$, which, at $t = 0$, gives the directional derivative of f at $\mathbf{x}_0$ in the direction of $\mathbf{v}$. Assume that

each $\frac{\partial f}{\partial x_i}(\mathbf{x})$ is differentiable, then we can apply the same chain rule to compute the second order derivative of $g(t)$ in t to get (6.12) in Carlen's notes:

$$g''(t) = \sum_{i,j=1}^{n} v_i v_j \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + t\mathbf{v}),$$

where

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + t\mathbf{v}) = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)(\mathbf{x}_0 + t\mathbf{v})$$

is the partial derivative of $\frac{\partial f}{\partial x_i}(\mathbf{x})$ with respect to x_j evaluated at $\mathbf{x}_0 + t\mathbf{v}$. Thus

$$g''(0) = \sum_{i,j=1}^{n} v_i v_j \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0),$$

which is a purely quadratic function in the variable $(v_1, v_2, \ldots, v_n)$, and gives the second order derivative of f at $\mathbf{x}_0$ in the direction of $\mathbf{v}$.

Example 6.2.1

For $f(x, y) = x^y$ for $x, y > 0$, we found from an earlier example that $\partial_x f(x, y) = yx^{y-1}$ and $\partial_y f(x, y) = x^y \ln x$. Therefore

$$\begin{aligned}
\frac{\partial^2 f}{\partial x \partial x} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) &&= \frac{\partial}{\partial x}\left(yx^{y-1}\right) = y(y-1)x^{y-2} \\
\frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) &&= \frac{\partial}{\partial y}\left(yx^{y-1}\right) = x^{y-1} + yx^{y-1}\ln x \\
\frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) &&= \frac{\partial}{\partial x}\left(x^y \ln x\right) = yx^{y-1}\ln x + x^{y-1} \\
\frac{\partial^2 f}{\partial y \partial y} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) &&= \frac{\partial}{\partial y}\left(x^y \ln x\right) = x^y (\ln x)^2
\end{aligned}$$

Then for any given $\mathbf{v} = (v_1, v_2)$, and $g(t) = (1 + tv_1)^{1+tv_2}$ would have

$$\begin{aligned}
g''(0) &= \frac{\partial^2 f}{\partial x \partial x}(1,1)v_1^2 + \frac{\partial^2 f}{\partial y \partial x}(1,1)v_1 v_2 + \frac{\partial^2 f}{\partial x \partial y}(1,1)v_1 v_2 + \frac{\partial^2 f}{\partial y \partial y}(1,1)v_2^2 \\
&= 0v_1^2 + 2v_1 v_2 + 0v_2^2.
\end{aligned}$$

6.2.2 Clairault's Theorem

In (6.1.2-4) we encounter a quadratic function in $\mathbf{v}$ whose coefficients are $\frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x}_0) := \frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} f\right)(\mathbf{x}_0)$*. **Clairault's Theorem** gives conditions under which these coefficients are symmetrical in i and j. Under the conditions here, the function $\sum_{i,j=1}^{n} v_i v_j \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$ can be written in a compact matrix form $\mathbf{v} \cdot [\text{Hess}_f(\mathbf{x}_0)]\mathbf{v} = (\mathbf{v})^T [\text{Hess}_f(\mathbf{x}_0)]\mathbf{v}$, where the entries of the matrix $[\text{Hess}_f(\mathbf{x}_0)]$, called the Hessian matrix of f at $\mathbf{x}_0$, are $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$, so $[\text{Hess}_f(\mathbf{x}_0)]$ would be a symmetric matrix.

In the above, we used the notation $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$, instead of $\frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} f(\mathbf{x}_0)\right)$, for the second derivative of f at $\mathbf{x}_0$. The choice of notation reflects the internal logic of the operation, as $\frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} f(\mathbf{x}_0)\right)$ may suggest that we are taking $\frac{\partial}{\partial x_j}$ partial derivative of $\frac{\partial}{\partial x_i} f(\mathbf{x}_0)$, which is already evaluated as a constant. Such kind of logic is important in the syntax of programming languages such as `Mathematica`. E.g., in `Mathematica`, for a generic function $f[x]$, `D[f[x],x]` would give the derivative of f at x, while `D[f[x_0],x]` would treat $f[x_0]$ as a constant, so give 0 as the output, not the derivative of f at x_0. The derivative of f at x_0 is obtained as `D[f[x],x]/.x->x_0`. Our notation $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$ treats $\frac{\partial^2 f}{\partial x_j \partial x_i}$ as the second derivative of f as a function, ready to be evaluated at any $\mathbf{x}$.

It is possible for $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \neq \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$ when the conditions for the Clairault's Theorem are not satisfied.

Example 6.2.2

Consider

$$f(x,y) = \begin{cases} xy\frac{x^2-y^2}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases}$$

*Some texts use the notation $f_{x_i}(\mathbf{x})$ for $\frac{\partial}{\partial x_i} f(\mathbf{x})$, and $f_{x_i x_j}(\mathbf{x})$ for $\frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} f\right)(\mathbf{x})$.

Then at $(x, y) \neq (0, 0)$,

$$\begin{aligned}
\frac{\partial f}{\partial x}(x, y) &= \frac{y\left(4x^2y^2 + x^4 - y^4\right)}{\left(x^2 + y^2\right)^2}, \\
\frac{\partial f}{\partial y}(x, y) &= \frac{-4x^3y^2 + x^5 - xy^4}{\left(x^2 + y^2\right)^2}, \\
\frac{\partial^2 f}{\partial y \partial x}(x, y) &= \frac{\left(x^2 - y^2\right)\left(10x^2y^2 + x^4 + y^4\right)}{\left(x^2 + y^2\right)^3} \\
\frac{\partial^2 f}{\partial x \partial y}(x, y) &= \frac{\left(x^2 - y^2\right)\left(10x^2y^2 + x^4 + y^4\right)}{\left(x^2 + y^2\right)^3}
\end{aligned}$$

so we see that

$$\frac{\partial^2 f}{\partial y \partial x}(x, y) = \frac{\partial^2 f}{\partial x \partial y}(x, y) \quad \text{when } (x, y) \neq (0, 0).$$

We can also verify directly by definition that

$$\frac{\partial f}{\partial x}(0, 0) = 0, \quad \frac{\partial f}{\partial y}(0, 0) = 0.$$

To compute $\frac{\partial^2 f}{\partial y \partial x}(0, 0)$, we only need to examine the derivative with respect to y of $\frac{\partial f}{\partial x}(0, y) = -y$, which gives -1; while to compute $\frac{\partial^2 f}{\partial x \partial y}(0, 0)$, we only need to examine the derivative with respect to x of $\frac{\partial f}{\partial y}(x, 0) = x$, which gives 1. Thus

$$\frac{\partial^2 f}{\partial y \partial x}(0, 0) = -1 \neq 1 = \frac{\partial^2 f}{\partial x \partial y}(0, 0).$$

6.2.3 A multivariable second order Taylor expansion

Treating $g(t) = f(\mathbf{x}_0 + t\mathbf{v})$ as a one variable function of t, and if f has continuous second derivatives with respect to $\mathbf{x}$, then $g(t)$ has continuous second derivatives in t, and we have a Taylor's expansion of the form $g(t) = g(0) + g'(0)t + \frac{g''(0)}{2}t^2 + R_2(t)$, where the remainder

$$R_2(t) = \int_0^t (t-s)[g''(s) - g''(0)]ds = t^2 \int_0^1 (1-\tau)[g''(t\tau) - g''(0)]d\tau \text{ (after setting } s = t\tau).$$

Setting $\mathbf{w} = t\mathbf{v}$, we get

$$f(\mathbf{x}_0+\mathbf{w}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\cdot\mathbf{w} + \frac{1}{2}\mathbf{w}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{w} + R_2,$$

where $R_2 = \mathbf{w}\cdot(\int_0^1 A(\tau)d\tau)\mathbf{w}$, $A(\tau) = [\mathrm{Hess}_f(\mathbf{x}_0+\tau\mathbf{w})] - [\mathrm{Hess}_f(\mathbf{x}_0)]$. Here

$$f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\cdot\mathbf{w} + \frac{1}{2}\mathbf{w}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{w}$$

provides the "best quadratic approximation" to $f(\mathbf{x})$ for $\mathbf{x} = \mathbf{x}_0+\mathbf{w}$ near $\mathbf{x}_0$.

If the second derivatives of f are continuous at $\mathbf{x}_0$, then we will be able to prove that

$$\lim_{\|\mathbf{w}\|\to 0}\frac{|R_2|}{\|\mathbf{w}\|^2} = 0. \tag{6.3}$$

This means that $|R_2|$ is vanishingly small compared with $\|\mathbf{w}\|^2$, which represents the size of the term $\frac{1}{2}\mathbf{w}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{w}$. It is in this sense we see that $f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\cdot\mathbf{w} + \frac{1}{2}\mathbf{w}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{w}$ provides a "best quadratic approximation" to f near $\mathbf{x}_0$.

(6.3) is seen as follows: for any $\epsilon > 0$, we can find $\delta > 0$ such that for all $\mathbf{w}$ with $\|\mathbf{w}\| < \delta$, and all $\tau\in[0,1]$, $\|A(\tau)\|_F < \epsilon$. For such $\mathbf{w}$, we thus have

$$\|R_2\| \le \|\mathbf{w}\|\|(\int_0^1 A(\tau)d\tau)\mathbf{w}\| \le \|\mathbf{w}\|\int_0^1 \|(A(\tau)\mathbf{w}\|d\tau \le \epsilon\|\mathbf{w}\|^2.$$

In the remainder of this section, we use the second order directional derivative

$$\left.\frac{d^2}{dt^2}\right|_{t=0} f(\mathbf{x}_0+t\mathbf{v}) = \mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v} = (\mathbf{v})^T[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}$$

and the Taylor expansion

$$f(\mathbf{x}_0+\mathbf{v}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\cdot\mathbf{v} + \frac{1}{2}\mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v} + R_2$$

to study the local behavior of f near $\mathbf{x}_0$.

Example 6.2.3

For the function $f(x,y) = x^y$ for $x, y > 0$, its quadratic Taylor expansion at

$(1,1)$ is

$$\begin{aligned} f(1+v_1,1+v_2) &= f(1,1)+\partial_x f(1,1)v_1+\partial_y f(1,1)v_2 \\ &\quad +\frac{\partial^2_{xx}f(1,1)}{2}v_1^2+\partial^2_{xy}f(1,1)v_1v_2+\frac{\partial^2_{yy}f(1,1)}{2}v_2^2 \\ &= 1+v_1+v_1v_2, \end{aligned}$$

while its quadratic Taylor expansion at $(2,1)$ is

$$\begin{aligned} f(2+v_1,1+v_2) &= f(2,1)+\partial_x f(2,1)v_1+\partial_y f(2,1)v_2 \\ &\quad +\frac{\partial^2_{xx}f(2,1)}{2}v_1^2+\partial^2_{xy}f(2,1)v_1v_2+\frac{\partial^2_{yy}f(2,1)}{2}v_2^2 \\ &= 2+v_1+2\ln 2\, v_2+(1+\ln 2)v_1v_2+(\ln 2)^2v_2^2. \end{aligned}$$

When $\mathbf{v}$ is an eigenvector of $[\mathrm{Hess}_f(\mathbf{x}_0)]$: $[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}=\mu\mathbf{v}$ for some μ, then the term $\mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}=\mu\|\mathbf{v}\|^2$ simplifies. The main results about the quadratic approximation are

(a). If all the eigenvalues of $[\mathrm{Hess}_f(\mathbf{x}_0)]$ are positive, then $\left.\frac{d^2}{dt^2}\right|_{t=0} f(\mathbf{x}_0+t\mathbf{v})>0$ for all $\mathbf{v}$, and the graph of f near $\mathbf{x}_0$ would be approximated by $f(\mathbf{x}_0)+\nabla f(\mathbf{x}_0)\cdot\mathbf{v}+\frac{1}{2}\mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}$, which is the sum of a convex paraboloid $\frac{1}{2}\mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}$ and the tangent plane $f(\mathbf{x}_0)+\nabla f(\mathbf{x}_0)\cdot\mathbf{v}$.

(b). Likewise, if all the eigenvalues of $[\mathrm{Hess}_f(\mathbf{x}_0)]$ are negative, then $\left.\frac{d^2}{dt^2}\right|_{t=0} f(\mathbf{x}_0+t\mathbf{v})<0$ for all $\mathbf{v}$, and the graph of f near $\mathbf{x}_0$ would be approximated by $f(\mathbf{x}_0)+\nabla f(\mathbf{x}_0)\cdot\mathbf{v}+\frac{1}{2}\mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}$, which is the sum of a concave paraboloid $\frac{1}{2}\mathbf{v}\cdot[\mathrm{Hess}_f(\mathbf{x}_0)]\mathbf{v}$ and the tangent plane $f(\mathbf{x}_0)+\nabla f(\mathbf{x}_0)\cdot\mathbf{v}$.

(c). In particular, at a critical point $\mathbf{x}_0$ of f, if all the eigenvalues of $[\mathrm{Hess}_f(\mathbf{x}_0)]$ are positive, then f has a local minimum at $\mathbf{x}_0$; while if all the eigenvalues of $[\mathrm{Hess}_f(\mathbf{x}_0)]$ are negative, then f has a local maximum at $\mathbf{x}_0$.

(d). One can check whether all the eigenvalues of $[\mathrm{Hess}_f(\mathbf{x}_0)]$ are positive (or negative) without computing the eigenvalues directly by applying the

Sylvester's criterion in **6.2.7**, which is based on computing the determinants of a number of submatrices constructed based on the given matrix. This is useful for 2×2 and 3×3 matrices, but not as useful for matrices of bigger sizes.

(e). At a critical point, if the eigenvalues of $[\text{Hess}_f(\mathbf{x}_0)]$ have opposite signs, then f behaves as having a minimum in directions of eigenvectors associated with negative eigenvalues, and as having a maximum in directions of eigenvectors associated positive eigenvalues. Such a critical point is called a saddle critical point.

(f). One can also use the diagonalization of all the eigenvalues of $[\text{Hess}_f(\mathbf{x}_0)]$ are positive to construct the local contour curve of f near $\mathbf{x}_0$, as is done in **6.2.5.**

6.3 A Brief Discussion of Determinant

In our earlier discussion on mixed product in $\mathbb{R}^3$ (in section 1.2), we already mentioned that relation between the mixed product and determinant of a 3×3 matrix with the signed volume of the parallelepiped. This latter relation is the geometric motivation for a general $n \times n$ matrix.

Let's first review the situation in $\mathbb{R}^2$. Let $[\mathbf{u}, \mathbf{v}]$ denote the parallelogram spanned by $\mathbf{u}$ and $\mathbf{v}$,. Analytically, this means the set of vectors $\{s\mathbf{u} + t\mathbf{v} : 0 \leq s, t \leq 1\}$. Let $A[\mathbf{u}, \mathbf{v}]$ denote the area of this parallelogram. Then we have

$$A[\mathbf{u}, c\mathbf{v}] = cA[\mathbf{u}, \mathbf{v}], \quad \text{for all } \mathbf{u}, \mathbf{v}, \tag{a1}$$

at least for $c \geq 0$,

$$A[\mathbf{u}, \mathbf{v} + c\mathbf{u}] = A[\mathbf{u}, \mathbf{v}], \quad \text{for all } \mathbf{u}, \mathbf{v}, \text{ and } c. \tag{a2}$$

(a2) is a reflection of the geometric principle that *parallelograms with the same base and equal heights have equal areas*. Here $[\mathbf{u}, \mathbf{v} + c\mathbf{u}]$ and $[\mathbf{u}, \mathbf{v}]$ share a common base $\mathbf{u}$, and have equal heights with base $\mathbf{u}$. The following is a Desmos graph illustration* of areas of parallelograms with adjacent edges$[\mathbf{u}, \mathbf{v}]$ and $[\mathbf{u}, \mathbf{v} + c\mathbf{u}]$.

In order to extend (a1) to all c, we can either replace c on the right by $|c|$, or replace $A[\mathbf{u}, \mathbf{v}]$ by the *signed* area $sgA[\mathbf{u}, \mathbf{v}]$. $sgA[\mathbf{u}, \mathbf{v}]$ satisfies (a1) for all c, and satisfies (a2) as well; furthermore, we also have

$$sgA[\mathbf{u}, \mathbf{v} + \mathbf{w}] = sgA[\mathbf{u}, \mathbf{v}] + sgA[\mathbf{u}, \mathbf{w}], \quad \text{for all } \mathbf{u}, \mathbf{v}, \mathbf{w}. \tag{a3}$$

*https://www.desmos.com/calculator/895gqiflo9

(Construct corresponding parallelograms to give a geometric proof to the above.) The same properties hold if we do the same operations on the first vector $\mathbf{u}$. We would rather keep the very useful properties (a1) and (a3) and accept that $sgA[\mathbf{u}, \mathbf{v}]$ can take on negative values, than insisting on working with a nonnegative area function; we can always define, in the end, that $A[\mathbf{u}, \mathbf{v}] = |sgA[\mathbf{u}, \mathbf{v}]|$.

Based on (a1–3) and the normalization requirement that $A[\mathbf{e}_1, \mathbf{e}_2] = 1$, it is easy to derive that $sgA[\mathbf{u}, \mathbf{v}]$ must be given by the usual formula for the determinant of the 2×2 matrix $[\mathbf{u}, \mathbf{v}]$ with $\mathbf{u}, \mathbf{v}$ as its columns (we are abusing notation using the same $[\mathbf{u}, \mathbf{v}]$ to denote both the geometric parallelogram and the matrix).

Here is how the argument goes. Suppose $\mathbf{u} = \begin{bmatrix} a \\ b \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} c \\ d \end{bmatrix}$. Let's first assume that $a \neq 0$, then $\mathbf{u} = a \begin{bmatrix} 1 \\ b/a \end{bmatrix}$, so according to (a1),

$$\det[\mathbf{u}, \mathbf{v}] = a \det \begin{bmatrix} 1 & c \\ b/a & d \end{bmatrix}.$$

Next $\mathbf{v}$ subtracting $c*$ the first column of the matrix on the right hand gives $\begin{bmatrix} 0 \\ d - cb/a \end{bmatrix}$, and according to (a2) and (a1),

$$\det \begin{bmatrix} 1 & c \\ b/a & d \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ b/a & d - cb/a \end{bmatrix} = (d - cb/a) \det \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix}.$$

Finally, the first column of the above matrix subtracting $b/a*$ the second column produces $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, so

$$\det \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1.$$

Putting these together, we get

$$\det[\mathbf{u}, \mathbf{v}] = a(d - cb/a) = ad - bc.$$

When $a = 0$ and $c \neq 0$, or when $a = c = 0$, we just need to adjust the above arguments and in both cases we arrive at the same conclusion.

The same argument also works for a 3×3 matrix, and can be used to define the determinant of any $n \times n$ matrix, namely,

$$\det[\mathbf{v}_1, \cdots, \mathbf{v}_n] = \text{signed volume}[\mathbf{v}_1, \cdots, \mathbf{v}_n],$$

where signed volume$[\mathbf{v}_1, \cdots, \mathbf{v}_n]$ is the n-dimensional analogue of parallelepiped and obeys similar properties as in (a1–3).

(a2) means that $\det A$ is unchanged when the columns of A undergo the kind of "elementary column operations" as described in (a2). (a1) implies that if A has a column of 0's, then $\det A = 0$. These together also imply that when any column of A can be expressed as a linear combination of some other columns (namely when A is not invertible), then $\det A = 0$, as one can then perform a number of operations as in (a2) to produce a column of 0's. Furthermore, for any diagonal matrix D with its diagonal entries $d_1, \cdots, d_n$,

$$\det D = d_1 \cdot \ldots \cdot d_n.$$

This follows by applying (a1) multiple times to produce

$$\det D = d_1 \cdot \ldots \cdot d_n \cdot \det I_n = d_1 \cdot \ldots \cdot d_n,$$

where we still demand the normalization condition that $\det I_n = 1$.

For any upper triangular matrix R, if it has n pivots, then we can repeatedly apply the column operations in (a2) to reduce R to the diagonal matrix whose entries on the diagonal are the same as those of R, therefore conclude that $\det R =$ product of its diagonal entries.

At this point we don't have a definition for the volume or signed volumes of a general region in $\mathbb{R}^n$, but any region spanned by n vectors in $\mathbb{R}^n$, the properties (a1–a3) are natural ones for the notion of signed volume. Based on this interpretation, $\det Q = \pm 1$ for any orthogonal matrix Q.

$\det A$ also plays the role of magnifying factor between the signed volumes of a region U and its image $A(U)$ under the linear function defined by A:

$$\text{signed volume of } A(U) = \det A \text{ (signed volume of } U).$$

This interpretation forces $\det(AB) = \det A \det B$, as

$$\begin{aligned} \text{signed volume of } AB(U) &= \det A \text{ (signed volume of } B(U)) \\ &= \det A \det B \text{ (signed volume of } U). \end{aligned}$$

As a consequence of this property,

$$\det(AB) = \det A \det B = \det(BA),$$

even though $AB \neq BA$ may happen. In addition, when A is invertible,

$$1 = \det I_n = \det(AA^{-1}) = \det A \det A^{-1},$$

so $\det(A^{-1}) = 1/\det(A)$.

As this point, when A has a QR factorization $A = QR$, the above properties imply that $\det A = \det Q \det R$, where $\det Q = \pm 1$, and $\det R$ is the product of its diagonal

entries. The only remaining issue is to determine how to tell whether $\det Q = 1$ or -1.

Another useful fact is that $\det A^T = \det A$.

As a consequence of the above properties, when a symmetric matrix A is diagonalized by an orthogonal matrix Q: $A = QDQ^T$,

$$\det A = \det Q \det D \det Q^T = (\det Q)^2 \det D = \det D,$$

where $\det D$ is the product of its diagonal entries, which is the product of the eigenvalues of A.

Exercise 6.3.1. *Applying (a1-3) to calculate*

$$\det \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Exercise 6.3.2. *Applying (a1-3) to calculate the determinant of the Jacobian matrix of the map from cylindrical coordinates (r, θ, z) to the rectangular coordinates (x, y, z):*

$$\det \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

*(*HINT: *The columns are orthogonal to each other.)*

Exercise 6.3.3. *Applying (a1-3) to calculate the determinant of the Jacobian matrix of the map from spherical polar coordinates (r, θ, ϕ) to the rectangular coordinates (x, y, z):*

$$\det \begin{bmatrix} \sin\phi\cos\theta & -r\sin\phi\sin\theta & r\cos\phi\cos\theta \\ \sin\phi\sin\theta & r\sin\phi\cos\theta & r\cos\phi\sin\theta \\ \cos\phi & 0 & -r\sin\phi \end{bmatrix}.$$

*(*HINT: *The columns are orthogonal to each other.)*

Chapter 7

INTEGRATION IN SEVERAL VARIABLES

7.1 Integration and summation

Although the rough ideas in defining integrals in several variables are similar to those in one variable, there are some major differences, with the major ones being

(a) the domain of integration in several variables can be much more varied and complicated, as opposed to an interval in one dimension;

(b) the partition of the domain of integration would involve complications: except when the domain is a rectangular box whose faces (or edges) are parallel to the coordinate axes, if we do partition of the domain using small rectangular boxes whose faces (or edges) are parallel to the coordinate axes, some boxes would *straddle* between the domain and its complement, and we would have to account for the impact of contributions or omissions from these boxes when forming the Riemann sum.

