

MULTIVARIABLE CALCULUS

Eric A. Carlen
Professor of Mathematics
Rutgers University

August, 2019

Contents

1 GEOMETRY, ALGEBRA AND CALCULUS IN SEVERAL VARIABLES	1
1.1 Algebra and Geometry in \mathbb{R}^n	1
1.1.1 Geometry, Algebra and Calculus	1
1.1.2 Vector variables and Cartesian coordinates	2
1.1.3 Parameterization	4
1.1.4 The vector space \mathbb{R}^n	11
1.1.5 Geometry and the dot product	17
1.1.6 Parallel and orthogonal components	21
1.1.7 Orthonormal subsets of \mathbb{R}^n	23
1.1.8 Householder reflections and orthonormal bases	25
1.2 Lines and planes in \mathbb{R}^3	29
1.2.1 The cross product in \mathbb{R}^3	29
1.2.2 Lines and planes in \mathbb{R}^3	34
1.2.3 Distance problems	40
1.3 The Gram-Schmidt Orthonormalization Algorithm	45
1.3.1 The Gram-Schmidt Orthonormalization Algorithm in \mathbb{R}^3	45
1.3.2 The Gram-Schmidt Algorithm in general	47
1.3.3 Subspaces of \mathbb{R}^n	50
1.3.4 Orthogonal complements	52
1.3.5 Higher dimensional analogs of lines and planes	55
1.4 Exercises	57
2 DESCRIPTION OF MOTION	63
2.1 Functions from \mathbb{R} to \mathbb{R}^n and the description of motion	63
2.1.1 Continuity of functions from \mathbb{R} to \mathbb{R}^n	63
2.1.2 Differentiability of functions from \mathbb{R} to \mathbb{R}^n	65
2.1.3 Velocity and acceleration	68
2.1.4 Torsion and the Frenet–Serret formulae for a curve in \mathbb{R}^3	72
2.1.5 Curvature and torsion are independent of parameterization.	79
2.1.6 Speed and arc length	82

2.1.7	Speed, curvature and torsion are independent of the choice of a right-handed coordinate system	84
2.1.8	Geodesics in \mathbb{R}^n and on the unit sphere	87
2.1.9	Rotations, continuity and the right hand rule	92
2.2	Exercises	97
3	CONTINUOUS FUNCTIONS	101
3.1	Continuity in several variables	101
3.1.1	Functions of several variables	101
3.1.2	Continuity in several variables	103
3.1.3	Continuity and limits	108
3.1.4	The Squeeze Principle in several variables	111
3.1.5	Continuity versus separate continuity	113
3.2	Continuity, compactness and maximizers	117
3.2.1	Open and closed sets in \mathbb{R}^n	117
3.2.2	Minimizers and maximizers	119
3.2.3	Compactness and existence of maximizers	120
3.2.4	The Squeeze Principle revisited	125
3.3	Exercises	126
4	DIFFERENTIABLE FUNCTIONS	129
4.1	Vertical slices and directional derivatives	129
4.1.1	Directional derivatives and partial derivatives	129
4.1.2	The gradient and a chain rule for functions of a vector variable	132
4.1.3	The geometric meaning of the gradient	135
4.1.4	Critical points	137
4.1.5	The gradient and tangent planes	138
4.2	Linear functions from \mathbb{R}^n to \mathbb{R}^m	145
4.2.1	The matrix representation of linear functions	146
4.2.2	Composition of linear functions and matrix multiplication	151
4.2.3	Solving the equation $A\mathbf{x} = \mathbf{b}$	153
4.2.4	<i>QR</i> factorization	155
4.2.5	Matrix inverses	160
4.2.6	Continuity of matrix inverses	165
4.3	Differentiability of functions from \mathbb{R}^n to \mathbb{R}^m	168
4.3.1	Differentiability and best linear approximation in several variables	168
4.3.2	The general chain rule	170
4.4	Newton's Method	172
4.4.1	Linear approximation and Newton's iterative scheme	172
4.4.2	The Geometry of Newton's Method	174
4.4.3	Starting and stopping the iteration	176