Issue (b) turns out to be the main source of difficulty in defining integrals in several variables*. One possible remedy is to enclose the domain U to be enclosed in a rectangular box $\mathcal{R}$ whose faces (or edges) are parallel to the coordinate axes (assuming U to be bounded), and extend the integrand $f(\mathbf{x})$ to be 0 in $\mathcal{R} \setminus U$. But the extended function is generally discontinuous at points on the boundary ∂U of U, and these points of discontinuity could create difficulties for having a well defined integral.

*Look up Osgood curves and Knopp curves, whose fractal geometry creates difficulties in defining the area of its complement in a rectangle enclosing the curve using Riemann's approximation. The area (integral) can be defined using a different procedure due to Lebesgue.

The issues arise as follows. Once we partition the given domain U as the non-overlapping union of small boxes $\mathcal{R}_k$, some of which are rectangular boxed contained within U, and some of which straddle between U and its complement, we need to sample a point $\mathbf{x}_k$ in $\mathcal{R}_k$ and form the Riemann sum

$$\sum_k f(\mathbf{x}_k) \text{ Area}(R_k)$$

and study whether this Riemann sum has a limit independent of how we do the partition and pick $\mathbf{x}_k$ with $\mathcal{R}_k$ as the partition size goes to 0.

But for those $\mathcal{R}_k$ which straddle between U and its complement, a more appropriate choice for its contribution to the Riemann sum would have been $f(\mathbf{x}_k)$ Area$(R_k \cap U)$. However, Area$(R_k \cap U)$ is not something that we know how to compute.

Our alternative strategy is to extend the integrand f to be 0 outside of U and use Area(R_k) directly; but then for these straddling $\mathcal{R}_k$, the term $f(\mathbf{x}_k)$ Area(R_k) would be 0 if $\mathbf{x}_k$ is chosen to be a point in $\mathcal{R}_k$ but outside of U, which may not reflect the actual contribution from such a region.

Take the case of $f \equiv 1$, then the term $f(\mathbf{x}_k)$ Area(R_k) could vary between 0 and Area$(R_k \cap U)$, and the overall accumulation of these terms in the Riemann sum would be varying between 0 and $\sum$ Area$(R_k \cap U)$, the latter may not approach 0 as the partition size goes to 0, as is the case if the boundary ∂U has a fractal nature.

We will not aim to define integrals on most general domains in multi-dimensions; instead, we will focus on defining integrals on domains which we usually encounter in applications. These include

(i). Rectangular boxes $\mathcal{R}$ whose faces (or edges) are parallel to the coordinate axes;

(ii). Domains in $\mathbb{R}^2$ which can be described as $\{(x, y) : c \le x \le d, a(x) \le y \le b(x)\}$, where $a(x) \le b(x)$ are given continuous functions defined on the interval $[c, d]$ whose graphs describe the upper and lower portion of the boundary of this domain;

(iii). Domains in $\mathbb{R}^2$ which can be described as $\{(x, y) : c \le y \le d, a(y) \le x \le b(y)\}$, where $a(y) \le b(y)$ are given continuous functions defined on the interval $[c, d]$ whose graphs describe the left and right portion of the boundary of this domain;

(iv). Domains which can be partitioned into a finite union of non-overlapping subdomains, each of which is of a type of the above;

(v). Domains which in a rotated rectangular coordinates, or in polar/spherical/cylindrical coordinates, or under a change of variables, have a structure similar to one of the above.

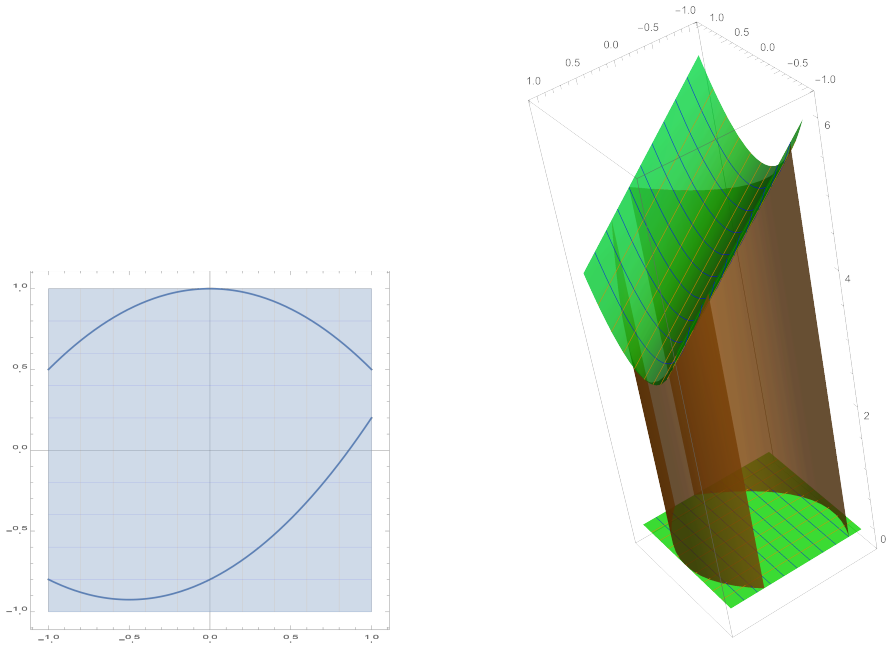


Figure 7.1: A region of integration and a partition of this region by small rectangles, as well as a graph over this region

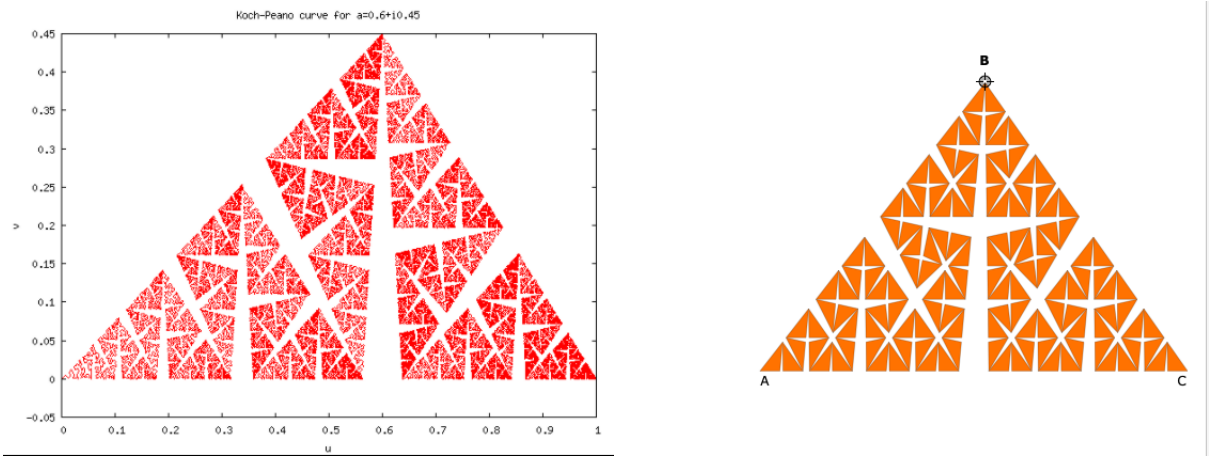


Figure 7.2: On the left is an Osgood curve which is obtained after an infinite iteration of removing a certain proportion from each triangle left from the previous iteration; a partition of the enclosing rectangle by small rectangles would always have a positive portion overlapping with the Osgood curve. On the right is the shape after only six iterations.

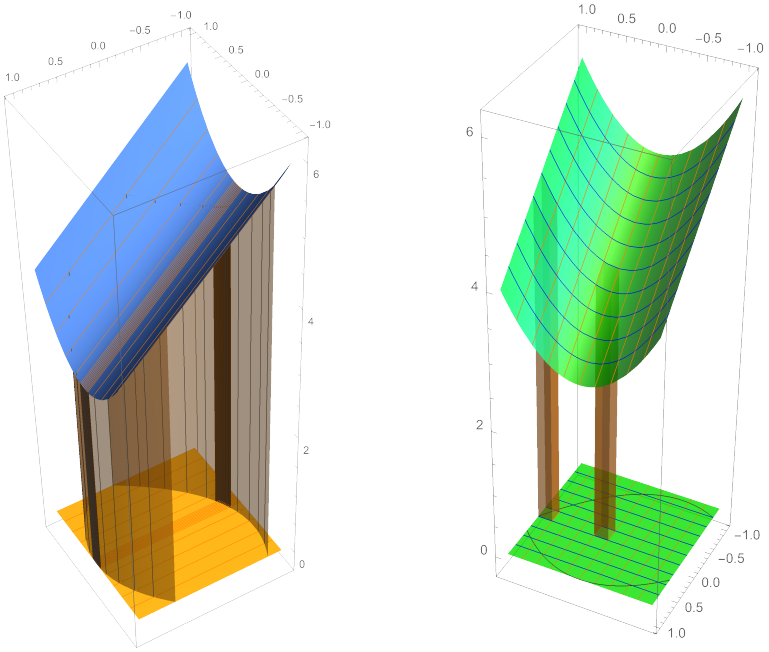


Figure 7.3: A partition of the region under the graph by thin slices or by small rectangular columns; some columns "straddle" on the boundary of the region

For the domains that we list above, the potential complications of contributions from the boundary terms would not arise when the integrand is a continuous function on the closed domain of integration. As a result the integral of a continuous function on such a domain is well defined. What we need to figure out is how to evaluate such an integral without having to use the definition.

Here are two basic properties of integrals, which are not stated explicitly in Professor Carlen's notes:

(I) If both $f(\mathbf{x})$ and $g(\mathbf{x})$ are integrable over the set U, then so is $af(\mathbf{x})+bg(\mathbf{x})$ for any constants a and b, and $\int_U (af(\mathbf{x})+bg(\mathbf{x}))dA = a\int_U f(\mathbf{x})dA + b\int_U g(\mathbf{x})dA$*

(II) If U and V are non-overlapping sets, and $\int_U f(\mathbf{x})dA$ and $\int_V f(\mathbf{x})dA$ are both well defined, then $\int_{U\cup V} f(\mathbf{x})dA$ is also well defined, and $\int_{U\cup V} f(\mathbf{x})dA = \int_U f(\mathbf{x})dA + \int_V f(\mathbf{x})dA$.

*$\int_U(\cdots)dA$ in Carlen's notes denotes an integral in two variables; but these two properties are the requirements for integrals in any number of variables. It is common to use $\iint_U(\cdots)dA$, or $\iint_U(\cdots)dx\,dy$ to denote an integral in two variables (x,y), and use $\iiint_U(\cdots)dV$, or $\iiint_U(\cdots)dx\,dy\,dz$ to denote an integral in three variables (x,y,z).

Another important point to note is that, in defining integrals in several variables, we need not restrict ourselves to partitioning the domain only using small rectangular boxes whose faces (or edges) are parallel to the coordinate axes; it is often more efficient to partition the domain using thin long slices or wedges, and there are often different ways of doing such partitions. This is the issue of appropriate "*disintegration*", as described in Professor Carlen's notes. Technically, this amounts to tallying up the sum of contributions from small rectangular boxes which constitute a thin slice or wedge, then tallying up these subtotals; in other words, compute certain one-variable integral according to the slicing scheme, then compute another integral with the previous integral as integrand. In summary, we often evaluate an integral of several variables via a certain choice of iterated (one-variable) integrals.

It would take some subtle discussions to prove properly that the integral of a continuous function on a domain as described above is well defined, and the different ways of disintegration (such as (7.7) and (7.9) in the notes, called iterated integrals in many other texts) would provide equivalent ways of evaluating the integral. Professor Carlen's discussion on pp.267-268 is only a heuristic argument, and should not be taken as an actual proof.

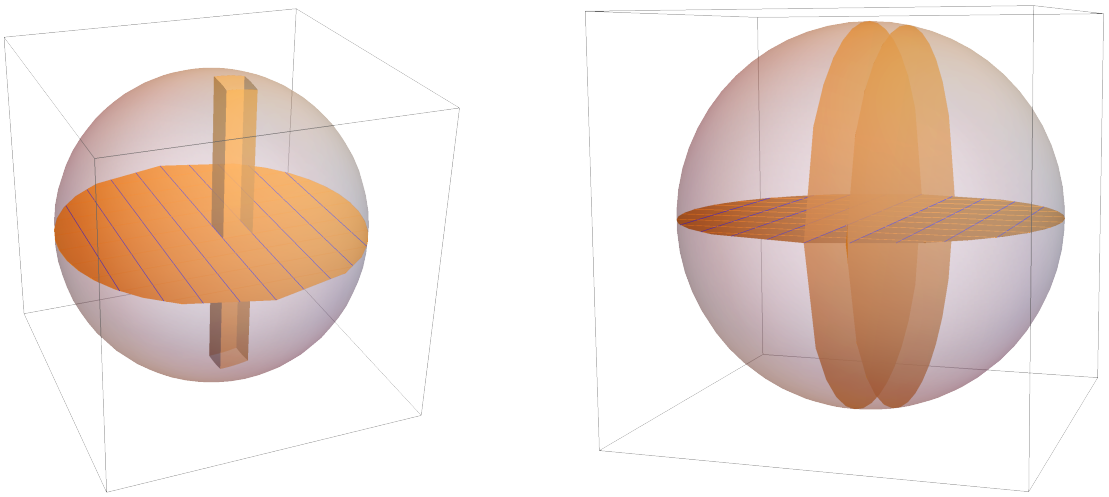


Figure 7.4: Finding the volume of a sphere by partitioning it into thin rectangular columns, or into thin cylindrical slices.

Example 7.1.1

We will use two different disintegration approaches to set up and evaluate the volume of a ball of radius R in $\mathbb{R}^3$. The equation for the ball will take a simple form when centered at the origin: $x^2 + y^2 + z^2 = R^2$.

We can consider the volume of the ball as the integration of thin rectangular columns over squares of dimensions $\Delta x \times \Delta y$ with (x, y) as one of its corners, where (x, y) lies in the disk $x^2 + y^2 \leq R^2$ in the $z = 0$ plane, and the height of the rectangular column is $\sqrt{R^2 - x^2 - y^2}$. Thus the volume of this ball is

$$2 \iint_{x^2+y^2 \leq R^2} \sqrt{R^2 - x^2 - y^2}\, dA.$$

To evaluate this integral, we can think of it as the integration in x of the area of the slice of the graph $z = \sqrt{R^2 - x^2 - y^2}$ as a function of y over $[-\sqrt{R^2 - x^2}, \sqrt{R^2 - x^2}]$. Thus

$$2 \iint_{x^2+y^2 \leq R^2} \sqrt{R^2 - x^2 - y^2}\, dA = 2 \int_{-R}^{R} \left(\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \sqrt{R^2 - x^2 - y^2}\, dy \right) dx.$$

In principle, this iterated integral can be carried out, although it is not that straightforward to compute $\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \sqrt{R^2 - x^2 - y^2}\, dy$ directly using only the standard calculus tools.

We can also disintegrate this volume by slicing the ball by planes perpendicular to the x-axis: suppose the slices cut the x-axis at $-R = x_0 < x_1 < \ldots < x_N = R$, with $\Delta x_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$, and ΔV_i denoting the volume of the ball between the slices $x = x_{i-1}$ and x_i. ΔV_i can be approximated by the volume of a cylindrical slice with a disk of radius $r_i = r(x_i) = \sqrt{R^2 - x_i^2}$ as its base and thickness Δx_i. Set $E_i = \Delta V_i - \pi r_i^2 \Delta x_i$. Then E_i is no more than the volume of a cylindrical ring with its cross section having area $|r_i - r_{i-1}|\Delta x_i \approx |r'(x_i)||\Delta x_i|^2$, which is vanishingly small percentage-wise in comparison to either $\pi r_i^2 \Delta x_i$ or $|\Delta x_i|$. In fact,

$$\sum_{i=1}^{N} \Delta V_i = \pi \sum_{i=1}^{N} r_i^2 \Delta x_i + \sum_{i=1}^{N} E_i,$$

and $\sum_{i=1}^{N} E_i \to 0$ as $\max |\Delta x_i| \to 0$. Thus the volume of this ball is also equal to $\pi \int_{-R}^{R} r(x)^2 dx = \pi \int_{-R}^{R} (R^2 - x^2) dx = \frac{4\pi}{3} R^3$.

In fact, if we interpret $\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \sqrt{R^2 - x^2 - y^2}\, dy$ as the area of the semi-disk of radius $\sqrt{R^2 - x^2}$, we know it is equal to $\pi(R^2 - x^2)/2$, which also enables us to conclude that

$$2 \int_{-R}^{R} \left(\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \sqrt{R^2 - x^2 - y^2}\, dy \right) dx = \int_{-R}^{R} \pi(R^2 - x^2)\, dx.$$

We will also discuss how to evaluate $\iint_{x^2+y^2\le R^2} \sqrt{R^2-x^2-y^2}\, dA$ using polar coordinates.
A third approach to formulate the volume of this ball is to set it up as an integration in three variables: it is the integral of the function 1 over the ball: $\iiint_{x^2+y^2+z^2\le R^2} 1\, dV$, as it is the limit of the sum of small three dimensional rectangular boxes with dimensions Δx, Δy, and Δz. We can "disintegrate" this integral in three variables as three iterated integrals

$$\int_{-R}^{R} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \int_{-\sqrt{R^2-x^2-y^2}}^{\sqrt{R^2-x^2-y^2}} 1\, dz\, dy\, dx,$$

where the limits of these iterated integrals are determined by the rule that for $-R \le x \le R$, $-\sqrt{R^2-x^2} \le y \le \sqrt{R^2-x^2}$ describes the range of the y variable along the segment within the disk $x^2+y^2 \le R^2$ in the $z=0$ plane (the intersection of the sphere with the $z=0$ plane), and for each such pair (x,y), the range $-\sqrt{R^2-x^2-y^2} \le z \le \sqrt{R^2-x^2-y^2}$ describes the range of the z variable along the vertical segment within the ball $x^2+y^2+z^2 \le R^2$. If we carry out the inner-most integral, $\int_{-\sqrt{R^2-x^2-y^2}}^{\sqrt{R^2-x^2-y^2}} 1\, dz$, we get $2\sqrt{R^2-x^2-y^2}$, and we have reduced the triple integral in three variables into a double integral in two variables.

Reading Quizzes/Questions: Let D denote the region enclosed by the circle $(x-1)^2+y^2=1$. Express the integral $\iint_D x\, dA$ in the following three different ways and then evaluate the iterated integrals: (a). Integrate in y-variable first; (b) Integrate in x-variable first; (c) Integrate in polar-coordinates.

Example 7.1.2

Sometimes one form of iterated integral is easier to compute than others. For example

$$\int_0^1 \int_{e^y}^{e} \frac{1}{\ln x}\, dx dy,$$

would take some effort to carry out the integration x-variable. But re-examining this integral makes one realize that the domain of integration can also be de-

scribed as $\{(x,y) : 1 \le x \le e, 0 \le y \le \ln x\}$, so

$$\int_0^1 \int_{e^y}^e \frac{1}{\ln x}\, dxdy = \int_1^e \int_0^{\ln x} \frac{1}{\ln x}\, dydx = \int_1^e \frac{y}{\ln x}\Big|_{y=0}^{y=\ln x} dx = e - 1.$$

Integration in Polar coordinates

Given an integration in rectangular coordinates $\iint_D f(x,y)\, dxdy$, if representing D in terms of polar coordinates $(r,\theta) \in U \mapsto (x,y) = (r\cos\theta, r\sin\theta) \in D$, how would we compute this integral in terms of (r,θ)?

This amounts to doing partition in terms of r and θ: for a rectangle in the (r,θ) coordinates: $r_i \le r \le r_{i+1} := r_i + \Delta r_i$ and $\theta_j \le \theta \le \theta_{j+1} := \theta_j + \Delta\theta_j$, its image in (x,y) is a truncated sector with angle opening $\Delta\theta_j$ and inner and outer radii r_i and r_{i+1} respectively. The area of the image is

$$\frac{1}{2}(r_i + \Delta r_i)^2 \Delta\theta_j - \frac{1}{2} r_i^2 \Delta\theta_j = r_i \Delta r_i \Delta\theta_j + \frac{1}{2}(\Delta r_i)^2 \Delta\theta_j.$$

Note that $r_i\Delta\theta_j$ is the length of the inner circular arc of the sector, so $r_i\Delta r_i\Delta\theta_j$ is an approximation for the area of this sector, if we treat it as a rectangle.

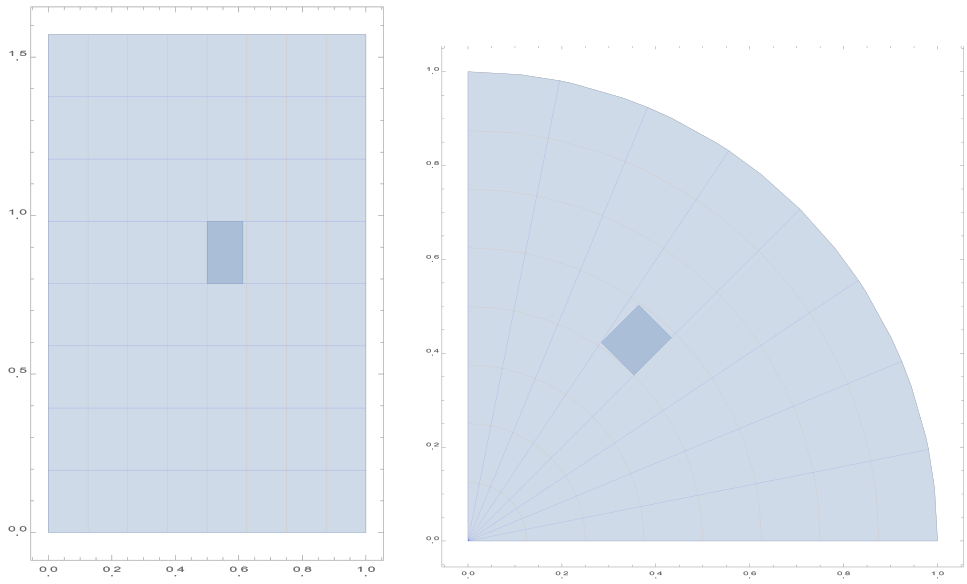


Figure 7.5: A rectangle in polar coordinates and its image in rectangular coordinates.

In defining $\iint_D f(x,y)\,dxdy$ using such partitions in U, we need to examine

$$\lim_{\Delta r_i,\Delta_j\to 0}\sum_{i,j} f(r_i\cos\theta_j, r_i\sin\theta_j)(\text{Area of Image of Polar Rectangle } [r_i, r_i+\Delta r_i]\times[\theta_j,\theta_j+\Delta\theta_j])$$

$$=\lim_{\Delta r_i,\Delta_j\to 0}\sum_{i,j} f(r_i\cos\theta_j, r_i\sin\theta_j)(r_i\Delta r_i\Delta\theta_j+\frac{1}{2}(\Delta r_i)^2\Delta\theta_j).$$

We assume that there exists some $M>0$ such that $|f(x,y)|\le M$ for all $(x,y)\in D$. We then claim that

$$\lim_{\Delta r_i,\Delta_j\to 0}\sum_{i,j} f(r_i\cos\theta_j, r_i\sin\theta_j)r_i\Delta r_i\Delta\theta_j,$$

is finite, and

$$\lim_{\Delta r_i,\Delta_j\to 0}\sum_{i,j} \frac{1}{2}f(r_i\cos\theta_j, r_i\sin\theta_j)(\Delta r_i)^2\Delta\theta_j=0.$$

It then follows that $\iint_D f(x,y)\,dxdy=\iint_U f(r\cos\theta,r\sin\theta)r\,dr\,d\theta$.

The claim is based on $\sum_{i,j}\Delta r_i\Delta\theta_j=$ Area of rectangle in polar coordinate, and

$$\left|\sum_{i,j}\frac{1}{2}f(r_i\cos\theta_j,r_i\sin\theta_j)(\Delta r_i)^2\Delta\theta_j\right|\le\frac{M}{2}\sum_{i,j}(\Delta r_i)^2\Delta\theta_j\le\frac{M}{2}\max|\Delta r_i|\sum_{i,j}\Delta r_i\Delta\theta_j,$$

so it tends to 0 as $\max|\Delta r_i|\to 0$.

The key of the above argument is that, of the two terms in $r_i\Delta r_i\Delta\theta_j+\frac{1}{2}(\Delta r_i)^2\Delta\theta_j$, the latter is a vanishingly small proportion of the former — both terms individually are tending to 0, but when summed up, the first term gives a finite limit, and the second term then has 0 as its limit.

We are going to do an analysis of change of variables in the more general context in the next section. The general idea is to use linear approximation to account for the change of area: when we use linear approximation to $X=(x,y)=(r\cos\theta,r\sin\theta)$ at (r_i,θ_j), we get

$$\begin{bmatrix} r_i\cos\theta_j\\ r_i\sin\theta_j\end{bmatrix}+\begin{bmatrix}\cos\theta_j & -r_i\sin\theta_j\\ \sin\theta_j & r_i\cos\theta_j\end{bmatrix}\begin{bmatrix} r-r_i\\ \theta-\theta_j\end{bmatrix}.$$

where $\begin{bmatrix}\cos\theta_j & -r_i\sin\theta_j\\ \sin\theta_j & r_i\cos\theta_j\end{bmatrix}$ is the Jacobian matrix of $(r,\theta)\mapsto(r\cos\theta,r\sin\theta)$ at (r_i,θ_j), and its determinant is r_i, which is the factor in front of $\Delta r_i\Delta\theta_j$ in $r_i\Delta r_i\Delta\theta_j$!

7.2 Jacobians and changing variables of integration in $\mathbb{R}^2$

7.2.1 Letting the boundary of D determine the disintegration strategy

The key underlying basis for the method of this subsection is **Theorem 83**. Here is a different proof of **Theorem 83** using the QR factorization of A: Suppose that $A = Q\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$, where we assume that both columns of A are pivotal, so $a, d > 0$. We further note that $\begin{bmatrix} a & b \\ 0 & d \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & b/(ad) \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}$, and use this to accomplish the transformation $\mathbf{x} = A\mathbf{u}$ as the composition of four transformations:

$$\mathbf{u} \overset{f_1}{\mapsto} \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}\mathbf{u} =: \mathbf{v} \overset{f_2}{\mapsto} \begin{bmatrix} 1 & b/(ad) \\ 0 & 1 \end{bmatrix}\mathbf{v} =: \mathbf{w} \overset{f_3}{\mapsto} \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}\mathbf{w} =: \mathbf{y} \overset{f_4}{\mapsto} Q\mathbf{y} =: \mathbf{x}.$$

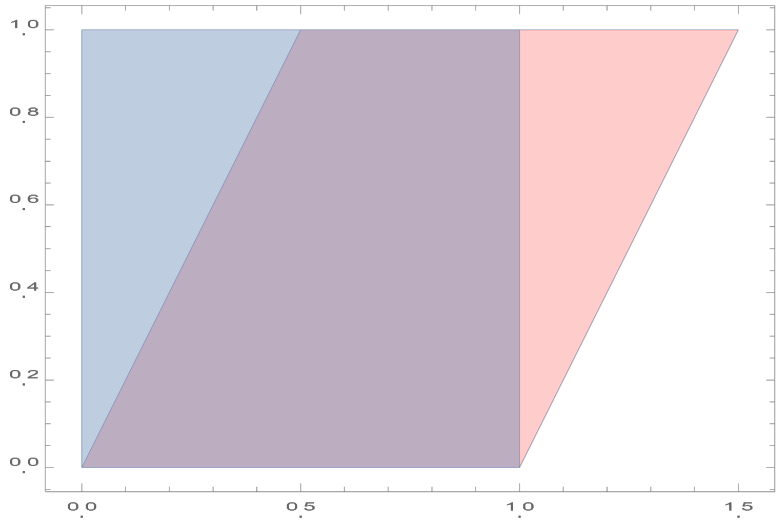


Figure 7.6: A square $[0, h] \times [0, h]$ and its image under $(x, y) \mapsto (x + cy, y)$.

It is now relatively easy to see that for any region U, $\text{area}(f_1(U)) = d\,\text{area}(U)$, $\text{area}(f_2(U)) = \text{area}(U)$, $\text{area}(f_3(U)) = a\,\text{area}(U)$, and $\text{area}(f_4(U)) = \text{area}(U)$. The second equality follows because f_2 maps any square whose sides are parallel to the x and y axis respectively into a parallelogram with the same base length and height, so the area of the parallelogram is equal to the area of the pre-image square (for example, the square with vertices $(0, 0), (h, 0), (h, h), (0, h)$ is mapped into the parallelogram

with vertices $(0,0),(h,0),((1+b/(ad))h,h),(bh/(ad),h))$, as a result, area$(f_2(U))$ = area(U). The fourth equality holds because f_4, arising from an orthogonal matrix, does not change the size and shape of any geometric figure.

It then follows that area$(f_4 \circ f_3 \circ f_2 \circ f_1(\widehat{D})) = ad\,$area$(\widehat{D})$. Finally, $\det(Q) = \pm 1$, and $\det(A) = \det(Q)ad = \pm ad$. Thus $ad = |\det(A)|$, and area$(f_4 \circ f_3 \circ f_2 \circ f_1(\widehat{D})) = |\det(A)|$area$(\widehat{D})$.

7.2.2 The change of variables formula for integrals in $\mathbb{R}^2$

The key here is the linear approximation of a continuously differentiable map $\mathbf{x} = \Phi(\mathbf{u})$ near any $\mathbf{u}_0$: $\Phi(\mathbf{u}) \approx \Phi(\mathbf{u}_0) + [D_{\mathbf{u}}\Phi(\mathbf{u}_0)](\mathbf{u} - \mathbf{u}_0)$ for $\mathbf{u}$ near $\mathbf{u}_0$. Then the image of a small rectangular box $\mathcal{R}$ with a vertex at $\mathbf{u}_0$ will be mapped into a shape by Φ, which can be approximated by the image of $\mathcal{R}$ by $[D_{\mathbf{u}}\Phi(\mathbf{u}_0)]$ (the action of $\Phi(\mathbf{u}_0)+$ merely translates every point by the same $\Phi(\mathbf{u}_0)$, so does not cause any change of area). But according to our discussion from the last subsection, the area of the latter is $|\det(D_{\mathbf{u}}\Phi(\mathbf{u}_0))|$ times the area of $\mathcal{R}$. This $|\det(D_{\mathbf{u}}\Phi(\mathbf{u}_0))|$ factor varies with $\mathbf{u}_0$, and it can be made rigorous to use the images of small rectangular boxes in $\mathbf{u}$ coordinates under the map Φ, or rather its approximation $D_{\mathbf{u}}\Phi(\mathbf{u})$, to disintegrate $\iint_D f(\mathbf{x})dA$ to prove

$$\iint_D f(\mathbf{x})\; d^2\mathbf{x} = \iint_{\widehat{D}} f(\Phi(\mathbf{u}))|\det(D_{\mathbf{u}}\Phi(\mathbf{u}))|\; d^2\mathbf{u},$$

where, instead of using dA in the two integrals, we use $d^2\mathbf{x}$ to stand for dA with respect to the $\mathbf{x}$ variable, and $d^2\mathbf{u}$ to stand for dA with respect to the $\mathbf{u}$ variable, and $D = \Phi(\widehat{D})$.

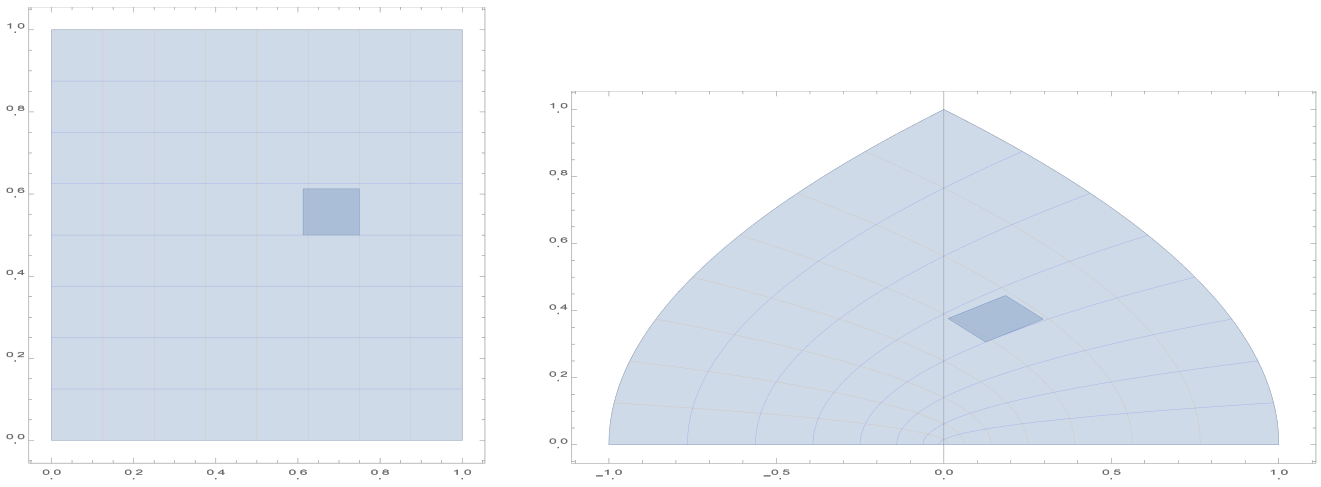


Figure 7.7: A square $[0,1]\times[0,1]$ and its image under $(x,y)\mapsto(x^2-y^2,xy)$.

Here are some more details about this process.

- Partition $\widehat{D}$ by a collection of small rectangular boxes $\{\mathcal{R}_i\}$ whose edges are parallel to the u and v axes, respectively. Suppose Δu_i and Δv_i are the dimensions of this box.

- Then $\{\Phi(\mathcal{R}_i)\}$ forms a partition of D. Let $\mathbf{u}_i$ denote a vertex of $\mathcal{R}_i$, then $\Phi(\mathcal{R}_i)$ can be approximated by the parallelogram with vertex $\Phi(\mathbf{u}_i)$, and with edges $\Phi_u(\mathbf{u}_i)\Delta u$ and $\Phi_v(\mathbf{u}_i)\Delta v$. Here $\Phi_u(\mathbf{u}_i) = \dfrac{\partial \Phi(\mathbf{u})}{\partial u}\Big|_{\mathbf{u}=\mathbf{u}_i}$ is the partial derivative of the vector-valued function $\Phi(\mathbf{u})$ with respect to u, and similarly for $\Phi_v(\mathbf{u}_i)$. Recall that the derivative $[D\Phi(\mathbf{u}_i)]$ is a matrix whose two columns are $\Phi_u(\mathbf{u}_i)$ and $\Phi_v(\mathbf{u}_i)$ respectively.

- The area of the parallelogram in the previous item is $|\Phi_u(\mathbf{u}_i) \times \Phi_v(\mathbf{u}_i)|\Delta u_i \Delta v_i = |\det[D\Phi(\mathbf{u}_i)]|\Delta u_i \Delta v_i$. Thus

$$\sum f(\Phi(\mathbf{u}_i))|\det[D\Phi(\mathbf{u}_i)]|\Delta u_i \Delta v_i$$

  provides an approximation of the Riemann sum of $\iint_D f(\mathbf{x})\ dA$ with the partition from $\{\Phi(\mathcal{R}_i)\}$. In the limit, we should get

$$\iint_D f(\mathbf{x})\ dA = \iint_{\widehat{D}} f(\Phi(\mathbf{u}))|\det(D_{\mathbf{u}}\Phi(\mathbf{u}))|\ d^2\mathbf{u}.$$

7.3 Integration in $\mathbb{R}^3$

7.3.1 Reduction to iterated integrals in lower dimension

The general idea of defining and computing integrals of three variables is the same as that for two variables: partition the domain into non-overlapping union of shapes (usually rectangular boxes or slabs) whose volume can be computed easily, account for the contribution from each piece, and take the limit to define the integral; the computation also follows a similar line — finding a good way to slice the domain to make the computation of the partitioned part easy to compute, and eventually reducing the computation to several iterated integrals of one variable. Another related idea is to introduce a change of variables to make it easier to do partition in the new variables, but modifying the integrand with the absolute value of the determinant of the Jacobian matrix of the change of variables.

Example 7.3.1

Let's compute the integration of x^2+y^2 inside the solid ball $x^2+y^2+z^2 \le R^2$. There are multiple ways to slice the solid ball.
One way is to treat the solid ball as thin columns (of varying heights) sitting on top of the disk $\{(x,y,0) : x^2+y^2 \le R^2\}$, namely, for each such (x,y), the range of z is determined by $-\sqrt{R^2-x^2-y^2} \le z \le \sqrt{R^2-x^2-y^2}$. Thus we can transform the triple integral into

$$\iiint_{x^2+y^2+z^2\le R^2} \left(x^2+y^2\right)\, dV = \iint_{x^2+y^2\le R^2} \int_{-\sqrt{R^2-x^2-y^2}}^{\sqrt{R^2-x^2-y^2}} \left(x^2+y^2\right)\, dz\, dx\, dy.$$

The integration in z is carried out easily as $2\left(x^2+y^2\right)\sqrt{R^2-x^2-y^2}$, so we need to evaluate the double integral

$$\iint_{x^2+y^2\le R^2} 2\left(x^2+y^2\right)\sqrt{R^2-x^2-y^2}\, dx\, dy.$$

which is most easily done by converting it into integration in polar coordinates. Another way to slice the solid ball is to treat it as stacks of discs of radius $\sqrt{R^2-z^2}$ as z varies from $-R$ to R, so

$$\iiint_{x^2+y^2+z^2\le R^2} \left(x^2+y^2\right)\, dV = \int_{-R}^{R} \iint_{x^2+y^2\le R^2-z^2} \left(x^2+y^2\right)\, dxdy\, dz$$

The double integral in x and y is also most easily done in polar coordinates:

$$\iint_{x^2+y^2\le R^2-z^2} \left(x^2+y^2\right)\, dxdy = \int_0^{2\pi} \int_0^{\sqrt{R^2-z^2}} r^2\, r\, dr\, d\theta = \frac{\pi}{2}(R^2-z^2)^2.$$

It is now easy to carry out the integration in z to get

$$\iiint_{x^2+y^2+z^2\le R^2} \left(x^2+y^2\right)\, dV = \int_{-R}^{R} \frac{\pi}{2}(R^2-z^2)^2\, dz = \frac{8\pi R^5}{15}.$$

Later on, we will discuss how to evaluate the triple integral in spherical coordinates.

Reading Quizzes/Questions:

(i). Carry out the integrations in the example above according to the first approach

to confirm that it gives the same answer.

(ii). Formulate the integral $\iiint_D 1\, dV$ as three different iterated integrals and evaluate them, where D is the region enclosed by $x \geq 0, y \geq 0, z \geq 0, x+2y+3z \leq 6$.

(iii). Formulate the volume of the solid given by $|x| + |y| + |z| \leq 1$ as an iterated integral and find the volume.

7.3.2 The Change of Variables Formula for integrals in $\mathbb{R}^3$

Here the idea is similar: If $\mathbf{u} \in \hat{D} \mapsto \mathbf{x} = \Phi(\mathbf{u}) \in D$ is the change of variable to be used, then we need to partition $\hat{D}$ into non-overlapping union of rectangular boxes, and need to examine how the volume of $\Phi(R)$ relates to that of R, here R is one of the small rectangular boxes with a vertex at $\mathbf{u}_i$ and with its axes parallel to the $\mathbf{u}$ coordinate axes. Again we use the linear approximation to $\Phi(\mathbf{u})$ near $\mathbf{u}_i$: $\Phi(\mathbf{u}_i) + [D_{\mathbf{u}}\Phi(\mathbf{u}_i)](\mathbf{u} - \mathbf{u}_i)$ to analyze this.

Suppose $\mathbf{u} = (u, v, w)$, and the lengths of the edges of R along the u, v, and w axes are Δu_i, Δv_i, and Δw_i respectively, then the linear approximation maps the edge along the u-axis to a segment starting at $\Phi(\mathbf{u}_i)$ and adding a displacement of $D_u\Phi(\mathbf{u}_i)\Delta u_i$ (Make sure to understand this point!), and similarly for the other two edges. In other words, the linear approximation maps R into a parallelepiped whose three adjacent edges at $\Phi(\mathbf{u}_i)$ are $D_u\Phi(\mathbf{u}_i)\Delta u_i, D_v\Phi(\mathbf{u}_i)\Delta v_i, D_w\Phi(\mathbf{u}_i)\Delta w_i$, respectively. The volume of this parallelepiped is

$$\begin{aligned} &|(D_u\Phi(\mathbf{u}_i) \times D_v\Phi(\mathbf{u}_i)) \cdot D_w\Phi(\mathbf{u}_i)||\Delta u_i \Delta v_i \Delta w_i| \\ =&|\det[D_u\Phi(\mathbf{u}_i), D_v\Phi(\mathbf{u}_i), D_w\Phi(\mathbf{u}_i)]|\text{Volume } (R) \\ =&|\det(D_{\mathbf{u}}\Phi(\mathbf{u}_i))|\text{Volume } (R). \end{aligned}$$

This explains why we need to have the Jacobian factor $|\det(D_{\mathbf{u}}\Phi(\mathbf{u}_i))|$ in changing variables in the integration:

$$\iiint_D f(\mathbf{x})\, d^3\mathbf{x} = \iiint_{\hat{D}} f(\Phi(\mathbf{u}))|\det(D_{\mathbf{u}}\Phi(\mathbf{u}_i))|\, d^3\mathbf{u}.$$

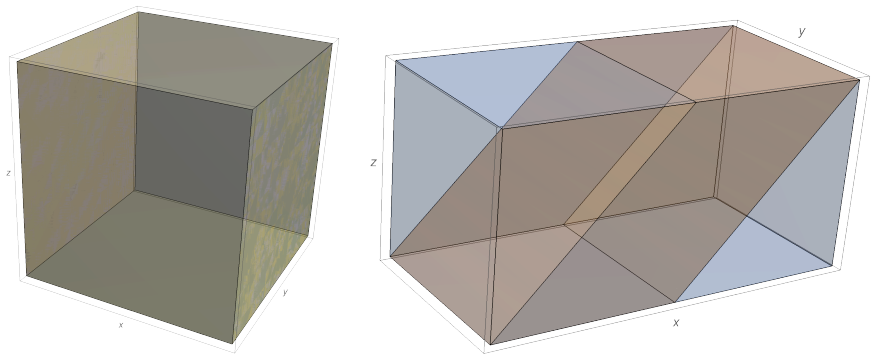


Figure 7.8: When a cube is mapped to a parallelepiped by a linear map $\mathbf{x} \mapsto A\mathbf{x}$, its volume is magnified by $|\det A|$.

Example 7.3.2

Suppose we use spherical coordinates (r, ϕ, θ) to evaluate some integral in rectangular coordinates (x, y, z), where

$$\begin{cases} x = r\sin\phi\cos\theta \\ y = r\sin\phi\sin\theta \\ z = r\cos\phi \end{cases}$$

Then, denoting this maps as Φ, we have

$$\begin{cases} \Phi_r = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi) \\ \Phi_\phi = r(\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi) \\ \Phi_\theta = r(-\sin\phi\sin\theta, \sin\phi\cos\theta, 0) \\ \Phi_\phi \times \Phi_\theta = r^2\sin\phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi) \end{cases}$$

So

$$(\Phi_\phi \times \Phi_\theta) \cdot \Phi_r = \det[\Phi_\phi, \Phi_\theta, \Phi_r] = r^2\sin\phi.$$

For 3×3 matrices, you may take the first equality as the definition for the determinant of the matrix. Using the symmetry properties of the mixed product on the left hand side, we see that $\det[\Phi_\phi, \Phi_\theta, \Phi_r] = \det[\Phi_r, \Phi_\phi, \Phi_\theta] = \det[D_{r,\phi,\theta}\Phi]$. So when an integral in the rectangular coordinates x, y, z is transformed into one in r, ϕ, θ, the integrand needs to be multiplied by the Jacobian factor $r^2\sin\phi$[a]

and

$$\begin{aligned}&\iiint_D f(x,y,z)\,dx\,dy\,dz\\ =&\iiint_{\hat{D}} f(r\sin\phi\cos\theta, r\sin\phi\sin\theta, r\cos\phi) r^2\sin\phi\,dr\,d\phi\,d\theta.\end{aligned}$$

In our earlier example computing $\iiint_{x^2+y^2+z^2\le R^2}(x^2+y^2)\,dx\,dy\,dz$, if we use spherical coordinates (r,ϕ,θ), then the solid ball $x^2+y^2+z^2\le R^2$ is mapped from $\hat{D}=\{(r,\phi,\theta): 0\le r\le R, 0\le\phi\le\pi, 0\le\theta\le 2\pi\}$, so

$$\begin{aligned}&\iiint_{x^2+y^2+z^2\le R^2}\left(x^2+y^2\right)\,dx\,dy\,dz\\ =&\int_0^{2\pi}\int_0^{\pi}\int_0^{R}\left(r^2\sin^2\phi\right)\,r^2\sin\phi\,dr\,d\phi\,d\theta\\ =&\int_0^{2\pi}\int_0^{\pi}\frac{R^5}{5}\sin^3\phi\,d\phi\,d\theta\\ =&\frac{8\pi R^5}{15}.\end{aligned}$$

[a] Some texts use ρ in place of r in spherical coordinates; other texts may swap the symbols between θ and ϕ. It is important to note that the ϕ in the $r^2\sin\phi$ factor refers to the angle with respect to the north pole.