4.5 Exercises	178
5 THE IMPLICIT FUNCTION THEOREM AND ITS CONSEQUENCES	183
5.1 Horizontal slices and contour curves	183
5.1.1 Implicit and explicit descriptions of planar curves	186
5.1.2 When is the contour curve actually a curve?	189
5.2 Constrained Optimization in Two variables	191
5.2.1 Lagrange's criterion for optimizers on the boundary	192
5.2.2 Application of Lagrange's Theorem	195
5.2.3 Optimization for regions with a piecewise smooth boundary	198
5.3 The Implicit Function Theorem via the Inverse Function Theorem	200
5.3.1 Inverting coordinate transformations	200
5.3.2 From the Inverse Function Theorem to the Implicit Function Theorem	203
5.3.3 Proof of the Inverse Function Theorem.	204
5.4 The general Implicit Function Theorem	207
5.5 Lagrange's Theorem in general	210
5.6 Exercises	215
6 CURVATURE AND QUADRATIC APPROXIMATION	219
6.1 Quadratic functions	219
6.1.1 The matrix form of a purely quadratic function	219
6.1.2 Purely quadratic functions as sums of squares	220
6.1.3 Eigenvalues and eigenvectors of a symmetric matrix	221
6.1.4 Computing eigenvectors and eigenvalues	226
6.2 The best quadratic approximation	228
6.2.1 Higher order directional derivatives and repeated partial differentiation	228
6.2.2 Clairault's Theorem	230
6.2.3 A multivariable second order Taylor expansion	231
6.2.4 Principal curvatures at a critical point	234
6.2.5 Contour plots near critical points	236
6.2.6 Types of critical points for real valued functions on \mathbb{R}^n	240
6.2.7 Sylvester's Criterion	241
6.3 Curvature of surfaces in \mathbb{R}^3	243
6.3.1 Parameterized surfaces in \mathbb{R}^3	243
6.3.2 The arclength of curves on a parameterized surface	248
6.3.3 Curvature and the second fundamental matrix	250
6.3.4 The Gauss map	256
6.4 Exercises	258

7 INTEGRATION IN SEVERAL VARIABLES	263
7.1 Integration and summation	263
7.1.1 A look back at integration in one variable	263
7.1.2 Integrals of functions on \mathbb{R}^2	266
7.1.3 Computing area integrals	269
7.1.4 Polar coordinates	272
7.2 Jacobians and changing variables of integration in \mathbb{R}^2	279
7.2.1 Letting the boundary of D determine the disintegration strategy	279
7.2.2 The change of variables formula for integrals in \mathbb{R}^2	285
7.2.3 An alternative computational method	290
7.3 Integration in \mathbb{R}^3	292
7.3.1 Reduction to iterated integrals in lower dimension	292
7.3.2 The change of variables formula for integrals in \mathbb{R}^3	294
7.4 Integration on parameterized surfaces	297
7.4.1 Parameterized surfaces	297
7.4.2 The surface area of a parameterized surface	299
7.5 Exercises	304
8 DETERMINANTS	309
8.1 Permuations	309
8.1.1 The permutation group	309
8.1.2 The character of a permutation	311
8.1.3 The permutation group as a metric space	314
8.2 Algebraic properties of the determinant	317
8.2.1 The determinant formula	317
8.3 The volume of sets in \mathbb{R}^n and the determinant	323
8.3.1 Volume in n -dimensional Euclidean space	324
8.3.2 The singular value decomposition	327
8.4 Exercises	329
9 FLUX AND CIRCULATION, DIVERGENCE AND CURL	331
9.1 Flows and flux	331
9.1.1 Vector fields and flows	331
9.1.2 Lipschitz vector fields on \mathbb{R}^n	333
9.1.3 Flux across an oriented curve in \mathbb{R}^2	336
9.1.4 Flux across oriented surfaces in \mathbb{R}^3	343
9.1.5 Computing flux integrals	347
9.2 The Divergence Theorem	348
9.2.1 The Divergence Theorem in the plane	348
9.2.2 The Divergence Theorem in \mathbb{R}^3	352
9.3 Line integrals, force fields and work	355