Remark 7.3.1

In computing the Jacobian matrix from the spherical coordinates (r,ϕ,θ) to the rectangular coordinates, note that $\Phi_r,\Phi_\phi,\Phi_\theta$ are orthogonal to each other, and $||\Phi_r||=1, ||\Phi_\phi||=r, ||\Phi_\theta||=r\sin\phi$. This has a clear geometric interpretation: as r moves at a unit speed, (x,y,z) moves in the radial direction at a unit speed, so $||\Phi_r||=1$; as ϕ moves at a unit speed, (x,y,z) moves along a circle of radius r at unit angular speed, so $||\Phi_\phi||=r$; as θ moves at a unit speed, (x,y,z) moves along a circle of radius $r\sin\phi$ at unit angular speed, so $||\Phi_\theta||=r\sin\phi$; and the three velocity vectors $\Phi_r,\Phi_\phi,\Phi_\theta$ are orthogonal to each other based on the motion of the point in relation to that of r,ϕ,θ.

The above observation can be used to produce a set of three orthonormal vectors $\{\Phi_r, r^{-1}\Phi_\phi, (r\sin\phi)^{-1}\Phi_\theta\}$, illustrated as $\{\hat{\mathbf{r}},\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}\}$ in the figure above, which

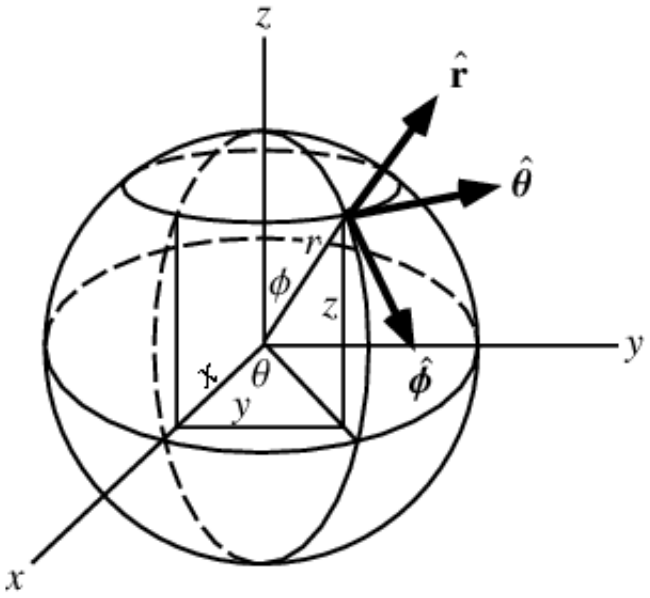


Figure 7.9: Relation between rectangular and spherical coordinates

then forms an orthogonal matrix, and

$$[\Phi_r, \Phi_\phi, \Phi_\theta] = Q \begin{bmatrix} 1 & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r\sin\phi \end{bmatrix}.$$

One basic property of determinant is that $\det(QR) = \det(Q)\det(R)$ *for any two square matrices* Q, R*, and* $\det(Q) = \det(Q^T)$*. It then follows that, if* Q *is an orthogonal matrix, using* $I = Q^TQ$*, we see that*

$$1 = \det(I) = \det(Q^TQ) = \det(Q^T)\det(Q) = \det(Q)\det(Q),$$

so $\det(Q) = \pm 1$*. In our situation here, it turns out that* $\det(Q) = 1$*, and*

$$\det[\Phi_r, \Phi_\phi, \Phi_\theta] = \det(Q)\det\begin{bmatrix} 1 & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r\sin\phi \end{bmatrix} = r^2\sin\phi.$$

Reading Quizzes/Questions:

(i) Determine the volume of the parallelepiped with vectors $(1,0,0), (2,2,0), (-5,6,3)$

as its three adjacent edges.

(ii) Determine the Jacobian matrix and its determinant for the change of variables from cylindrical coordinates (r, θ, z) to (x, y, z), where $x = r\cos\theta, y = r\sin\theta$.

7.4 Integration on parameterized surfaces

7.4.1 Parameterized surfaces

7.4.2 The surface area of a parameterized surface

The definition of the surface area of a parameterized surface S is motivated again by partition of the surface as the non-overlapping union of images of small rectangular boxes which form a partition of the domain D, each of such an image of a small box with vertex at (u, v) in D and box side length h is approximated by a parallelogram with a vertex at $X(u, v)$ and with sides $X_u(u, v)h$ and $X_v(u, v)h$. Thus the magnifying factor is $\|X_u(u, v) \times X_v(u, v)\|$, which leads us to define the surface area as

$$\iint_S 1\, dS = \iint_D \|X_u(u, v) \times X_v(u, v)\|\, du\, dv,$$

and the integral of a function $f(X)$ on S as

$$\iint_S f(\mathbf{x})\, dS = \iint_D f(X(u, v))\|X_u(u, v) \times X_v(u, v)\|\, du\, dv.$$

The most useful cases for computations include

(I). (Surface as a graph $z = h(x, y)$ over a two dimensional domain D) $(x, y) \mapsto X(x, y) = (x, y, h(x, y))$ provides a parametrization. $X_x = (1, 0, h_x(x, y))$, $X_y = (0, 1, h_y(x, y))$, and $X_x \times X_y = (-h_x(x, y), -h_y(x, y), 1)$. So

$$\|X_x \times X_y\| = \sqrt{1 + h_x(x, y)^2 + h_y(x, y)^2} = \sqrt{1 + \|\nabla h(x, y)\|^2},$$

and

$$\iint_S f(\mathbf{x})\, dS = \iint_D f(x, y, h(x, y))\sqrt{1 + \|\nabla h(x, y)\|^2}\, dx\, dy.$$

The simplest case of such a surface is a plane $z = ax + by + c$ defined over $(x, y) \in D \subset \mathbb{R}^2$. Then $\sqrt{1 + \|\nabla h(x, y)\|^2} = \sqrt{1 + a^2 + b^2}$, so the area of this graph is $\iint_D \sqrt{1 + a^2 + b^2}\, dx\, dy = \sqrt{1 + a^2 + b^2}$(Area of D).

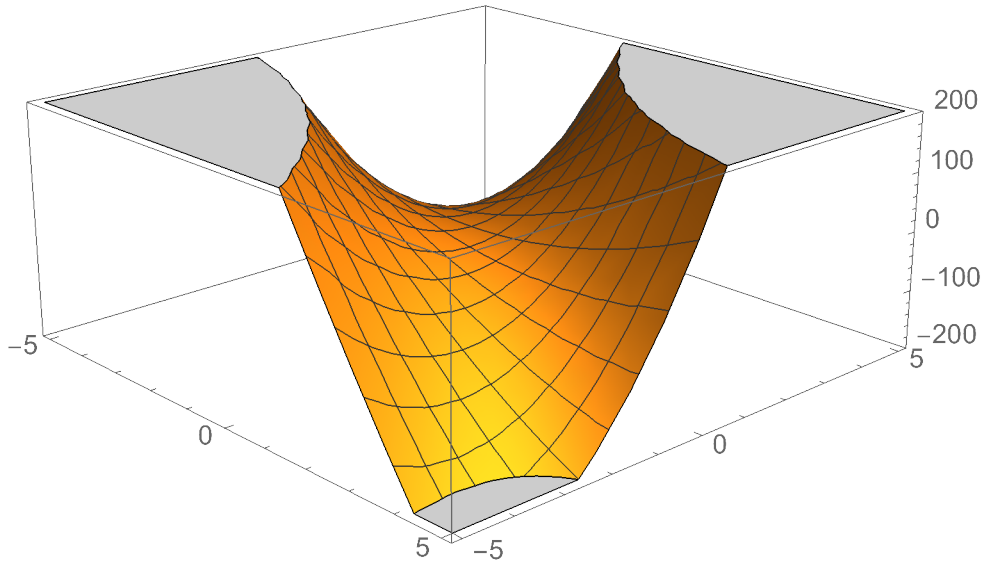


Figure 7.10: A parametric surface showing the partition by images of the partitioned rectangles in the parameter domain.

Example 7.4.1

The surface area of the sphere $x^2 + y^2 + z^2 = R^2$ is twice of the area of the upper hemi-sphere, which is given as the graph of $z = \sqrt{R^2 - x^2 - y^2}$ over the disc $\{(x, y) : x^2 + y^2 \leq R^2\}$. According to our set up, we need to compute

$$\|\nabla z\|^2 = \frac{x^2}{R^2 - x^2 - y^2} + \frac{y^2}{R^2 - x^2 - y^2} = \frac{x^2 + y^2}{R^2 - x^2 - y^2},$$

and get

$$\begin{aligned}\text{Area of sphere of radius } R &= 2\iint_{x^2+y^2\le R^2}\sqrt{1+\frac{x^2+y^2}{R^2-x^2-y^2}}\,dx\,dy\\ &= 2\iint_{x^2+y^2\le R^2}\frac{R}{\sqrt{R^2-x^2-y^2}}\,dx\,dy\\ &= 2\int_0^{2\pi}\int_0^R\frac{R}{\sqrt{R^2-r^2}}\,r\,dr\,d\theta\\ &= 2\int_0^{2\pi}-R\sqrt{R^2-r^2}\Big|_{r=0}^{r=R}\,d\theta\\ &= 4\pi R^2.\end{aligned}$$

(II). For the purpose of integrating a function on a surface, what matters is to find $\|X_u(u,v)\times X_v(u,v)\|$, the vector $X_u(u,v)\times X_v(u,v)$ is not needed explicitly. Note that $\|X_u(u,v)\times X_v(u,v)\| = \sqrt{\|X_u(u,v)\|^2\|X_v(u,v)\|^2-(X_u(u,v)\cdot X_v(u,v))^2}$. This follows from

$$\|X_u(u,v)\times X_v(u,v)\| = \|X_u(u,v)\|\|X_v(u,v)\|\sin\theta,$$

where θ is the angle between $X_u(u,v)$ and $X_v(u,v)$; and

$$\|X_u(u,v)\|^2\|X_v(u,v)\|^2-(X_u(u,v)\cdot X_v(u,v))^2 = \|X_u(u,v)\|^2\|X_v(u,v)\|^2(1-\cos^2\theta).$$

The latter is often easier to compute. Thus we can also compute $\iint_D f(X(u,v))\|X_u(u,v)\times X_v(u,v)\|\,du\,dv$ as

$$\iint_D f(X(u,v))\sqrt{\|X_u(u,v)\|^2\|X_v(u,v)\|^2-(X_u(u,v)\cdot X_v(u,v))}\,du\,dv.$$

Remark 7.4.1

Note that

$$\begin{aligned}&\|X_u(u,v)\|^2\|X_v(u,v)\|^2-(X_u(u,v)\cdot X_v(u,v))^2\\ =&\det\begin{bmatrix}X_u(u,v)\cdot X_u(u,v) & X_u(u,v)\cdot X_v(u,v)\\ X_v(u,v)\cdot X_u(u,v) & X_v(u,v)\cdot X_v(u,v)\end{bmatrix}\\ =&\det\left(\begin{bmatrix}X_u(u,v) & X_v(u,v)\end{bmatrix}^T\begin{bmatrix}X_u(u,v) & X_v(u,v)\end{bmatrix}\right)\end{aligned}$$

where

$$\begin{bmatrix} X_u(u,v) & X_v(u,v) \end{bmatrix}$$

is the Jacobian matrix of the parametrization map $(u,v) \mapsto X(u,v)$. *This formulation works for more general settings such as in higher dimensions.*

Example 7.4.2

We use the spherical coordinates to parametrize the sphere of radius R centered at $(0,0,0)$: $x = R\sin\phi\cos\theta, y = R\sin\phi\sin\theta, z = R\cos\phi$, where $0 \le \phi \le \pi, 0 \le \theta \le 2\pi$. Denoting this parametrization by $X(\phi,\theta)$, then $X_\phi = R(\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi)$, $X_\theta = R(-\sin\phi\sin\theta, \sin\phi\cos\theta, 0)$. $X_\phi \cdot X_\theta = 0$, $\|X_\phi\| = R$, and $\|X_\theta\| = R\sin\phi$, so

$$\begin{aligned} \text{Area of sphere of radius } R &= \iint_{0\le\phi\le\pi, 0\le\theta\le 2\pi} 1\, R^2\sin\phi\, d\phi\, d\theta \\ &= \int_0^{2\pi}\int_0^{\pi} R^2\sin\phi\, d\phi\, d\theta \\ &= \int_0^{2\pi} 2R^2 d\theta \\ &= 4\pi R^2. \end{aligned}$$

We could also use cylindrical coordinates (r,θ,z) to parametrize the sphere of radius R centered at $(0,0,0)$: $x = r\cos\theta, y = r\sin\theta, z = z$, where the equation $x^2+y^2+z^2 = R^2$ turns into $r^2+z^2 = R^2$. Geometrically, we should take $0 \le r \le R$ and $-R \le z \le R$, and each r corresponds to two possible values of $z : \pm\sqrt{R^2-r^2}$. It's easier to treat r as a function of z: $r = \sqrt{R^2-z^2}$. for $-R \le z \le R$. Thus we have the parametrization

$$\mathbf{X}(z,\theta) = \begin{bmatrix} \sqrt{R^2-z^2}\cos\theta \\ \sqrt{R^2-z^2}\sin\theta \\ z \end{bmatrix},$$

so

$$\mathbf{X}_z(z,\theta) = \begin{bmatrix} -\frac{z}{\sqrt{R^2-z^2}}\cos\theta \\ -\frac{z}{\sqrt{R^2-z^2}}\sin\theta \\ 1 \end{bmatrix}, \quad \mathbf{X}_\theta(z,\theta) = \begin{bmatrix} -\sqrt{R^2-z^2}\sin\theta \\ \sqrt{R^2-z^2}\cos\theta \\ 0 \end{bmatrix}.$$

It follows that

$$\|\mathbf{X}_z(z,\theta)\| = \frac{R}{\sqrt{R^2-z^2}}, \quad \|\mathbf{X}_\theta(z,\theta)\| = \sqrt{R^2-z^2}.$$

Since $\mathbf{X}_z(z,\theta)\cdot\mathbf{X}_\theta(z,\theta)=0$, it follows that

$$\|\mathbf{X}_z(z,\theta)\times\mathbf{X}_\theta(z,\theta)\| = \|\mathbf{X}_z(z,\theta)\|\|\mathbf{X}_\theta(z,\theta)\| = R.$$

Thus

$$\text{Area of sphere of radius } R = \int_0^{2\pi}\int_{-R}^{R} R\,dz\,d\theta = 4\pi R^2.$$

This can be used to prove Archimedes' Theorem, for the area of the section of the sphere of radius R between $z=z_1$ and $z=z_2$ is

$$\int_0^{2\pi}\int_{z_1}^{z_2} R\,dz\,d\theta = 2\pi R(z_2-z_1),$$

which is the area of the circumscribed cylinder of radius R between $z=z_1$ and $z=z_2$.

Remark 7.4.2

In Carlen's Example 117 and 118, he chooses to work with cylindrical coordinate parametrization, and needs to compute $\mathbf{X}_r\times\mathbf{X}_\theta$ for the specific surface. This requires a good amount of computation. It is easier to work with rectangular coordinates to set up the integral as

$$\int f(x,y,h(x,y))\sqrt{1+\|\nabla h(x,y)\|^2}\,dx\,dy$$

then change the rectangular integral into polar coordinates for such round domains of integration. In doing this, $dx\,dy = r\,dr\,d\theta$ is the standard change of variables, and the computation for $\sqrt{1+\|\nabla h(x,y)\|^2}$ is fairly routine.

In general, when a surface is represented as a graph $z=h(x,y)$, we know $X_x\times X_y = (-h_x,-h_y,1)$. If we decide to use a new parametrization $\hat{X}(u,v)$, it means that $x=\phi(u,v), y=\psi(u,v)$ for some (differentiable)

ϕ, ψ such that $\hat{X}(u,v) = X(\phi(u,v), \psi(u,v))$. Then by the chain rule

$$\hat{X}_u \times \hat{X}_v = X_x \times X_y \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}.$$

Since we only need $\|\hat{X}_u \times \hat{X}_v\|$ in this context, we find

$$\|\hat{X}_u \times \hat{X}_v\| = \|X_x \times X_y\| |\det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}|,$$

and the two factors $\|X_x \times X_y\| = \sqrt{1 + h_x^2 + h_y^2}$ and $\det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$ are often easier to compute than computing $\|\hat{X}_u \times \hat{X}_v\|$ directly.
For the case of Example 117, $h(x,y) = 1 - x^2 - y^2$, so $\sqrt{1 + \|\nabla h(x,y)\|^2} = \sqrt{1 + 4x^2 + 4y^2}$.
For the lower portion of the surface in Example 118, $h(x,y) = \frac{1}{\sqrt{x^2+y^2}}$, and $\sqrt{1 + \|\nabla h(x,y)\|^2} = \sqrt{1 + r^{-4}}$. In both cases, the integral can be evaluated easily using polar coordinates.

(III). (Surface of revolution) Such a surface is defined through a curve $t \in (a,b) \mapsto (y,z) = (\phi(t), \psi(t))$, and rotating each point $(\phi(t), \psi(t))$ around the z-axis. This means that we will assume $\phi(t) > 0$ and will use it as the radius of rotation at that point, so $X(t,\theta) := (x,y,z) = (\phi(t)\cos\theta, \phi(t)\sin\theta, \psi(t))$, $t \in (a,b)$, $\theta \in [0, 2\pi)$ provides a parametrization for this surface of revolution. Then

$$X_t = (\phi'(t)\cos\theta, \phi'(t)\sin\theta, \psi'(t)), \quad X_\theta = (-\phi(t)\sin\theta, \phi(t)\cos\theta, 0).$$

$\|X_t\|^2 = |\phi'(t)|^2 + |\psi'(t)|^2$, $\|X_\theta\|^2 = |\phi(t)|^2$, and $X_t \cdot X_\theta = 0$, so its area is given by

$$\int_a^b \int_0^{2\pi} \sqrt{|\phi'(t)|^2 + |\psi'(t)|^2} |\phi(t)| \, d\theta \, dt = 2\pi \int_a^b \sqrt{|\phi'(t)|^2 + |\psi'(t)|^2} |\phi(t)| \, dt.$$

Special cases include $\phi(t) = t$, namely, we can identify t to be the radial variable in the cylindrical coordinates; we then label $\psi(r)$ as $z = z(r)$, thus the area is

$$2\pi \int_a^b \sqrt{1 + |z'(r)|^2} \, r \, dr.$$

Recall that $\sqrt{1+|z'(r)|^2}\,dr = ds$, where s is the arc length parameter of the curve $z = z(r)$, so the area can also be written as

$$2\pi \int r\, ds,$$

which, geometrically, means that it is the integration of cylindrical slices with width ds and radius r.

Another such case is $t = z$, $\psi(t) = t$, so we can treat $r = \phi(z)$ as the radial variable in cylindrical coordinates. Then the area is $2\pi \int_a^b \sqrt{1+|r'(z)|^2}\; r\; dz$. Note that $\sqrt{1+|r'(z)|^2}\; dz$ still equals ds, where s is the arc length parameter.

In the special case of the sphere $x^2+y^2+z^2 = R^2$, which can be written in terms of cylindrical coordinates as $r^2+z^2 = R^2$, we get by implicit differentiation that $2rr'(z) + 2z = 0$, so $r'(z) = -z/r$, and $\sqrt{1+|r'(z)|^2} = \sqrt{1+z^2/r^2} = R/r$, so the area of this sphere between $z = z_1 < z = z_2$ is $2\pi \int_{z_1}^{z_2} R\; dz = 2\pi R(z_2 - z_1)$, which is the area of the circumscribed cylinder between $z = z_1 < z = z_2$. This was originally due to Archimedes.

In the case of a section of a cone given by $z^2 = k^2(x^2+y^2)$, $z_1 \le z \le z_2$, we can parametrize it in terms of cylindrical coordinates $(z,\theta) \mapsto (k^{-1}z\cos\theta, k^{-1}z\sin\theta, z)$, and according to our discussion, the surface area is

$$2\pi \int_{z_1}^{z_2} r\sqrt{1+\left(\frac{dr}{dz}\right)^2}\, dz = 2\pi \int_{z_1}^{z_2} k^{-1}\sqrt{1+k^{-2}}z\, dz = \pi(z_2^2 - z_1^2)k^{-1}\sqrt{1+k^{-2}}.$$

If we express this area in terms of the radius $r_i = z_i/k$ of the two cross sectional disks, then the area is $\pi(r_2^2 - r_1^2)\sqrt{k^2+1}$.