9.3.1	Conservative vector fields	356
9.3.2	Curl, circulation and Stokes' Theorem	358
9.3.3	Curl and conservative vector fields	366
9.3.4	Vector potentials	370
9.4	The Laplace operator and Poisson's Equation	372
9.4.1	The basic problem of electrostatics	372
9.4.2	Harmonic functions	373
9.4.3	The Hodge Decomposition of vector fields	378
9.5	Exercises	380

Chapter 1

GEOMETRY, ALGEBRA AND CALCULUS IN SEVERAL VARIABLES

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

1.1.1 Geometry, Algebra and Calculus

The basic questions studied in single variable calculus – finding the slope of the tangent line to the graph of a function, and finding the area under the graph of a function – involve geometry in an obvious way. However, the geometry involved is by and large simple planar geometry. The essential core of the subject may appear therefore to be the theory of limits, in which most of the subtlety of the subject resides.

In multivariable calculus, geometry plays a much more central role, and the geometric issues that we must contend with in order to solve problems in multivariable calculus are more challenging than the simple planar geometric issues that typically arise in the solution of single variable problems. Fortunately, powerful algebraic and analytic methods have been devised that facilitate treatment of these geometric problems.

Consider a geometric object, such a sphere in three dimensional space. The surface of the sphere is a two dimensional object in the obvious intuitive sense. The simplest sort of multivariable differential calculus problems concern such matters as finding the “tangent plane” to such a surface at a given point. Already, the problem of determining a “tangent plane” raises interesting issues. Tangent lines to the graph of a single variable function $f(x)$ are a simple matter: If f is differentiable at x_0 , then the point $(x_0, y_0) = (x_0, f(x_0))$ is on the tangent line at x_0 , and the slope of this tangent line is $f'(x_0)$, the derivative of f at x_0 . A point (x, y) in the plane lies on the tangent line if and only

if x and y solve the equation

$$y = y_0 + f'(x_0)(x - x_0) .$$

Tangent planes can be described in an analogous manner, but will require a *system* of two equations, and the analog of the slope will be a “vector” and not a number.

The methods with which we deal with tangent planes and their higher dimensional analogs are largely algebraic, and this is a very fortunate thing: While it is at least easy to visualize tangent planes in three dimensional space, we will be concerned with problems involving arbitrarily many variables. While the location of a *point* in three dimensional space can be described by specifying three “coordinates”, we are often concerned with both the position and attitude of objects in three dimensional space. If the object is an airplane, it likely matters where the nose is pointed and if the wings are horizontal or not. The position and attitude of an airplane – even treating it in the first approximation as a rigid body – requires 6 variables; 3 variables to specify the position of the center of mass, and 3 other variables to specify the attitude. (The three other variables may be taken to be Euler angles, which will be introduced later.) More complicated but equally natural problems require many more variables. It will not be enough to be able to deal with tangent planes in three dimensional space. To solve many natural problems, we will need to be able to deal with their higher dimensional analogs in arbitrarily many dimensions. We now lay the groundwork for this.

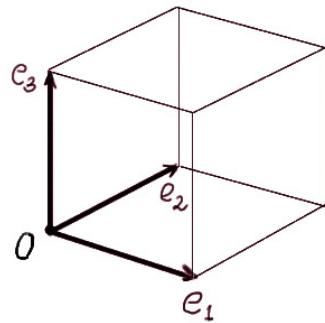
1.1.2 Vector variables and Cartesian coordinates

Many problems in science and engineering lead to the consideration of functions taking several variables as input, and returning several variables of output.

- *Multivariable functions are simply functions that take an ordered list of numbers as their input, or return an ordered list as output, or both.*

For instance, one such function might give the current temperature, barometric pressure, and relative humidity at a given point on the earth, as specified by latitude and longitude. In this case, there are two input variables, and three output variables. You will also recognize the input variables in this example as *coordinates*. Our subject has its real beginning with a fundamental idea of Rene Descartes, for whom *Cartesian coordinates* are named.