(IV). Often a geometric surface S in $\mathbb{R}^3$ does not come with a canonical choice of parametrization, so one may choose different parametrizations to represent the surface and compute its area or integrals on it. A natural question is whether our way of defining the area of a surface and integrals on it depends on how we choose to parametrize it? The answer is no. To verify this statement, suppose $(u,v) \in D \mapsto X(u,v) \in S$ is an initial (differentiable) parametrization, and $(\hat{u},\hat{v}) \in \widehat{D} \mapsto (u,v) \in \Phi(\hat{u},\hat{v}) \in D$ is a change of variables from $(\hat{u},\hat{v}) \in \widehat{D}$ to $(u,v) \in D$, so that $\widehat{X}(\hat{u},\hat{v}) := X \circ \Phi(\hat{u},\hat{v})$ is another parametrization of the surface S. Then we now have two ways of computing the area of S:

$$\iint_D \|X_u \times X_v\|\, du\, dv \quad \text{and} \quad \iint_{\widehat{D}} \|\widehat{X}_{\hat{u}} \times \widehat{X}_{\hat{v}}\|\, d\hat{u}\, d\hat{v}.$$

According to the chain rule of differentiation we have

$$\begin{cases} \widehat{X}_{\hat{u}} = X_u \dfrac{\partial u}{\partial \hat{u}} + X_v \dfrac{\partial v}{\partial \hat{u}}, \\ \widehat{X}_{\hat{v}} = X_u \dfrac{\partial u}{\partial \hat{v}} + X_v \dfrac{\partial v}{\partial \hat{v}}, \end{cases}$$

so

$$\widehat{X}_{\hat{u}} \times \widehat{X}_{\hat{v}} = \left(X_u \frac{\partial u}{\partial \hat{u}} + X_v \frac{\partial v}{\partial \hat{u}} \right) \times \left(X_u \frac{\partial u}{\partial \hat{v}} + X_v \frac{\partial v}{\partial \hat{v}} \right) = X_u \times X_v \det \begin{bmatrix} \frac{\partial u}{\partial \hat{u}} & \frac{\partial u}{\partial \hat{v}} \\ \frac{\partial v}{\partial \hat{u}} & \frac{\partial v}{\partial \hat{v}} \end{bmatrix},$$

and

$$\|\widehat{X}_{\hat{u}} \times \widehat{X}_{\hat{v}} = \|X_u \times X_v\| |\det \begin{bmatrix} \frac{\partial u}{\partial \hat{u}} & \frac{\partial u}{\partial \hat{v}} \\ \frac{\partial v}{\partial \hat{u}} & \frac{\partial v}{\partial \hat{v}} \end{bmatrix}| = \|X_u \times X_v\| |\det(D_{\hat{u},\hat{v}}\Phi)|.$$

Therefore, by the change of variables formula for integration in two variables, we have

$$\iint_D \|X_u \times X_v\| \, du \, dv = \iint_{\widehat{D}} \|\widehat{X}_{\hat{u}} \times \widehat{X}_{\hat{v}}\| \, d\hat{u} \, d\hat{v}.$$

(V). A piece of surface often can be represented as a graph over (x, y) and as a graph over (y, z) or (x, z). This often arises when the surface Σ is given implicitly by $F(x, y, z) = 0$. If we assume $F_x(x_0, y_0, z_0) \neq 0$, $F_y(x_0, y_0, z_0) \neq 0$, and $F_z(x_0, y_0, z_0) \neq 0$, then by Implicit Function Theorem in Chapter 5, the set of points satisfying $F(x, y, z) = 0$ near $\mathbf{x}_0 = (x_0, y_0, z_0)$ can be represented as a graph over (x, y) and as a graph over (y, z) as well as a graph over (x, z). Recall that $(F_z(x, y, z), F_y(x, y, z), F_z(x, y, z))$ is a normal vector to the surface Σ, so we can use

$$\begin{aligned} N(x, y, z) &= \frac{(F_z(x, y, z), F_y(x, y, z), F_z(x, y, z))}{\|\nabla F(x, y, z)\|} \\ &= \frac{(F_z(x, y, z), F_y(x, y, z), F_z(x, y, z))}{\sqrt{F_x(x, y, z)^2 + F_y(x, y, z)^2 + F_z(x, y, z)^2}} \end{aligned}$$

as the unit normal vector to Σ at (x, y, z).

If we treat Σ as a graph of z over (x, y), then implicit differentiation gives

$$F_x(x, y, z) + F_z(x, y, z) \frac{\partial z}{\partial x} = 0, \quad F_y(x, y, z) + F_z(x, y, z) \frac{\partial z}{\partial y} = 0,$$

from which it follows that

$$dS = \sqrt{1 + (\frac{\partial z}{\partial x})^2 + (\frac{\partial z}{\partial y})^2} \, dx \, dy = \frac{\sqrt{F_x(x, y, z)^2 + F_y(x, y, z)^2 + F_z(x, y, z)^2}}{|F_z(x, y, z)|} \, dx \, dy.$$

This relation can be rewritten as

$$|N_3(x,y,z)|\ dS = dx\ dy.$$

Geometrically, this means that when a piece of (small) surface with unit normal vector $N = (N_1, N_2, N_3)$ is projected orthogonally onto the x-y plane, then the area of the projected region is $|N_3|$ times the area of the original surface. The same applies to the other coordinate planes, namely,

$$|N_1|\ dS = dy\ dz, \quad \text{and} \quad |N_2|\ dS = dx\ dz.$$

When N is a constant vector, namely, when Σ is a piece on a plane in $\mathbb{R}^3$, let Σ_{xy} denotes the orthogonal projection of Σ onto the x-y plane, Σ_{xz} denotes the orthogonal projection of Σ onto the x-z plane, and Σ_{yz} denotes the orthogonal projection of Σ onto the y-z plane. Let $|\Sigma_*|$ denote the area of Σ_*, then using $N_1^2 + N_2^2 + N_3^2 = 1$, we have

$$|\Sigma|^2 = |\Sigma_{xy}|^2 + |\Sigma_{yz}|^2 + |\Sigma_{xz}|^2.$$

In Chapter 9, we will take into account the orientation of the surface, and work with $(N_1(x,y,z), N_2(x,y,z), N_3(x,y,z))dS = (\pm dx\ dy, \pm dy\ dz, \pm dz\ dx)$, by removing the absolute value signs on $N_i(x,y,z)$ when computing the flux integral $F(x,y,z) \cdot (N_1(x,y,z), N_2(x,y,z), N_3(x,y,z))dS$.

Reading Quizzes/Questions:

1. Suppose Γ is a closed curve in $\mathbb{R}^2$ parametrized by $x = \phi(t), y = \psi(t)$ for $t \in [0, l]$. We define a right cylinder $\mathcal{C}$ over Γ by

$$\mathbf{X}(t,z) = (x,y,z) = (\phi(t), \psi(t), z), 0 \le t \le l, a \le z \le b.$$

   (i) Compute $\partial_t \mathbf{X} \times \partial_z \mathbf{X}$ and use it to determine a unit normal vector $\mathbf{n}(\mathbf{X})$ to $\mathcal{C}$ at $\mathbf{X}(t,z)$.

   (ii) Is there a good way to determine whether your unit normal is outward pointing or inward pointing?

   (iii) Compute the area of this surface.

2. Suppose we have two differentiable functions $h_1(x,y) \le h_2(x,y)$ defined over a bounded domain D in $\mathbb{R}^2$. The graphs of these two functions, together with the right cylinder over the boundary ∂D as defined in the previous problem, enclose a three dimensional solid. For any point $(x, y, h_i(x,y))$ on either the top or the bottom portion of the boundary of this solid, determine an outward pointing unit normal to the surface there.

3. Suppose that $\mathbf{X}(u,v)$ for $(u,v) \in D$ is a parametrization for a surface $\mathcal{S}$, that $\mathbf{n}(\mathbf{X})$ denotes the unit normal to $\mathcal{S}$ in the direction of $\partial_u \mathbf{X} \times \partial_v \mathbf{X}$. Suppose that we are given a vector valued function $F(\mathbf{X}) = (F_1(\mathbf{X}), F_2(\mathbf{X}), F_3(\mathbf{X}))$ for $\mathbf{X} \in \mathcal{S}$, and want to use $F(\mathbf{X}) \cdot \mathbf{n}(\mathbf{X})$ as the integrand to compute $\int_{\mathcal{S}} F(\mathbf{X}) \cdot \mathbf{n}(\mathbf{X})\, dS$. Verify that

$$\int_{\mathcal{S}} F(\mathbf{X}) \cdot \mathbf{n}(\mathbf{X})\, dS = \int_D F(\mathbf{X}) \cdot (\partial_u \mathbf{X} \times \partial_v \mathbf{X})\; dA$$

where dA refers to integration in the (u,v) variables.

4. Using the set up of the previous problem, suppose $\mathcal{S}$ is given by a graph $z = h(x,y)$ for $(x,y) \in D$, $\mathbf{n}(\mathbf{X})$ is upward pointing, and $F(x,y,z) = (0,0,F_3(x,y,z))$ is a vector field oriented vertically, verify that

$$\int_{\mathcal{S}} F(\mathbf{X}) \cdot \mathbf{n}(\mathbf{X})\, dS = \int_D F_3(x,y,h(x,y))\, dA$$

where dA refers to integration in the (x,y) variables.

Chapter 9

FLUX AND CIRCULATION, DIVERGENCE AND CURL

For this chapter, due to our time constraint, we won't have time to follow Professor Carlen's notes to fully develop the material as he does; we will use several sections from Chapter 16 of the Rogawski and Adam Calculus textbook to quickly introduce the basic concepts and relevant computations. I have uploaded a scanned copy of the relevant sections under the Sakai Resource Tab. I will also provide some comments to facilitate students' reading of Professor Carlen's notes.

9.1 Flows and flux

Here are some main concepts.

- A **vector field** X over a domain D is a vector-valued function defined on D, taking values in $\mathbb{R}^n$, where $n = 2$ if D is two dimensions, and $n = 3$ if D is three dimensional. Geometrically, it assigns a vector $X(\mathbf{x})$ at each point $\mathbf{x} \in D$. Typically we want $X(\mathbf{x})$ to be continuous, or continuously differentiable; or at least piecewise continuous or continuously differentiable. The latter means that we can partition D into the union of several (non-overlapping) domains $\cup D_i$, such that the restriction to the interior of each is continuous.

- To define the **flux** of a (continuous) vector field X in a two dimensional domain D over a curve $\mathcal{C}$ in D, we typically work with a well defined and continuous unit normal vector $\mathbf{n}(\mathbf{x})$ along the $\mathcal{C}$. If we parametrize the curve $\mathcal{C}$ by its arc length parameter s, $0 \leq s \leq l$, then the flux of X across $\mathcal{C}$, with this designated

orientation is defined as

$$\int_0^l X(\mathbf{x}(s)) \cdot \mathbf{n}(\mathbf{x}(s))\ ds.$$

It is also denoted as $\int_{\mathcal{C}} X(\mathbf{x}(s)) \cdot \mathbf{n}(\mathbf{x}(s))\ ds$. It is important to keep in mind that this integral is independent of how the curve is parametrized, as long as the parametrization gives a unit normal vector field that agrees with the designated orientation.

The curve is usually not given in terms of its arc length parameter, but is given in terms of a general parametrization: $\mathbf{x} = \mathbf{r}(t)$ for $a \leq t \leq b$ for some $a < b$ (in some situations one may be working with a parametrization where $a > b$). To compute the flux, we note that $\mathbf{r}'(t) = (x'(t), y'(t))$ is tangent to the curve at $\mathbf{r}(t)$, and $\mathbf{r}'(t)^{\perp} = (-y'(t), x'(t))$ is a normal to the curve at $\mathbf{r}(t)$, so $\mathbf{n}(\mathbf{x}(s))$ is either $\mathbf{r}'(t)^{\perp}/\|\mathbf{r}'(t)\|$ or its opposite. Note that we would take the plus sign if $\mathbf{n}(\mathbf{x}(s))$ relates to $\mathbf{r}'(t)$ by a counterclockwise rotation, and take the negative sign otherwise.

Using $ds = \|\mathbf{r}'(t)\|\ dt$, we conclude that the flux is

$$\pm \int_a^b X(\mathbf{r}(t)) \cdot (-y'(t), x'(t))\ dt,$$

where the sign is determined by whether $\mathbf{n}(\mathbf{x}(s)) = \pm \mathbf{r}'(t)^{\perp}/\|\mathbf{r}'(t)\|$.

If we write out $X(\mathbf{r}(t))$ in terms of its components $(X_1(\mathbf{r}(t)), X_2(\mathbf{r}(t)))$, and take note that $x'(t)\ dt = dx$, $y'(t)\ dt = dy$, we have the flux equal to

$$\pm \int_{\mathcal{C}} X_2(x, y)dx - X_1(x, y)dy.$$

This notation also indicates that the integral is independent of how the curve is parametrized, and has the advantage that if the curve can be parametrized as a graph of the form $y = h(x)$, $a \leq x \leq b$, then $dy = h'(x)\ dx$, and the flux reduces to a one variable integral over $[a, b]$:

$$\pm \int_a^b \left[X_2(x, h(x)) - X_1(x, h(x))h'(x)\right]\ dx,$$

while if the curve can be parametrized as a graph of the form $x = g(y)$, $c \leq y \leq d$, then $dx = g'(y)\ dy$, and the flux reduces to

$$\pm \int_c^d \left[X_2(g(y), y)g'(y) - X_1(g(y), y)\right]\ dy.$$

- A more commonly used line integral of a (continuous) vector field X along an oriented curve $\mathcal{C}$ with continuously varying unit tangent vectors, parametrized as $\mathbf{x} = \mathbf{r}(t)$ over $a \le t \le b$, is the vector line integral (circulation in the case when the curve is a closed one) of X along $\mathcal{C}$ defined as

$$\int_{\mathcal{C}} X(\mathbf{r}(t)) \cdot d\mathbf{r}'(t) = \int_a^b X(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\ dt.$$

  Here if we set $\mathbf{T}(r(t)) = \mathbf{r}'(t)/\|\mathbf{r}'(t)\|$ to be the unit tangent vector to the curve at $\mathbf{r}(t)$ in the given orientation, then $\mathbf{r}'(t)\ dt = \mathbf{T}(r(t))\|\mathbf{r}'(t)\|\ dt = \mathbf{T}(r(t))\ ds$, with s being the arc length parameter. Therefore, using s to parametrize the curve: $\gamma(s) = \mathbf{r}(t)$, we have the line integral equal to $\int_0^l X(\gamma(s)) \cdot \mathbf{T}(\gamma(s))\ ds$, where l is the arc length of the curve from $\mathbf{r}(a)$ to $\mathbf{r}(b)$. A more conceptual notation for this integral is $\int_{\mathcal{C}} X(\gamma(s)) \cdot \mathbf{T}(\gamma(s))\ ds$, which indicates that the integral is independent of how the curve is parametrized as long as its tangent vector agrees with the designated orientation.

  Our way of defining this vector line integral/circulation makes it ordinary integral in the single variable t over $[a, b]$. For the case that $\mathcal{C}$ is a curve in $\mathbb{R}^2$, using $d\mathbf{r}'(t) = \mathbf{r}'(t)\ dt = (x'(t), y'(t))dt = (dx, dy)$, the integral can be written as

$$\int_{\mathcal{C}} X(\mathbf{r}(t)) \cdot d\mathbf{r}'(t) = \int_a^b X(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\ dt = \int_{\mathcal{C}} X_1(x, y)\ dx + X_2(x, y)\ dy,$$

  where the latter notation suggests that it is independent of how we parametrize the curve—as long as the parametrization gives the set orientation. Again in actual computations, we have can evaluate this integral in terms of an ordinary integral over an x interval or an y interval, if the curve can be parametrized as a graph over an x interval or an y interval.

  This integral depends on the orientation of the curve $\mathcal{C}$. If $\mathbf{x} = \mathbf{r}(t)$ is a parametrization of $\mathcal{C}$ in its given orientation, then $\mathbf{x} := \mathbf{r}^-(t) := \mathbf{r}(b + a - t)$ is a parametrization for the opposite orientation of $\mathcal{C}$, which is denoted as $-\mathcal{C}$. Note that

$$\int_{-\mathcal{C}} X_1(x, y)\ dx + X_2(x, y)\ dy = -\int_{\mathcal{C}} X_1(x, y)\ dx + X_2(x, y)\ dy.$$

- When the vector field is conservative, namely, $X(\mathbf{x}) = \nabla f(\mathbf{x})$ for some differentiable function $f(\mathbf{x})$, then

$$X(\mathbf{r}'(t)) \cdot \mathbf{r}'(t)\ dt = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = \frac{df(\mathbf{r}(t))}{dt},$$

by the chain rule, so

$$\int_{\mathcal{C}} X(\mathbf{r}(t)) \cdot d\mathbf{r}'(t) = \int_a^b X(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\ dt = f(\mathbf{r}(b)) - f(\mathbf{r}(a)),$$

which depends only on $(\mathbf{r}(b)) - f(\mathbf{r}(a))$, not on the particular path from $\mathbf{r}(a)$ to $\mathbf{r}(b)$. In particular, when $\mathcal{C}$ is a closed curve, we see that for a conservative vector field,

$$\int_{\mathcal{C}} X(\mathbf{r}(t)) \cdot d\mathbf{r}'(t) = 0.$$

- When a curve $\mathcal{C}$ does not have continuously varying unit normal vectors, as in the case of polygonal segments, it is possible to partition $\mathcal{C}$ into the non-overlapping union of several segments $\cup\mathcal{C}_i$, $i = 1, 2, \cdots, k$, such that the end point P_i of $\mathcal{C}_i$ coincides with the beginning point of $\mathcal{C}_{i+1}$ — this prescribes how the orientation of $\mathcal{C}_i$ relates to that of $\mathcal{C}_{i+1}$, even though there may be a discontinuity of the tangent at P_i. Then we define the line integral of X along $\mathcal{C}$ as

$$\sum_{i=1}^{k} \int_{\mathcal{C}_i} X_1(x, y)\ dx + X_2(x, y)\ dy.$$

Exercise 9.1.1. *Evaluate the flux $\int_{\mathcal{C}} X(x, y) \cdot \mathbf{n}(x, y) ds$ and the circulation $\int_{\mathcal{C}} X(x, y) \cdot \mathbf{T}(x, y) ds$, where $\mathcal{C}$ is the circle of radius 2 centered at the origin, oriented counterclockwise, $X(x, y) = (0, x)$, and the normal $\mathbf{n}(x, y)$ points outward.*

SOLUTION: *We use the parametrization $x = 2\cos t, y = 2\sin t, 0 \leq t \leq 2\pi$. It gives us $(x'(t), y'(t)) = (-2\sin t, 2\cos t) = (-y, x)$, which points counterclockwise, and $dx = -2\sin t dt = -y dt, dy = 2\cos t dt = x dy$. Based on our discussion, $\mathbf{T}(x, y) = (-y, x)/2$, and $\mathbf{T}(x, y) ds = (x'(t), y'(t)) dt = (dx, dy)$ so*

$$\int_{\mathcal{C}} X(x, y) \cdot \mathbf{T}(x, y) ds = \int_{\mathcal{C}} x\ dy.$$

$\mathcal{C}$ is the union of the right half circle $\mathcal{C}_1$ and the left half circle $\mathcal{C}_2$. $x = \sqrt{4 - y^2}$, $-2 \leq y \leq 2$, is a parametrization for $\mathcal{C}_1$, while $x = -\sqrt{4 - y^2}$, $-2 \leq y \leq 2$, is a parametrization for $-\mathcal{C}_2$! So we need to add a negative sign when using this parametrization for the left half circle to get

$$\int_{\mathcal{C}} x\ dy = \int_{-2}^{2} \sqrt{4 - y^2}\ dy - \int_{-2}^{2} (-\sqrt{4 - y^2})\ dy.$$

But the latter is seen to be the area of the disk of radius 2, so the result is 4π.

To compute $\int_{\mathcal{C}} X(x,y)\cdot\mathbf{n}(x,y)ds$, *note that* $\mathbf{n}(x,y)$ *is in the opposite direction from* $(-y'(t),x'(t))$, *if* $(x(t),y(t))$ *is a parametrization for the counterclockwise oriented* $\mathcal{C}$, *so*

$$\int_{\mathcal{C}} X(x,y)\cdot\mathbf{n}(x,y)ds = \int_{\mathcal{C}}(0,x)\cdot(dy,-dx) = \int_{\mathcal{C}}(-x)dx.$$

But $x\ dx = x(t)x'(t)\ dt = \frac{1}{2}(x(t)^2)'\ dt$, *so*

$$\int_{\mathcal{C}}(-x)dx = -\frac{1}{2}(x(t)^2)\Big|_{t=0}^{t=l}.$$

But $\mathcal{C}$ *is a closed curve, so* $x(0)=x(l)$, *so* $\int_{\mathcal{C}} X(x,y)\cdot\mathbf{n}(x,y)ds = 0$. □

Exercise 9.1.2. *Find the circulation of the vector field* $(x, x-2y)$ *along the edges of the triangle with* $(0,0),(1,0),(0,1)$ *as its vertices, oriented counterclockwise.*

SOLUTION: *The edges of the triangle has three segments,* C_1 *running from* $(0,0)$ *to* $(1,0)$, C_2 *running from* $(1,0)$ *to* $(0,1)$, *and* C_3 *running from* $(0,1)$ *to* $(0,0)$.

The parametrization for C_1 *given by* $(x,0), 0\le x\le 1$, *gives the correct orientation, and* $(dx,dy)=(dx,0)$ *in this parametrization, so*

$$\int_{C_1}(x,x-2y)\cdot d\mathbf{r}'(t) = \int_{C_1}(x,x-2y)\cdot(dx,0) = \int_0^1 x dx = 1/2.$$

The parametrization for C_2 *given by* $(x,1-x), 0\le x\le 1$, *gives the opposite orientation of the designated one, and* $(dx,dy)=(dx,-dx)$ *in this parametrization, so*

$$\int_{C_2}(x,x-2y)\cdot(dx,dy) = -\int_0^1 x dx + (x-2y)(-dx) = -\int_0^1 2(1-x)\,dx = -1.$$

The parametrization for C_3 *given by* $(0,y), 0\le y\le 1$, *gives the opposite orientation of the designated one, and* $(dx,dy)=(0,dy)$ *in this parametrization, so*

$$\int_{C_3}(x,x-2y)\cdot(dx,dy) = -\int_0^1(0,-2y)\cdot(0,dy) = -\int_0^1 -2y\,dy = 1.$$

Thus we have

$$\int_{C_1\cup C_2\cup C_3}(x,x-2y)\cdot(dx,dy) = 1/2-1+1 = 1/2.$$

□

Exercise 9.1.3. *Evaluate the vector line integral of the vector field* $(x, -2y, x+z)$ *along the segment from* $(1,0,0)$ *to* $(1,2,3)$.

SOLUTION: *We parametrize the segment as* $(x,y,z) = (1-t)(1,0,0)+t(1,2,3) = (1,2t,3t)$ *for* $0 \le t \le 1$. *Then* $(dx,dy,dz) = (0,2,3)dt$, *so the vector line integral is*

$$\begin{aligned}\int_S x\,dx - 2y\,dy + (x+z)\,dz &= \int_0^1 (x,-2y,x+z)\cdot(0,2,3)\,dt \\ &= \int_0^1 [-4y + 3(x+z)]\,dt \\ &= \int_0^1 [-4(2t) + 3(1+3t)]\,dt = 3\frac{1}{2}.\end{aligned}$$

Note that part of the vector field, $(x,-2y,z)$, *is a conservative vector field, for* $(x,-2y,z)\cdot(dx,dy,dz) = x\,dx - 2y\,dy + z\,dz = d(x^2/2 - y^2 + z^2/2)$, *so*

$$\int_C x\,dx - 2y\,dy + z\,dz = (x^2/2 - y^2 + z^2/2)\Big|_{\textit{initial position}}^{\textit{end position}} = (1/2 - 2^2 + 3^2/2) - (1/2) = 1/2.$$

And the remaining part gives

$$\int_C (0,0,x)\cdot(dx,dy,dz) = \int_C x\,dz = \int_0^1 1\cdot 3dt = 3.$$

Thus

$$\int_C (x,-2y,x+z)\cdot(dx,dy,dz) = \frac{1}{2} + 3 = 3\frac{1}{2}.$$

□

9.2 Flux of a vector field across an oriented surface

To define the flux of a vector field across an oriented surface, we first need to discuss the notion of orientation of a surface. We have an intuitive notion that many surfaces have two sides: any closed surface we see in three dimensions such as a closed sphere or a rectangular box has inside and outside; a graph has up side and down side.

One intuitive way to designate a side is to choose a unit normal vector at each point on the surface to point to the side one is referring to. For instance, if the surface is given by the graph of a (differentiable) function $z = h(x,y)$, then $(-h_x(x,y), -h_y(x,y), 1)$ is its upward pointing normal at $(x,y,h(x,y))$, and $(h_x(x,y), h_y(x,y), -1)$ is its downward pointing normal there.

Many textbooks define an oriented surface as one with a chosen continuously varying unit normal vector at each of its point. A rectangular box does not have a continuously varying unit normal vector at each of its point, but it does have a well defined notion of outside and inside. Here is how we can make this notion precise.

(a). A surface such as a rectangular box can be decomposed as the non-overlapping union of several pieces, such that each piece has a continuously varying unit normal vector at each of its point.

(b). When each piece is partitioned as the non-overlapping union of triangular regions (including curvilinear triangles), the chosen continuously varying unit normal vector field on this piece causes an orientation on the sides of each such triangular region by the right hand rule such that if the thumb points in the direction of the chosen normal direction, the remaining four fingers point to the orientation of the sides of each triangular region. Note that if a side is shared by two triangular regions on the same differentiable piece, then the orientations of this side induced by the chosen normal vector field on the two adjacent triangular regions are opposite of each other, as indicated in the figure below.

(c). When two pieces abut, they do so along a differentiable curve, and if a segment of this curve is shared by two triangular regions on the two neighboring differentiable pieces, then the orientations on this segment induced by the chosen normal vector field on the two triangular regions are also opposite of each other.

With this discussion, each piece has a chosen continuously varying unit normal vector field, and the neighboring pieces have a coherent choice according to (c) above.

Now for a given vector field $F(x, y, z)$ whose domain of definition includes the given oriented surface $\mathcal{S}$, we can define the flux of $F(x, y, z)$ across $\mathcal{S}$ by defining its flux across each of its differentiable piece $\mathcal{S}_i$ as discussed above, with the normal on $\mathcal{S}_i$ chosen according to (c) above. The formal definition of the flux across $\mathcal{S}_i$ is defined as

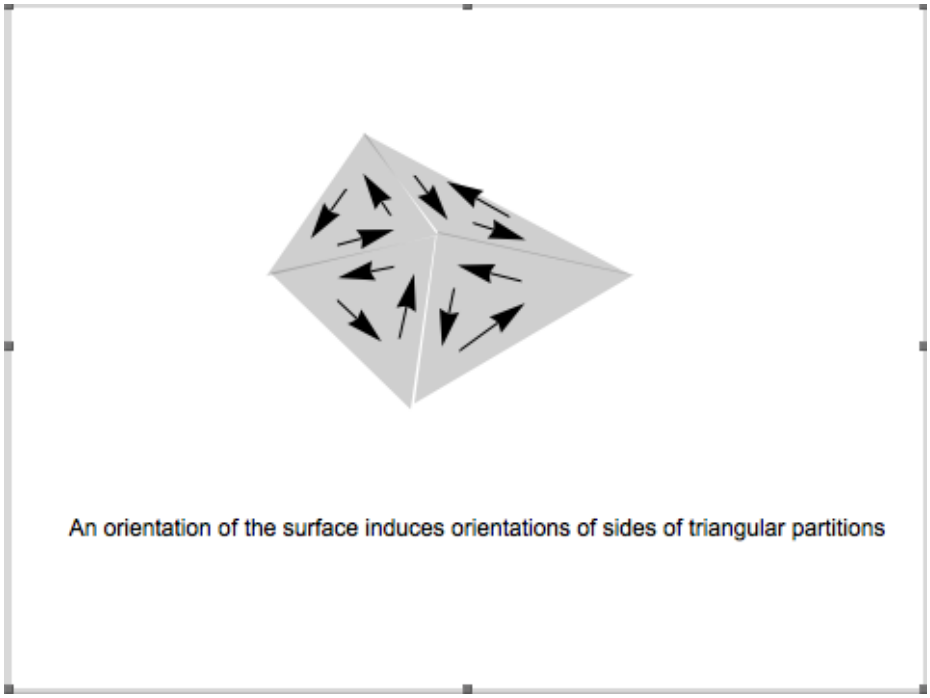
$$\int_{\mathcal{S}_i} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS,$$

where $\mathbf{n}(\mathbf{x})$ denotes the chosen unit normal to $\mathcal{S}_i$. Then

$$\int_{\mathcal{S}} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS = \int_{\mathcal{S}_1} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS + \cdots + \int_{\mathcal{S}_k} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS,$$

if $\mathcal{S} = \cup_{i=1}^{k} \mathcal{S}_i$ in the sense discussed above.

To compute this integral, we first need to choose a parametrization of $\mathcal{S}_i$ consistent with the chosen orientation of $\mathcal{S}_i$. This means that, the parametrization $\mathbf{x} = X(u, v)$, $(u, v) \in D \subset \mathbb{R}^2$ of $\mathcal{S}_i$, gives rise to a normal vector field $X_u(u, v) \times X_v(u, v)$, which



points in the same direction as the chosen unit normal $\mathbf{n}(X(u,v))$. Analytically this means

$$X_u(u,v) \times X_v(u,v) = \|X_u(u,v) \times X_v(u,v)\|\mathbf{n}(X(u,v)).$$

Since $dS = \|X_u(u,v) \times X_v(u,v)\|\, du\, dv$, we have

$$\int_{\mathcal{S}} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\, dS = \int_D F(X(u,v)) \cdot (X_u(u,v) \times X_v(u,v))\, dudv,$$

which can now be evaluated as an ordinary double integral in two dimensions.

One important observation is that this integral is independent of the particular parametrization for $\mathcal{S}_i$, as long as it provides the right orientation. This is seen as follows. Suppose that $(\hat{u}, \hat{v}) \in \hat{D} \subset \mathbb{R}^2$ is another parametrization of $\mathcal{S}_i$. This means that there exists a one-to-one differentiable $G : \hat{D} \mapsto D$ such that $\hat{X} = X \circ G$ is the new parametrization. Then by the chain rule

$$\hat{X}_{\hat{u}} = X_u \frac{\partial u}{\partial \hat{u}} + X_v \frac{\partial v}{\partial \hat{u}}$$
$$\hat{X}_{\hat{v}} = X_u \frac{\partial u}{\partial \hat{v}} + X_v \frac{\partial v}{\partial \hat{v}}$$

so

$$\hat{X}_{\hat{u}} \times \hat{X}_{\hat{v}} = \left[X_u \frac{\partial u}{\partial \hat{u}} + X_v \frac{\partial v}{\partial \hat{u}}\right] \times \left[X_u \frac{\partial u}{\partial \hat{v}} + X_v \frac{\partial v}{\partial \hat{v}}\right] = \left[\frac{\partial u}{\partial \hat{u}}\frac{\partial v}{\partial \hat{v}} - \frac{\partial v}{\partial \hat{u}}\frac{\partial u}{\partial \hat{v}}\right] X_u \times X_v.$$

The assumption that $(\hat{u}, \hat{v})$ provides the right orientation means that $\frac{\partial u}{\partial \hat{u}}\frac{\partial v}{\partial \hat{v}} - \frac{\partial v}{\partial \hat{u}}\frac{\partial u}{\partial \hat{v}} = \frac{\partial(u,v)}{\partial(\hat{u},\hat{v})} > 0$ over $(\hat{u}, \hat{v}) \in \hat{D}$. Then

$$\begin{aligned}
&\int_{\hat{D}} F(X \circ G(\hat{u}, \hat{v})) \cdot \left(\hat{X}_{\hat{u}} \times \hat{X}_{\hat{v}} \right) \, d\hat{u}d\hat{v} \\
&= \int_{\hat{D}} F(X \circ G(\hat{u}, \hat{v})) \cdot (X_u \times X_v) \frac{\partial(u,v)}{\partial(\hat{u},\hat{v})} \, d\hat{u}d\hat{v} \\
&= \int_{D} F(X(u,v)) \cdot (X_u \times X_v) \, dudv
\end{aligned}$$

by the change of variables formula. It is crucial that $\frac{\partial(u,v)}{\partial(\hat{u},\hat{v})} > 0$ over $(\hat{u}, \hat{v}) \in \hat{D}$ in the above discussion. The positivity of the Jacobian determinant between two sets of parametrizations turn out to be a useful criterion for checking that they provide the same orientation.

Exercise 9.2.1. *Evaluate the flux of $F(x, y, z) = (x, -2y, 3z)$ across the unit sphere $\mathbb{S}^2$: $x^2 + y^2 + z^2 = 1$, with the outward unit normal.*

SOLUTION: (METHOD 1) Note that $\mathbf{n}(x, y, z) = (x, y, z)$, and $F(x, y, z) \cdot \mathbf{n}(x, y, z) = (x, -2y, 3z) \cdot (x, y, z) = x^2 - 2y^2 + 3z^2$, so

$$\int_{\mathbb{S}^2} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \, dS = \int_{\mathbb{S}^2} (x^2 - 2y^2 + 3z^2) \, dS.$$

This now becomes the integral of a scalar function on $\mathbb{S}^2$. We can evaluate this integral using spherical coordinates as

$$\int_0^{2\pi} \int_0^{\pi} (\sin^2\phi \cos^2\theta - 2\sin^2\phi \sin^2\theta + 3\cos^2\phi) \sin\phi \, d\phi d\theta.$$

But we can also exploit the special symmetry of this surface $\mathbb{S}^2$. It is geometrically convincing that

$$\int_{\mathbb{S}^2} x^2 \, dS = \int_{\mathbb{S}^2} y^2 \, dS = \int_{\mathbb{S}^2} z^2 \, dS.$$

Thus

$$\int_{\mathbb{S}^2} x^2 \, dS = \int_{\mathbb{S}^2} y^2 \, dS = \int_{\mathbb{S}^2} z^2 \, dS = \frac{1}{3}\int_{\mathbb{S}^2} \left(x^2 + y^2 + z^2\right) \, dS = \frac{1}{3}\int_{\mathbb{S}^2} 1 \, dS = \frac{4\pi}{3}.$$

From this we can evaluate $\int_{\mathbb{S}^2} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \, dS = \frac{16\pi}{3}$.

(METHOD 2) We use spherical parametrization $\mathbf{x} = G(\theta, \phi)$ for $\mathbb{S}^2$ in terms of (θ, ϕ) given by

$$\begin{cases} x = \sin\phi\cos\theta \\ y = \sin\phi\sin\theta \\ z = \cos\phi \end{cases}$$

Then

$$\begin{cases} G_\theta = (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0) \\ G_\phi = (\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi) \\ G_\theta \times G_\phi = -\sin\phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi) \end{cases}$$

Note that $\|G_\theta \times G_\phi\| = \sin\phi$, which is what we expected, as $\|G_\theta \times G_\phi\|\ d\phi d\theta = \sin\phi\ d\phi d\theta$ is the formula for computing area on the unit sphere using spherical coordinates. But note that this gives the opposite orientation as the designated one! Conceptually this can be fixed by treating (ϕ, θ), in stead of (θ, ϕ), as the correct parametrization, just as the rotation from the y-axis to the x-axis is opposite of the rotation from the x-axis to y-axis. From computational point of view, we just need to make sure to use $G_\phi \times G_\theta$, instead of $G_\theta \times G_\phi$, in computing $F \cdot \mathbf{n}$. Then the flux is computed as

$$\int_0^{2\pi}\int_0^{\pi} F(x,y,z)\cdot(G_\phi \times G_\theta)\ d\phi d\theta = \int_0^{2\pi}\int_0^{\pi} (x, 2y, 3z)\cdot[\sin\phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)]\ d\phi d\theta.$$

This turns out to be the same as $\int_0^{2\pi}\int_0^{\pi}(x, 2y, 3z)\cdot(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)\sin\phi\ d\phi d\theta$.
□

We next discuss another view on

$$\int_{\mathcal{S}} F(\mathbf{x})\cdot n(\mathbf{x})\ dS = \int_{\mathcal{S}} (F_1(\mathbf{x})n_1(\mathbf{x}) + F_2(\mathbf{x})n_2(\mathbf{x}) + F_3(\mathbf{x})n_3(\mathbf{x}))\, dS.$$

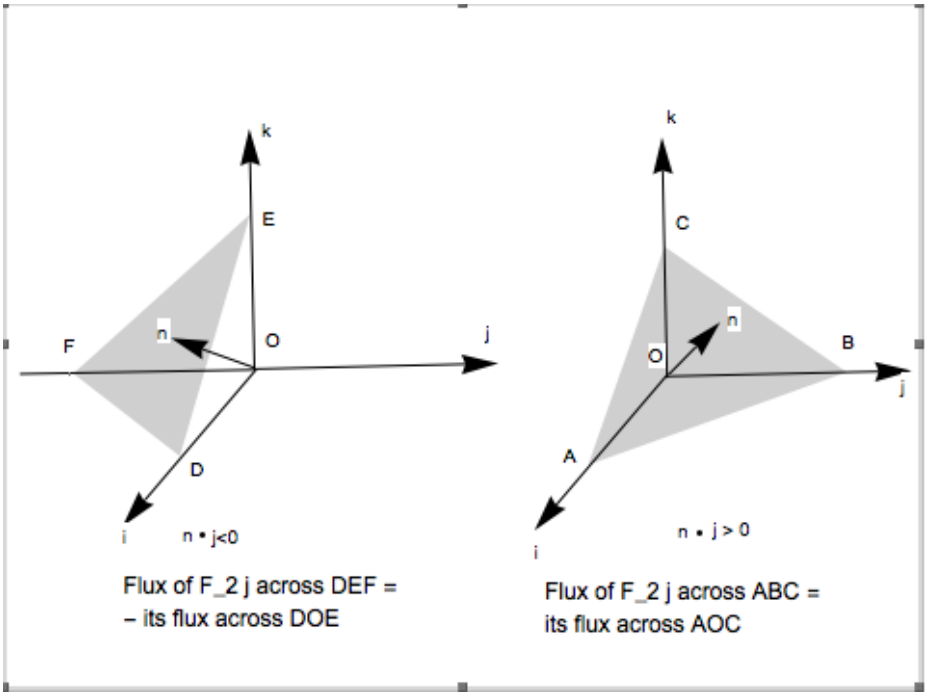
It is more illuminating to analyze each one term of the integrand separately. We will take $\int_{\mathcal{S}} F_2(\mathbf{x})n_2(\mathbf{x})\ dS$ as an example. Suppose that $\mathcal{S}$ can be represented as a graph $y = h(x, z)$ over some $D \subset \mathbb{R}^2$. Then

$$\mathbf{n}(x,y,z) = \pm\frac{(-h_x(x,z), 1, -h_z(x,z))}{\sqrt{1 + h_x(x,z)^2 + h_z(x,z)^2}} \quad \text{depending on whether } n_2(x,y,z) > 0 \text{ or } < 0.$$

Since $dS = \sqrt{1 + h_x(x,z)^2 + h_z(x,z)^2}\ dx\ dz$, we end up with

$$\int_{\mathcal{S}} F_2(\mathbf{x})n_2(\mathbf{x})\ dS = \pm\int_D F_2(x, h(x,z), z)\ dx\ dz,$$

namely, it has been reduced to an ordinary double integral in the x-z plane, with $F_2(x, h(x,z), z)$ as integrand. The geometric meaning of this is that $F_2(\mathbf{x})n_2(\mathbf{x})\ dS$ can be interpreted as the flux of $F_2(\mathbf{x})\mathbf{j}$ across the surface element $\mathbf{n}(\mathbf{x})\ dS$, and instead of projecting $F_2(\mathbf{x})\mathbf{j}$ in the direction of $\mathbf{n}(\mathbf{x})$, we interpret it as projecting $\mathbf{n}(\mathbf{x})\ dS$ along the $\mathbf{j}$ axis, namely, projecting the area element into the x-z plane, which is orthogonal to the $\mathbf{j}$ axis; and the other two terms carry a similar meaning.



The other two terms have a similar behavior, except that $F_1(\mathbf{x})n_1(\mathbf{x})\ dS$ is interpreted as flux of $F_1(\mathbf{x})\mathbf{i}$ across the projection of $\mathbf{n}(\mathbf{x})\ dS$ into the y-z plane, and $F_3(\mathbf{x})n_3(\mathbf{x})\ dS$ is interpreted as flux of $F_3(\mathbf{x})\mathbf{k}$ across the projection of $\mathbf{n}(\mathbf{x})\ dS$ into the x-y plane. In traditional textbooks, one often sees the notation

$$\int_{\mathcal{S}} F_1(\mathbf{x})\ dy\ dz + F_2(\mathbf{x})\ dz\ dx + F_3(\mathbf{x})\ dx\ dy$$

for

$$\int_{\mathcal{S}} (F_1(\mathbf{x})n_1(\mathbf{x}) + F_2(\mathbf{x})n_2(\mathbf{x}) + F_3(\mathbf{x})n_3(\mathbf{x}))\, dS.$$

The order $F_2(\mathbf{x})\ dz\ dx$ is meant that $\int_{\mathcal{S}} F_2(\mathbf{x})\ dz\ dx$ would be evaluated as $\int_D F_2(\mathbf{x})\ dz\ dx$ if the z-axis, x-axis, and $\mathbf{n}(\mathbf{x})$ form a right handed system, and would be evaluated as $-\int_D F_2(\mathbf{x})\ dz\ dx$ if the z-axis, x-axis, and $\mathbf{n}(\mathbf{x})$ form a left handed system.

In the case of the example above, $\mathbb{S}^2$ can be split into the union of $\mathbb{S}^2_{\text{right}}$ and $\mathbb{S}^2_{\text{left}}$, on each of which it can be represented as a graph $y = \pm\sqrt{1-x^2-z^2}$. On $\mathbb{S}^2_{\text{right}}$, $n_2(x,y,z) > 0$, and on $\mathbb{S}^2_{\text{left}}$, $n_2(x,y,z) < 0$, so we have

$$\begin{aligned}\int_{\mathbb{S}^2}(-2y)n_2(x,y,z)\ dS &= \int_{x^2+z^2\le 1}(-2y_{\text{right}})\ dx\ dz - \int_{x^2+z^2\le 1}(-2y_{\text{left}})\ dx\ dz \\ &= -2\int_{x^2+z^2\le 1}(y_{\text{right}} - y_{\text{left}})\ dx\ dz,\end{aligned}$$

which is $-2\times$ the volume of the unit ball.

9.3 Why are the divergence and curl of a vector field define that way?

If $F(x,y,z) = (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z))$ is a differentiable vector field in a region $U \subset \mathbb{R}^3$. Then we define its divergence at (x,y,z) as the scalar

$$\frac{\partial F_1(x,y,z)}{\partial x} + \frac{\partial F_2(x,y,z)}{\partial y} + \frac{\partial F_3(x,y,z)}{\partial z}.$$

It has the pattern of $(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}) \cdot (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z))$, and is often denoted as $\nabla \cdot F$. Another commonly used notation is div F.

We define curl of F at (x,y,z) as the vector

$$(\frac{\partial F_3(x,y,z)}{\partial y} - \frac{\partial F_2(x,y,z)}{\partial z}, \frac{\partial F_1(x,y,z)}{\partial z} - \frac{\partial F_3(x,y,z)}{\partial x}, \frac{\partial F_2(x,y,z)}{\partial x} - \frac{\partial F_1(x,y,z)}{\partial y}).$$

The curl has the pattern of the cross product

$$(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}) \times (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z)),$$

which we denote as $\nabla \times F$. Note that when the vector field is two dimensional: $F = (F_1(x,y), F_2(x,y), 0)$, $\nabla \times F = (0, 0, \frac{\partial F_2(x,y)}{\partial x} - \frac{\partial F_1(x,y)}{\partial y})$. But what motivates the definition of these two quantities?

It turns out the divergence of F at a location $\mathbf{x}_0 = (x_0, y_0, z_0)$ gives a measurement of infinitesimal flux of F across small closed surfaces surrounding (x_0, y_0, z_0) (such as a sphere or a box). More precisely, if $B_\epsilon(\mathbf{x}_0)$ denotes a ball of radius ϵ centered at $\mathbf{x}_0$ (or box of of side length ϵ centered at $\mathbf{x}_0$), then

$$\iint_{\partial B_\epsilon(\mathbf{x}_0)} F(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS \sim (\text{div}F(\mathbf{x}_0))\text{vol}(B_\epsilon)(\mathbf{x}_0).$$

More precisely,

$$\lim_{\epsilon\to 0}\frac{1}{\mathrm{vol}(B_\epsilon(\mathbf{x}_0))}\iint_{\partial B_\epsilon(\mathbf{x}_0)}F(\mathbf{x})\cdot\mathbf{n}(\mathbf{x})\;dS=\mathrm{div}F(\mathbf{x}_0).$$

Namely, $\mathrm{div}F(\mathbf{x}_0)$ equals the infinitesimal flux of F across small closed surfaces surrounding (x_0,y_0,z_0) per unit volume.

When we compute $\iint_{\partial B_\epsilon(\mathbf{x}_0)}F(\mathbf{x})\cdot\mathbf{n}(\mathbf{x})\;dS$ for $\epsilon>0$ small, it is natural to use the linear approximation of F at $\mathbf{x}_0$:

$$F(\mathbf{x})\sim F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0).$$

We will see easily that if we replace F by $F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)$, then

$$\iint_{\partial B_\epsilon(\mathbf{x}_0)}[F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)]\cdot\mathbf{n}(\mathbf{x})\;dS=[\mathrm{div}F(\mathbf{x}_0)]\,\mathrm{vol}(B_\epsilon).$$

This is the easiest to see when $B_\epsilon(\mathbf{x}_0)$ is a cube centered at $\mathbf{x}_0$ with side length $\epsilon>0$. In such a case $\partial B_\epsilon(\mathbf{x}_0)$ has 6 congruent square faces with area ϵ^2. The top and bottom faces have unit normal vector $(0,0,\pm1)$ respectively, so

$$\begin{aligned}&[F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)]\cdot\mathbf{n}(\mathbf{x})\\ =&F_3(\mathbf{x}_0)+\frac{\partial F_3}{\partial x}(\mathbf{x}_0)(x-x(0))+\frac{\partial F_3}{\partial y}(\mathbf{x}_0)(y-y(0))+\frac{\partial F_3}{\partial z}(\mathbf{x}_0)\epsilon/2.\end{aligned}$$

Since $(x-x(0))$ and $(y-y(0))$ have odd symmetry with respect to the square centered at $(x(0),y(0))$, we see that

$$\iint_{\mathrm{top}}[F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)]\cdot\mathbf{n}(\mathbf{x})\;dS=\left[F_3(\mathbf{x}_0)+\frac{\partial F_3}{\partial z}(\mathbf{x}_0)\epsilon/2\right]\epsilon^2.$$

Similarly,

$$\iint_{\mathrm{bot}}[F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)]\cdot\mathbf{n}(\mathbf{x})\;dS=-\left[F_3(\mathbf{x}_0)-\frac{\partial F_3}{\partial z}(\mathbf{x}_0)\epsilon/2\right]\epsilon^2.$$

Thus

$$\iint_{\mathrm{top}\cup\mathrm{bot}}[F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)]\cdot\mathbf{n}(\mathbf{x})\;dS=\frac{\partial F_3}{\partial z}(\mathbf{x}_0)\epsilon^3.$$

The flux of F across the other two pairs of opposite squares are handled similarly, and we are led to

$$\iint_{\partial B_\epsilon(\mathbf{x}_0)}[F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)]\cdot\mathbf{n}(\mathbf{x})\;dS=\left[\frac{\partial F_1}{\partial x}(\mathbf{x}_0)+\frac{\partial F_2}{\partial y}(\mathbf{x}_0)+\frac{\partial F_3}{\partial z}(\mathbf{x}_0)\right]\epsilon^3,$$

which confirms that div $F = \frac{\partial F_1}{\partial x}(\mathbf{x}_0) + \frac{\partial F_2}{\partial y}(\mathbf{x}_0) + \frac{\partial F_3}{\partial z}(\mathbf{x}_0)$ measures the infinitesimal rate of flux of F across closed cubes centered around $\mathbf{x}_0$.

The above conclusion will also follow from the divergence theorem directly, but it is instructive to do this computation to understand the origin for the notion of divergence.

Let's also verify the case when $B_\epsilon(\mathbf{x}_0)$ is the ball of radius $\epsilon > 0$ centered at $\mathbf{x}_0$. Let's set $\mathbf{x}_0 = (0,0,0)$ for notational simplicity, then $\mathbf{n}(\mathbf{x}) = \mathbf{x}/\epsilon$, and

$$\iint_{\partial B_\epsilon(\mathbf{0})} F(\mathbf{0}) \cdot \mathbf{n}(\mathbf{x}) \, dS = \iint_{\partial B_\epsilon(\mathbf{0})} \epsilon^{-1} \left[F_1(\mathbf{0})x + F_2(\mathbf{0})y + F_3(\mathbf{0})z \right] \, dS.$$

Due to the odd symmetry of x, y, and z on $\partial B_\epsilon(\mathbf{0})$, we see that

$$\iint_{\partial B_\epsilon(\mathbf{0})} x \, dS = \iint_{\partial B_\epsilon(\mathbf{0})} y \, dS = \iint_{\partial B_\epsilon(\mathbf{0})} z \, dS = 0.$$

At this point we see that $\iint_{\partial B_\epsilon(\mathbf{0})} F(\mathbf{0}) \cdot \mathbf{n}(\mathbf{x}) \, dS = 0$.

Next, we note that

$$\begin{aligned}
[DF(\mathbf{0})]\mathbf{x} \cdot \mathbf{n}(\mathbf{x}) = &\begin{bmatrix} \frac{\partial F_1}{\partial x}(\mathbf{0})x + \frac{\partial F_1}{\partial y}(\mathbf{0})y + \frac{\partial F_1}{\partial z}(\mathbf{0})z \\ \frac{\partial F_2}{\partial x}(\mathbf{0})x + \frac{\partial F_2}{\partial y}(\mathbf{0})y + \frac{\partial F_2}{\partial z}(\mathbf{0})z \\ \frac{\partial F_3}{\partial x}(\mathbf{0})x + \frac{\partial F_3}{\partial y}(\mathbf{0})y + \frac{\partial F_3}{\partial z}(\mathbf{0})z \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \epsilon^{-1} \\
= &\epsilon^{-1} \left\{ \frac{\partial F_1}{\partial x}(\mathbf{0})x^2 + \frac{\partial F_1}{\partial y}(\mathbf{0})yx + \frac{\partial F_1}{\partial z}(\mathbf{0})zx \right. \\
&+ \frac{\partial F_2}{\partial x}(\mathbf{0})xy + \frac{\partial F_2}{\partial y}(\mathbf{0})y^2 + \frac{\partial F_2}{\partial z}(\mathbf{0})zy \\
&\left. + \frac{\partial F_3}{\partial x}(\mathbf{0})xz + \frac{\partial F_3}{\partial y}(\mathbf{0})yz + \frac{\partial F_3}{\partial z}(\mathbf{0})z^2 \right\}
\end{aligned}$$

Using the symmetry of x^2, y^2, z^2, xy, yz, zx and that of $\partial B_\epsilon(\mathbf{0})$ we see that

$$\begin{aligned}
&\iint_{\partial B_\epsilon(\mathbf{0})} x^2 \, dS = \iint_{\partial B_\epsilon(\mathbf{0})} y^2 \, dS \quad = \iint_{\partial B_\epsilon(\mathbf{0})} z^2 \, dS = \iint_{\partial B_\epsilon(\mathbf{0})} \frac{x^2+y^2+z^2}{3} \, dS = \frac{4\pi\epsilon^4}{3} \\
&\iint_{\partial B_\epsilon(\mathbf{0})} xy \, dS = \iint_{\partial B_\epsilon(\mathbf{0})} yz \, dS \quad = \iint_{\partial B_\epsilon(\mathbf{0})} zx \, dS = 0.
\end{aligned}$$

Thus we see that

$$\iint_{\partial B_\epsilon(\mathbf{0})} [DF(\mathbf{0})]\mathbf{x} \cdot \mathbf{n}(\mathbf{x}) \, dS = \frac{4\pi\epsilon^3}{3} \left(\frac{\partial F_1}{\partial x}(\mathbf{0}) + \frac{\partial F_2}{\partial y}(\mathbf{0}) + \frac{\partial F_3}{\partial z}(\mathbf{0}) \right) = \frac{4\pi\epsilon^3}{3} \left(\text{div } F(\mathbf{0}) \right).$$

Next we discuss the origin for the notion of the curl of a vector field. On any plane passing through $\mathbf{x}_0$ with $\mathbf{n}$ as a unit normal, let $\mathcal{C}$ be a closed curve surrounding $\mathbf{x}_0$—we will take it to be a circle of radius $\epsilon > 0$ centered at $\mathbf{x}_0$, and orient it so that the orientations of $\mathcal{C}$ and $\mathbf{n}$ follow the right hand rule. We will see that in computing $\int_{\mathcal{C}} F \cdot d\mathbf{r} = \int_{\mathcal{C}} F_1\, dx + F_2\, dy + F_3\, dz$, the leading order term will be $\pi\epsilon^2(\nabla \times F)(\mathbf{x}_0) \cdot \mathbf{n}$. Thus, among circles with radius ϵ centered at $\mathbf{x}_0$, when $\mathbf{n}$ is aligned with $(\nabla \times F)(\mathbf{x}_0)$, one gets maximal circulation $\int_{\mathcal{C}} F \cdot d\mathbf{r}$. This gives a geometric meaning for the notion of curl of a vector field: $(\nabla \times F)(\mathbf{x}_0) \cdot \mathbf{n}$ equals the infinitesimal circulation of F per unit area along closed loops surrounding $\mathbf{x}_0$ in the plane with unit normal vector $\mathbf{n}$, and the direction of $(\nabla \times F)(\mathbf{x}_0)$ in some sense gives the direction of (maximal) axis of rotation of the vector field F around $\mathbf{x}_0$. When the vector field is two dimensional: $F = (F_1(x, y), F_2(x, y), 0)$, $\nabla \times F = (0, 0, \frac{\partial F_2(x,y)}{\partial x} - \frac{\partial F_1(x,y)}{\partial y})$, which is consistent with our intuition that any axis of of rotation of a two vector field is perpendicular to the plane in which the vector field lies.

To verify the claim, we will again use the linear approximation $F(\mathbf{x}_0)+[DF(\mathbf{x}_0)](\mathbf{x}-\mathbf{x}_0)$ to approximate $F(\mathbf{x})$ when doing the integration along small circles surrounding $\mathbf{x}_0$. For notational simplicity, we will take $\mathbf{x}_0 = 0$. Then

$$
\begin{aligned}
& F_1\, dx + F_2\, dy + F_3\, dz \\
\approx & \left[F_1(\mathbf{0}) + \frac{\partial F_1}{\partial x}(\mathbf{0})x + \frac{\partial F_1}{\partial y}(\mathbf{0})y + \frac{\partial F_1}{\partial z}(\mathbf{0})z\right] dx \\
& + \left[F_2(\mathbf{0}) + \frac{\partial F_2}{\partial x}(\mathbf{0})x + \frac{\partial F_2}{\partial y}(\mathbf{0})y + \frac{\partial F_2}{\partial z}(\mathbf{0})z\right] dy \\
& + \left[F_3(\mathbf{0}) + \frac{\partial F_3}{\partial x}(\mathbf{0})x + \frac{\partial F_3}{\partial y}(\mathbf{0})y + \frac{\partial F_3}{\partial z}(\mathbf{0})z\right] dz \\
= & [F_1(\mathbf{0})dx + F_2(\mathbf{0})dy + F_3(\mathbf{0})dz] \\
& + \left[\frac{\partial F_1}{\partial x}(\mathbf{0})\ xdx + \frac{\partial F_2}{\partial y}(\mathbf{0})\ ydy + \frac{\partial F_3}{\partial z}(\mathbf{0})\ zdz\right] \\
& + \left[\frac{\partial F_1}{\partial y}(\mathbf{0})\ ydx + \frac{\partial F_2}{\partial x}(\mathbf{0})\ xdy\right] \\
& + \left[\frac{\partial F_1}{\partial z}(\mathbf{0})\ zdx + \frac{\partial F_3}{\partial x}(\mathbf{0})\ xdz\right] \\
& + \left[\frac{\partial F_2}{\partial z}(\mathbf{0})\ zdy + \frac{\partial F_3}{\partial y}(\mathbf{0})\ ydz\right]
\end{aligned}
$$

We will see that the integrals along the closed curve $\mathcal{C}$ have the following properties

$$\begin{aligned}
&\int_{\mathcal{C}} 1\ dx = \int_{\mathcal{C}} 1\ dy = \int_{\mathcal{C}} 1\ dz = 0,\\
&\int_{\mathcal{C}} x dx = \int_{\mathcal{C}} y dy = \int_{\mathcal{C}} z dz = 0,\\
&\int_{\mathcal{C}} x dy = -\int_{\mathcal{C}} y dx,\\
&\int_{\mathcal{C}} y dz = -\int_{\mathcal{C}} z dy,\\
&\int_{\mathcal{C}} z dx = -\int_{\mathcal{C}} x dz.
\end{aligned}$$

It then follows that

$$\begin{aligned}
&\int_{\mathcal{C}} F \cdot d\mathbf{r}\\
\approx &\int_{\mathcal{C}} [F(\mathbf{0}) + [DF(\mathbf{0})](\mathbf{x})] \cdot d\mathbf{r}\\
= &\left[\frac{\partial F_2}{\partial x}(\mathbf{0}) - \frac{\partial F_1}{\partial y}(\mathbf{0})\right] \int_{\mathcal{C}} x dy\\
&+ \left[\frac{\partial F_1}{\partial z}(\mathbf{0}) - \frac{\partial F_3}{\partial x}(\mathbf{0})\right] \int_{\mathcal{C}} z dx\\
&+ \left[\frac{\partial F_3}{\partial y}(\mathbf{0}) - \frac{\partial F_2}{\partial z}(\mathbf{0})\right] \int_{\mathcal{C}} y dz
\end{aligned}$$

Finally we will see that

$$\int_{\mathcal{C}} y dz = \pi\epsilon^2 n_1, \quad \int_{\mathcal{C}} z dx = \pi\epsilon^2 n_2, \quad \int_{\mathcal{C}} x dy = \pi\epsilon^2 n_3.$$

It then follows that

$$\int_{\mathcal{C}} [F(\mathbf{0}) + [DF(\mathbf{0})](\mathbf{x})] \cdot d\mathbf{r} = (\nabla \times F)(\mathbf{0}) \cdot \mathbf{n}\ \pi\epsilon^2.$$

We now sketch verifications for several of the integral properties used. Take any parametrization $t : [0, l] \mapsto \mathbf{x}(t)$ of a closed curve $\mathcal{C}$, we have

$$\int_{\mathcal{C}} 1 dx = \int_0^l x'(t)\ dt = x(t)\Big|_{t=0}^{t=l} = 0, \quad \int_{\mathcal{C}} x dx = \int_0^l x(t)x'(t)\ dt = \frac{x(t)^2}{2}\Big|_{t=0}^{t=l} = 0,$$

$$\int_{\mathcal{C}} x dy + y dx = \int_{\mathcal{C}} [x(t)y'(t) + x'(t)y(t)]\, dt = [x(t)y(t)]\,\Big|_{t=0}^{t=l} = 0, \quad \text{so } \int_{\mathcal{C}} x dy = -\int_{\mathcal{C}} y dx.$$

Finally, choose two orthonormal vectors $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ in the plane such that $\mathbf{u} \times \mathbf{v} = \mathbf{n}$. We can parametrize the circle $\mathcal{C}$ in this plane centered at $\mathbf{x}_0$ with radius $\epsilon > 0$ by $\mathbf{x}(t) = \epsilon \cos t\ \mathbf{u} + \epsilon \sin t\ \mathbf{v}$ (we have chosen $\mathbf{x}_0 = \mathbf{0}$). Then

$$\begin{cases} x(t) = \epsilon \cos t\ u_1 + \epsilon \sin t\ v_1 \\ y(t) = \epsilon \cos t\ u_2 + \epsilon \sin t\ v_2 \\ y'(t) = -\epsilon \sin t\ u_2 + \epsilon \cos t\ v_2 \\ x(t)y'(t) = \epsilon^2 \left[(v_1 v_2 - u_1 u_2) \cos t \sin t + u_1 v_2 \cos^2 t - u_2 v_1 \sin^2 t \right] \end{cases}$$

Using $\int_0^{2\pi} \cos t \sin t\ dt = 0$, and $\int_0^{2\pi} \cos^2 t\ dt = \int_0^{2\pi} \sin^2 t\ dt = \pi$, we see that

$$\int_{\mathcal{C}} x dy = \int_0^{2\pi} x(t) y'(t)\ dt = \pi \epsilon^2 (u_1 v_2 - u_2 v_1) = \pi \epsilon^2 n_3.$$

Once Stokes' theorem has been at one's disposal, one can also regard $\int_{\mathcal{C}} x dy$ as the circulation of the vector field $x\mathbf{j}$ along $\mathcal{C}$ and applies Stokes's theorem, with the disk D spanned by $\mathcal{C}$ in the plane as the surface, then

$$\int_{\mathcal{C}} x dy = \iint_D \nabla \times (x\mathbf{j}) \cdot \mathbf{n}\ dS.$$

$\nabla \times (x\mathbf{j}) = \mathbf{k}$, so $\nabla \times (x\mathbf{j}) \cdot \mathbf{n} = n_3$, and $\iint_D \nabla \times (x\mathbf{j}) \cdot \mathbf{n}\ dS = n_3 \text{Area } (D) = \pi \epsilon^2 D$.

Remark 9.3.1

Computing the curl of a vector field using the cross product pattern is tedious and prone to errors, a more efficient way to compute it is the following.

(i). Compute the differential of any multi-variable function as in one variable calculus: $df(x, y, z) = f_x\ dz + f_y\ dy + f_z\ dz$.

(ii). For a given vector field $F(x, y, z) = (F_1(x, y, z), F_2(x, y, z), F_3(x, y, z))$, write $F(x, y, z) \cdot d\mathbf{r}$ in differential form $F_1(x, y, z)\ dx + F_2(x, y, z)\ dy + F_3(x, y, z)\ dz$.

(iii). Apply the following rules

$$\begin{aligned}
&d(F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz)\\
=&dF_1(x,y,z)\ dx + dF_2(x,y,z)\ dy + dF_3(x,y,z)\ dz\\
=&\left(\frac{\partial F_1(x,y,z)}{\partial x}\ dx + \frac{\partial F_1(x,y,z)}{\partial y}\ dy + \frac{\partial F_1(x,y,z)}{\partial z}\ dz\right)\ dx\\
&+\left(\frac{\partial F_2(x,y,z)}{\partial x}\ dx + \frac{\partial F_2(x,y,z)}{\partial y}\ dy + \frac{\partial F_2(x,y,z)}{\partial z}\ dz\right)\ dy\\
&+\left(\frac{\partial F_3(x,y,z)}{\partial x}\ dx + \frac{\partial F_3(x,y,z)}{\partial y}\ dy + \frac{\partial F_3(x,y,z)}{\partial z}\ dz\right)\ dz
\end{aligned}$$

and treat $dx\ dx = 0$, $dy\ dy = 0$, $dz\ dz = 0$, $dy\ dx = -dx\ dy$, $dz\ dy = -dy\ dz$, *and* $dx\ dz = -dz\ dx$, *so*

$$\begin{aligned}
&d(F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz)\\
=&\left[\frac{\partial F_2(x,y,z)}{\partial x} - \frac{\partial F_1(x,y,z)}{\partial y}\right]\ dx\ dy\\
&+\left[\frac{\partial F_3(x,y,z)}{\partial y} - \frac{\partial F_2(x,y,z)}{\partial z}\right]\ dy\ dz\\
&+\left[\frac{\partial F_1(x,y,z)}{\partial z} - \frac{\partial F_3(x,y,z)}{\partial x}\right]\ dz\ dx
\end{aligned}$$

In advanced courses, an exterior product, also called wedge product, is introduced among dx, dy, dz, *which obeys the anti-symmetry described above, so the rules above are written as* $dx \wedge dx = 0$, $dy \wedge dx = -dx \wedge dy$, *etc.*

Finally, using $n_1\ dS = dy\ dz$, $n_2\ dS = dz\ dz$, *and* $n_3\ dS = dx\ dy$ *to identify the above expression as*

$$(\nabla \times F) \cdot \mathbf{n}\ dS = (\nabla \times F)_1\ dy\ dz + (\nabla \times F)_2\ dz\ dx + (\nabla \times F)_3\ dx\ dy,$$

so

$$\begin{aligned}
&\nabla \times F(\mathbf{x})\\
=&\left(\frac{\partial F_3(x,y,z)}{\partial y} - \frac{\partial F_2(x,y,z)}{\partial z}, \frac{\partial F_1(x,y,z)}{\partial z} - \frac{\partial F_3(x,y,z)}{\partial x}, \frac{\partial F_2(x,y,z)}{\partial x} - \frac{\partial F_1(x,y,z)}{\partial y}\right),
\end{aligned}$$

and

$$\begin{aligned}
&d(F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz)\\
=&dF_1(x,y,z)\ dx + dF_2(x,y,z)\ dy + dF_3(x,y,z)\ dz\\
=&(\nabla \times F) \cdot \mathbf{n}\ dS.
\end{aligned}$$

This procedure may look complicated, but in concrete cases, it's fairly straightforward. E.g. for $F = x\mathbf{j}$, we would compute $d(x\ dy) = dx\ dy$, so $\nabla(x\mathbf{j}) = k$. For a vector field $(P(x,y), Q(x,y))$ in two dimensions, this procedure also makes it easy to compute the curl easily, as by the rules above,

$$d\left[P(x,y)dx + Q(x,y)dy\right] = P_y(x,y)\,dy\,dx + Q_x(x,y)\,dx\,dy = \left[Q_x(x,y) - P_y(x,y)\right]dxdy,$$

so $\int_{\mathcal{C}} P(x,y)dx + Q(x,y)dy = \iint_D \left[Q_x(x,y) - P_y(x,y)\right]dxdy$, if D is the region enclosed by the closed curve $\mathcal{C}$.
In more advanced courses, Green's Theorem and Stokes' Theorem are formulated using differential form as above and take on a very simple form

$$\int_{\partial S} Pdx + Qdy + Rdz = \iint_S d(Pdx + Qdy + Rdz).$$

9.4 Green's Theorem and Its Proof

Green's theorem relates the integral of a differentiable vector field along a closed piece-wise differentiable curve in the plane to the integral of the curl of this vector field in the region enclosed by the closed curve. It's more precise formulation is

Theorem 9.4.1

Suppose that $\mathcal{C}$ is a closed piece-wise differentiable curve in $\mathbb{R}^2$. Then it is a fact that it encloses a region D in it. Let $\mathcal{C}$ be oriented such that as a point traverses along it, the region D stays on its left. Let $F(x,y) = (P(x,y), Q(x,y))$ be a vector field differentiable in D. Then

$$\int_{\mathcal{C}} F(\mathbf{x}) \cdot d\,\mathbf{x} = \int_{\mathcal{C}} P(x,y)\,dx + Q(x,y)\,dy = \iint_D \left(\frac{\partial Q(x,y)}{\partial x} - \frac{\partial P(x,y)}{\partial y}\right) dxdy, \tag{9.1}$$

In many contexts, a region D in $\mathbb{R}^2$ is given first, then it determines an orientation of its boundary curve according to the rule above (domains with holes have more than one boundary curves, the orientation of each needs to be determined according to the rule above), and we use ∂D to denote the boundary curve with the designated orientation. Then the Green's theorem takes the form

$$\int_{\partial D} P(x,y)\,dx + Q(x,y)\,dy = \iint_D \left(\frac{\partial Q(x,y)}{\partial x} - \frac{\partial P(x,y)}{\partial y}\right) dxdy.$$

We note that, when D is the unit square $[0,1]\times[0,1]$, Green's Theorem is simply a straightforward consequence of the Fundamental Theorem of Calculus. Since the boundary ∂D of the unit square D consists of two horizontal segments and two vertical segments with opposite directions, we have

$$\begin{aligned}
&\int_{\partial D} P(x,y)\ dx + Q(x,y)\ dy \\
=&\int_0^1 \left[P(x,0) - P(x,1)\right]\ dx + \int_0^1 \left[Q(1,y) - Q(0,y)\right]\ dy \\
=&\int_0^1\int_0^1 -\frac{P(x,y)}{\partial y}\ dydx + \int_0^1\int_0^1 \frac{\partial Q(x,y)}{\partial x}\ dxdy \\
=&\iint_D \left[\frac{\partial Q(x,y)}{\partial x} - \frac{P(x,y)}{\partial y}\right]\ dxdy.
\end{aligned}$$

When D has a curvy side, say, given by $D=\{(x,y): a\le x\le b, c\le y\le h(x)\}$, essentially the same argument works. The top and bottom pieces give

$$\int_a^b \left(P(x,c) - P(x,h(x)) - Q(x,h(x))h'(x)\right)\ dx = -\int_a^b\int_c^{h(x)} \frac{P(x,y)}{\partial y}\,dy\,dx - \int_a^b Q(x,h(x))h'(x)\,dx;$$

while the two vertical pieces give

$$\int_c^{h(b)} Q(b,y)\,dy - \int_c^{h(a)} Q(a,y)\,dy.$$

Setting $q(x)=\int_c^{h(x)} Q(x,y)\,dy$, we see by the differentiation rules of integrals that

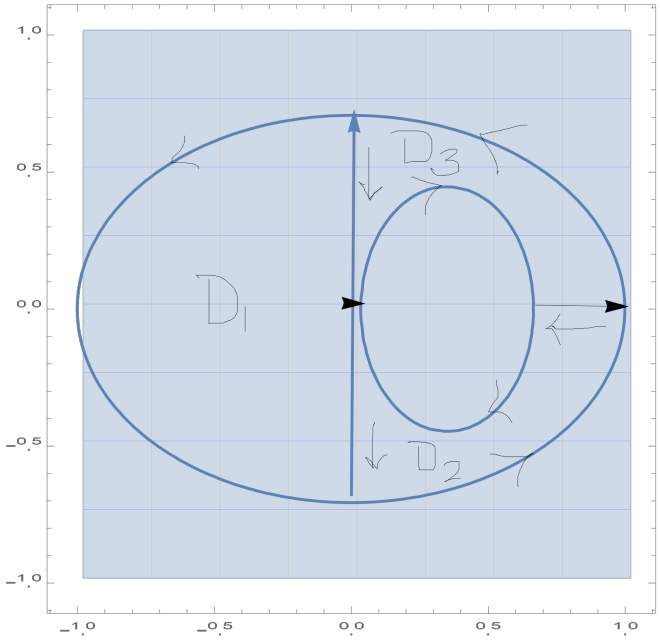
$$q'(x) = Q(x,h(x))h'(x) + \int_c^{h(x)} \frac{\partial Q(x,y)}{\partial x}\,dy,$$

so

$$\begin{aligned}
&\int_c^{h(b)} Q(b,y)\,dy - \int_c^{h(a)} Q(a,y)\,dy \\
=&q(b)-q(a) = \int_a^b q'(x)\,dx \\
=&\int_a^b Q(x,h(x))h'(x)\,dx + \int_a^b\int_c^{h(x)} \frac{\partial Q(x,y)}{\partial x}\,dy\,dx.
\end{aligned}$$

Putting these together we have established Green's theorem for such a case. This kind of arguments easily adapts to other domains of similar properties.

For a more general domain D, we can add some auxiliary lines to partition it into the non-overlapping union of several regions such that the above argument applies to each of the subregion. We can then use the "Additivity of circulations" to add up the all the equalities, and note that each of the added auxiliary lines appears exactly twice in the line integrals, but with opposite orientations, so in the outcome of the sum of the line integrals, only the contributions along the boundary curve of D remain. Note also how this kind of partition determines the orientation of the "inner" portion of the boundary curve, as illustrated in the figure here.

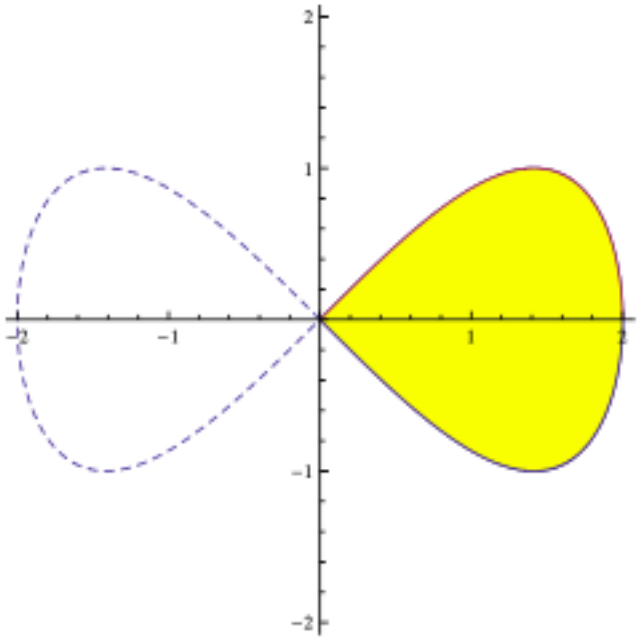


Exercise 9.4.1. *Let D be a bounded region in $\mathbb{R}^2$ with piecewise differentiable boundary curve ∂D. Then, according to the Green's Theorem, $\int_{\partial D} x dy = \iint_D 1\, dxdy = Area(D)$, as the curl of $(0, x)$ is 1. Likewise, $\int_{\partial D} y dx = \iint_D (-1)\, dxdy = -Area(D)$. Thus we can compute $Area(D)$ by either $\int_{\partial D} x dy$ or $-\int_{\partial D} y dx$.*

Here is a specific example. The Gerono lemniscate is given by $\mathbf{r}(t) = (2\sin t, 2\sin t \cos t)$. Its portion in the right half plane is a closed curve given by $0 \le t \le \pi$. The enclosed area can be computed by $\int x\, dy = \int_0^\pi x(t)y'(t)\, dt = \int_0^\pi 2\sin t\, 2\cos(2t)\, dt = 8/3$.

Remark 9.4.1

In applying Green's theorem, it is important that the vector field is continuously differentiable in the enclosed region. E.g., the vector field $(-\frac{y}{x^2+y^2}, \frac{x}{x^2+y^2})$ has



its curl equal to 0 *whenever* $(x, y) \neq (0, 0)$, *as*

$$\partial_x \left(\frac{x}{x^2 + y^2} \right) - \partial_y \left(-\frac{y}{x^2 + y^2} \right) = 0, \quad \text{for } (x, y,) \neq (0, 0);$$

On the other hand, we know

$$\int_{x^2+y^2=1} -\frac{y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy = \int_{x^2+y^2=1} -y\, dx + x\, dy = 2\pi.$$

The issue here is that the vector field is not continuously differentiable near $(0, 0)$. *Any application of the Green's theorem has to exclude a region near* $(0, 0)$. *If we are asked to calculate* $\int_C -\frac{y}{x^2+y^2} dx + \frac{x}{x^2+y^2} dy$, *where* C *is a somewhat complicated curve enclosing* $(0, 0)$, *say, an ellipse. Instead of trying to evaluate such an integral directly, we can choose a small* $\epsilon > 0$ *such that the circle* $x^2 + y^2 = \epsilon^2$ *is enclosed in* C, *and apply the Green's theorem in the region enclosed by both* C *and* $x^2 + y^2 = \epsilon^2$. *This would give us*

$$\int_C -\frac{y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy = \int_{x^2+y^2=\epsilon^2} -\frac{y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy,$$

where the orientation for the circle $x^2+y^2=\epsilon^2$ is also counterclockwise based on our discussion. But the latter can be evaluated easily, as

$$\begin{aligned}
&\int_{x^2+y^2=\epsilon^2} -\frac{y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy \\
=&\epsilon^{-2}\int_{x^2+y^2=\epsilon^2} -y\,dx + x\,dy \\
=&\epsilon^{-2}\int_{x^2+y^2\le\epsilon^2} 2\,dxdy \quad \text{by Green's theorem applied to } (-y,x) \\
=&2\pi.
\end{aligned}$$

Remark 9.4.2

Green's theorem is only for the line integral of a vector field along a closed curve. Sometimes one needs to calculate the line integral of a vector field along a curve which is not closed, but can be made closed by adjoining another curve. E.g., one may be asked to evaluate $\int -y\,dx + x\,dy$ along the portion of the Gerono lemniscate from $t=0$ to $t=\pi/2$. The curve starts at $(0,0)$ and ends at $(2,0)$. By adjoining the segment from $(2,0)$ to $(0,0)$, one forms a closed curve C, and $\int_C -y\,dx + x\,dy = \int_{\text{enclosed region by } C} 2\,dx\,dy$. The latter was known to be $8/3$ by the example earlier, while $\int_C -y\,dx + x\,dy$ also includes the contribution from the segment from $(2,0)$ to $(0,0)$. But on that segment, $y=0$, so the line integral there is 0. Thus the line integral being asked is $8/3$.

9.5 Stokes' Theorem and Its Proof

Stokes' theorem is an extension of Green's theorem when the curve is not necessarily a planar curve, or the enclosed surface is not a planar region. We now state Stokes' Theorem and outline a proof.

Theorem 9.5.1

Suppose that S is a bounded, oriented, piece-wise differentiable surface in $\mathbb{R}^3$, that $X(\mathbf{x}) = (X_1(\mathbf{x}), X_2(\mathbf{x}), X_3(\mathbf{x}))$ is a continuously differentiable vector field in a region Ω of $\mathbb{R}^3$ which contains S and its boundary curve ∂S (the boundary curve of S may consists of more than one component as in the case of a cylinder).

Then

$$\int_{\partial S} X(\mathbf{x}) \cdot d\,\mathbf{x} = \iint_S (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\; dS, \tag{9.2}$$

where $\mathbf{n}(\mathbf{x})$ stands for the designated unit normal to S at $\mathbf{x}$, and the orientation of ∂S is induced by that of S, namely, when S is formed as the non-overlapping union of several differentiable pieces S_i with continuously varying unit normals on each, it induces an orientation on the boundary curve of S_i by the right hand rule, and if two such pieces abut each other along a curve, then the induced orientation on that curve must be opposite to each other. This partition also partitions the boundary curve ∂S into the non-overlapping union of several curves, each of which is a boundary portion of some S_i with its induced orientation.

Remark 9.5.1

The most difficult part of this theorem is to make precise the notion of a bounded, oriented, piece-wise differentiable surface in $\mathbb{R}^3$, and its induced orientation on its boundary curve. Even though the geometric intuition may be clear to many, it is a challenge to find a mathematically precise way to describe such a surface. One difficulty is that, if we treat a surface as given by a single parametric representation, then rarely can we find an appropriate two dimensional domain in $\mathbb{R}^2$ as the domain of this representation. In more advanced treatment, one gives up using a single parametric representation to represent a surface; rather, one tries to find characterizations which are independent of specific parametric representation.

If we accept our geometric description for a partition of such a surface, we may assume that each piece S_i has a parametric representation as $\mathbf{x} = \Phi_i(u, v) : D \mapsto S_i$ for some simple domain D in $\mathbb{R}^2$. We will take D to be a unit square $\{(u, v) : 0 \le u, v \le 1\}$. Suppose we can establish Stokes' Theorem for each S_i:

$$\int_{\partial S_i} X(\mathbf{x}) \cdot d\,\mathbf{x} = \iint_{S_i} (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\; dS,$$

then

$$\begin{aligned}&\iint_S (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS\\ =&\sum_i \iint_{S_i} (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS\\ =&\sum_i \int_{\partial S_i} X(\mathbf{x}) \cdot d\ \mathbf{x}.\end{aligned}$$

In the last sum, the integrals over any sides which are shared by two abutting S_i's are canceled due to their opposite induced orientations; what remains is the integral on ∂S, with the appropriate orientation.
Thus it suffices to establish Stokes' Theorem for each such S_i, which we will do by transforming both sides to appropriate integrals on the square D, which is the domain of Φ_i. The proof in Rogawski and Adams uses a similar strategy, except they treat each S_i as a graph, and do computations using the graph.

Proof. We make one further reduction. Let $L[X]$ denote $\int_{\partial S_i} X(\mathbf{x}) \cdot d\ \mathbf{x}$. Then it is clear that $L[X+Y] = L[X] + L[Y]$ and $L[cX] = cL[X]$ for any continuously differentiable vector fields X and Y and any constant c. This property is called linearity. A similar property holds for the right hand side. Based on this property, it suffices to prove Stokes' Theorem for $X(\mathbf{x})$ of the form $X_1(\mathbf{x})\mathbf{i}$, $X_2(\mathbf{x})\mathbf{j}$, and $X_3(\mathbf{x})\mathbf{k}$ separately.

We will take $X(\mathbf{x}) = X_1(\mathbf{x})\mathbf{i}$, and prove

$$\int_{\partial S_i} X(\mathbf{x}) \cdot d\ \mathbf{x} = \int_{\partial D} X_1(\Phi_i(u,v)) \left[\frac{\partial x(\Phi_i(u,v))}{\partial u}\ du + \frac{\partial x(\Phi_i(u,v))}{\partial v}\ dv\right],$$

$$\iint_{S_i} (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS = \iint_D \left[\frac{\partial X_1(\Phi_i(u,v))}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial X_1(\Phi_i(u,v))}{\partial v}\frac{\partial x}{\partial u}\right]\ dudv. \tag{9.3}$$

(9.4)

Namely, we convert both sides into integrals in a two dimensional domain. Now if we set $P(u,v) = X_1(\Phi_i(u,v))\frac{\partial x(\Phi_i(u,v))}{\partial u}$ and $Q(u,v) = X_1(\Phi_i(u,v))\frac{\partial x(\Phi_i(u,v))}{\partial v}$, then we

apply Green's Theorem to 9.3 to obtain

$$\begin{aligned}
&\int_{\partial S_i} X(\mathbf{x}) \cdot d\ \mathbf{x} \\
&= \int_{\partial D} P(u,v)\ du + Q(u,v)\ dv \\
&= \iint_D \left[\frac{\partial Q(u,v)}{\partial u} - \frac{\partial P(u,v)}{\partial v} \right]\ dudv \\
&= \iint_D \left[\frac{\partial X_1(\Phi_i(u,v))}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial X_1(\Phi_i(u,v))}{\partial v}\frac{\partial x}{\partial u} \right]\ dudv \\
&= \iint_{S_i} (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS.
\end{aligned}$$

To prove 9.3 and 9.4, we note

$$X(\mathbf{x}) \cdot d\ \mathbf{x} = X_1(\mathbf{x})dx = X_1(\Phi_i(u,v)) \left[\frac{\partial x(\Phi_i(u,v))}{\partial u}\ du + \frac{\partial x(\Phi_i(u,v))}{\partial v} \right]\ dv,$$

and $(\nabla \times X)(\mathbf{x}) = (0, \partial_z X_1(\mathbf{x}), -\partial_y X_1(\mathbf{x}))$. But in the parametrization $\mathbf{x} = \Phi_i(u,v)$,

$$\mathbf{n}(\mathbf{x}) = \frac{\partial_u \Phi_i(u,v) \times \partial_v \Phi_i(u,v)}{\|\partial_u \Phi_i(u,v) \times \partial_v \Phi_i(u,v)\|},$$

and

$$\partial_u \Phi_i \times \partial_v \Phi_i = (\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}, \frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}, \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}).$$

So

$$\begin{aligned}
&\iint_{S_i} (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\ dS \\
&= \iint_D (\nabla \times X)(\mathbf{x}) \cdot (\partial_u \Phi_i \times \partial_v \Phi_i) du\ dv \\
&= \iint_D \left[\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial x}{\partial u}\frac{\partial z}{\partial v} \right) \partial_z X_1(\mathbf{x}) - \left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v} \right) \partial_y X_1(\mathbf{x}) \right] du\ dv \\
&= \iint_D \left[\left(\frac{\partial z}{\partial u}\partial_z X_1(\mathbf{x}) + \frac{\partial y}{\partial u}\partial_y X_1(\mathbf{x}) \right) \frac{\partial x}{\partial v} - \left(\frac{\partial z}{\partial v}\partial_z X_1(\mathbf{x}) + \frac{\partial y}{\partial v}\partial_y X_1(\mathbf{x}) \right) \frac{\partial x}{\partial u} \right] du\ dv \\
&= \iint_D \left[\frac{\partial X_1(\Phi_i(u,v))}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial X_1(\Phi_i(u,v))}{\partial v}\frac{\partial x}{\partial u} \right] du\ dv,
\end{aligned}$$

which concludes 9.4.

In fact, the proof works without breaking it down into different components of X by using

$$\iint_{S_i} (\nabla \times X)(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\, dS = \iint_D (\nabla \times X)(\mathbf{x}) \cdot (\Phi_u(u,v) \times \Phi_v(u,v))\, du\, dv,$$

$$\int_{\partial S_i} X(\mathbf{x}) \cdot d\mathbf{x} = \int_{\partial D} X(\mathbf{x}) \cdot \Phi_u(u,v)\, du + X(\mathbf{x}) \cdot \Phi_v(u,v)\, dv,$$

and the Green's theorem, together with the following relation

$$\mathrm{Curl}_z(X(\mathbf{x}) \cdot \Phi_u(u,v), X(\mathbf{x}) \cdot \Phi_v(u,v)) = (\nabla \times X)(\mathbf{x}) \cdot (\Phi_u(u,v) \times \Phi_v(u,v)).$$

□

Exercise 9.5.1. *Suppose one is asked to evaluate $\int_C z\,dx + x\,dy + (y+2z)\,dz$, where C is the intersection of the cylinder $x^2+y^2=1$ and the plane $z=x$, with the counterclockwise orientation when viewed from the top. One strategy is to find a parametrization of C and evaluate this line integral directly. Since C is a closed curve, it spans many surfaces, in particular, the planar portion on $z=x$ cut-out by the cylinder, so it may be worthwhile to explore the Stokes' Theorem. First, Curl $(z,x,y+2z) = (1,1,1)$, and a normal to $z=x$ is $(1,0,-1)$, which is orthogonal to $(1,1,1)$, so for any point (x,y,z) in the said planar region, Curl $(z,x,y+2z)\cdot\mathbf{n}(x,y,z) = 0$! This renders*

$$\int_C z\,dx + x\,dy + (y+2z)\,dz = \iint_{enclosed\ region} Curl\ (z,x,y+2z)\cdot\mathbf{n}(x,y,z) = 0.$$

If one is asked to evaluate $\int_{C_1} z\,dx + x\,dy + (y+2z)\,dz$, where C_1 is the portion of C in the upper half space running from $(0,-1,0)$ to $(0,1,0)$, then C_1 is no longer a closed curve, so we can't directly apply Stokes' theorem. But adjoining the segment from $(0,1,0)$ to $(0,-1,0)$ would create a closed curve C_1', which encloses one-half of the planar region encountered above, and we still have $\int_{C_1'} z\,dx+x\,dy+(y+2z)\,dz = 0$. But

$$\int_{C_1'} z\,dx+x\,dy+(y+2z)\,dz = \int_{C_1} z\,dx+x\,dy+(y+2z)\,dz + \int_{(0,1,0)\mapsto(0,-1,0)} z\,dx+x\,dy+(y+2z)\,dz,$$

and the letter can be evaluated easily as the segment from $(0,1,0)$ to $(0,-1,0)$ can be parametrized as $(0,(1-t)1+t(-1),0)$, $t\in[0,1]$, so

$$\int_{(0,1,0)\mapsto(0,-1,0)} z\,dx + x\,dy + (y+2z)\,dz = \int_0^1 0(-2)\,dt = 0.$$

This still leads to $\int_{C_1} z\,dx + x\,dy + (y+2z)\,dz = 0$.

Remark 9.5.2

Computing the curl of a vector field using the cross product pattern is tedious and prone to errors, a more efficient way to compute it is the following.

(i). Compute the differential of any multi-variable function as in one variable calculus: $df(x,y,z) = f_x\ dz + f_y\ dy + f_z\ dz$.

(ii). For a given vector field $F(x,y,z) = (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z))$, *write* $F(x,y,z)\cdot d\mathbf{r}$ *in differential form* $F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz$.

(iii). Apply the following rules

$$\begin{aligned}
&d(F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz)\\
=&dF_1(x,y,z)\ dx + dF_2(x,y,z)\ dy + dF_3(x,y,z)\ dz\\
=&\left(\frac{\partial F_1(x,y,z)}{\partial x}\ dx + \frac{\partial F_1(x,y,z)}{\partial y}\ dy + \frac{\partial F_1(x,y,z)}{\partial z}\ dz\right)\ dx\\
&+\left(\frac{\partial F_2(x,y,z)}{\partial x}\ dx + \frac{\partial F_2(x,y,z)}{\partial y}\ dy + \frac{\partial F_2(x,y,z)}{\partial z}\ dz\right)\ dy\\
&+\left(\frac{\partial F_3(x,y,z)}{\partial x}\ dx + \frac{\partial F_3(x,y,z)}{\partial y}\ dy + \frac{\partial F_3(x,y,z)}{\partial z}\ dz\right)\ dz
\end{aligned}$$

and treat $dx\ dx = 0$, $dy\ dy = 0$, $dz\ dz = 0$, $dy\ dx = -dx\ dy$, $dz\ dy = -dy\ dz$, *and* $dx\ dz = -dz\ dx$, *so*

$$\begin{aligned}
&d(F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz)\\
=&\left[\frac{\partial F_2(x,y,z)}{\partial x} - \frac{\partial F_1(x,y,z)}{\partial y}\right]\ dx\ dy\\
&+\left[\frac{\partial F_3(x,y,z)}{\partial y} - \frac{\partial F_2(x,y,z)}{\partial z}\right]\ dy\ dz\\
&+\left[\frac{\partial F_1(x,y,z)}{\partial z} - \frac{\partial F_3(x,y,z)}{\partial x}\right]\ dz\ dx
\end{aligned}$$

In advanced courses, an exterior product, also called wedge product, is introduced among dx, dy, dz, *which obeys the anti-symmetry described above, so the rules above are written as* $dx\wedge dx = 0$, $dy\wedge dx = -dx\wedge dy$, *etc.*

Finally, using $n_1\ dS = dy\ dz$, $n_2\ dS = dz\ dz$, *and* $n_3\ dS = dx\ dy$ *to identify the above expression as*

$$(\nabla\times F)\cdot\mathbf{n}\ dS = (\nabla\times F)_1\ dy\ dz + (\nabla\times F)_2\ dz\ dx + (\nabla\times F)_3\ dx\ dy,$$

so

$$
\begin{aligned}
(\nabla \times F(\mathbf{x}))_1 &= \frac{\partial F_3(x,y,z)}{\partial y} - \frac{\partial F_2(x,y,z)}{\partial z}, \\
(\nabla \times F(\mathbf{x}))_2 &= \frac{\partial F_1(x,y,z)}{\partial z} - \frac{\partial F_3(x,y,z)}{\partial x}, \\
(\nabla \times F(\mathbf{x}))_3 &= \frac{\partial F_2(x,y,z)}{\partial x} - \frac{\partial F_1(x,y,z)}{\partial y},
\end{aligned}
$$

and

$$
\begin{aligned}
& d(F_1(x,y,z)\ dx + F_2(x,y,z)\ dy + F_3(x,y,z)\ dz) \\
=& dF_1(x,y,z)\ dx + dF_2(x,y,z)\ dy + dF_3(x,y,z)\ dz \\
=& (\nabla \times F) \cdot \mathbf{n}\ dS.
\end{aligned}
$$

This procedure may look complicated, but in concrete cases, it's fairly straightforward. E.g. for $F = x\mathbf{j}$, we would compute $d(x\ dy) = dx\ dy$, so $\nabla(x\mathbf{j}) = k$. For a vector field $(P(x,y), Q(x,y))$ in two dimensions, this procedure also makes it easy to compute the curl easily, as by the rules above,

$$
\begin{aligned}
& d\left[P(x,y)dx + Q(x,y)dy\right] \\
=& P_y(x,y)\ dy\ dx + Q_x(x,y)\ dx\ dy = \left[Q_x(x,y) - P_y(x,y)\right] dxdy,
\end{aligned}
$$

so $\int_{\mathcal{C}} P(x,y)dx + Q(x,y)dy = \iint_D \left[Q_x(x,y) - P_y(x,y)\right] dxdy$, if D is the region enclosed by the closed curve $\mathcal{C}$.
In more advanced courses, Green's Theorem and Stokes' Theorem are formulated using differential form as above and take on a very simple form

$$
\int_{\partial \mathcal{S}} Pdx + Qdy + Rdz = \iint_{\mathcal{S}} d(Pdx + Qdy + Rdz).
$$