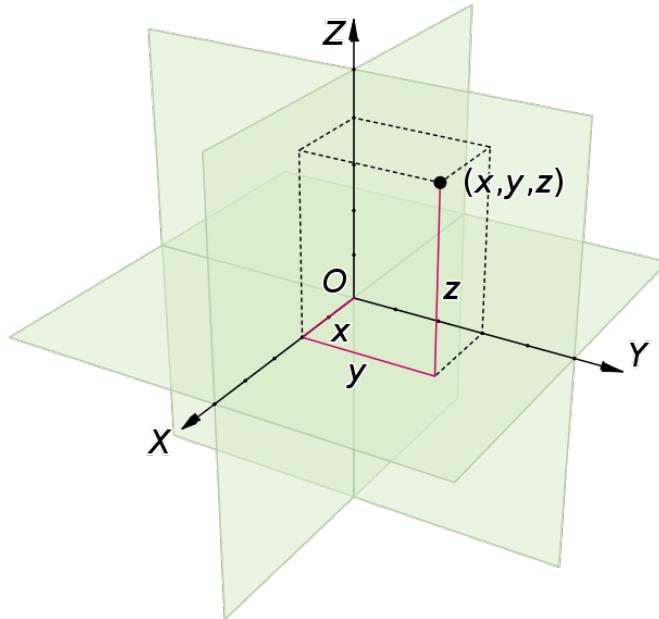
Descartes’ idea was to specify points in three dimensional Euclidean space using lists of three numbers (x, y, z) , such lists are now known as *vectors*. To do this one first fixes a *reference system*, by specifying a “base point” or “origin” that we shall denote by $\mathbf{0}$, and also a set of *three orthogonal directions*. For instance, if you are standing somewhere on the surface of the Earth, you might take the point at which you stand as the origin $\mathbf{0}$, and you might take East to be the first direction, North to be the second, and “straight up” to be the third. All of these directions are orthogonal to one another. Let us use the symbols \mathbf{e}_1 , \mathbf{e}_2 and \mathbf{e}_3 to denote these three directions.



The brilliant idea of Descartes is this:

- We can describe the exact position of any point in physical space by telling how to reach it by moving in the directions e_1 , e_2 and e_3 . This is simply a matter of “giving directions”: Start at the origin $\mathbf{0}$, and go out x units of distance in the e_1 direction, then go out y units of distance in the e_2 direction, and finally go out z units of distance in the e_3 direction. The numbers x , y and z may be positive or negative (or zero). If, say, x is negative, this means that you should go $|x|$ units of distance in the direction opposite to e_1 .

Thus, following Descartes’ idea, we can specify the exact position of any point in physical space by giving the ordered list of numbers (x, y, z) that describes how to reach it from the origin of our reference system. The three numbers x , y and z are called the *coordinates* of the point with respect to the given reference system. (It is important to note that the reference system is part of the description too: Knowing how far to go in each direction is not much use if you do not know the directions, or the starting point.)



This representation of points in space as ordered triples of numbers, such as (x, y, z) allows one to use algebra and calculus to solve problems in geometry, and it literally revolutionized mathematics.

We now define a *three dimensional vector* to be an ordered triple (x, y, z) of real numbers. The geometric interpretation is that we regard x , y and z as the *coordinates* of a unique point in physical space – the one you get to by starting from the origin and moving x units of distance in the \mathbf{e}_1 direction, y units of distance in the \mathbf{e}_2 direction, and z units of distance in the \mathbf{e}_3 direction. We may identify the vector (x, y, z) with this point in physical space, once again keeping in mind that this identification depends on the reference system, and that what the vector really represents is not the point itself, but the translation that carries the origin to that point.

As we have said, this way of identifying three dimensional vectors with points in physical space is extremely useful because it brings algebra to bear on geometric problems. For instance, referring to the previous diagram, you see from (two application of) the Pythagorean Theorem that the distance from the origin $\mathbf{0}$ to the point represented by the vector (x, y, z) is

$$\sqrt{x^2 + y^2 + z^2} .$$

This distance is called the *length* or *magnitude* of the vector $\mathbf{x} = (x, y, z)$. The vector \mathbf{x} also has a *direction*; namely the direction of the displacement that would carry one directly from $\mathbf{0}$ to \mathbf{x} . It is useful to associate to each direction the vector corresponding to a unit displacement in that direction. This provides a one-to-one correspondence between directions and *unit vectors*, i.e., vectors of unit length.

The unit sphere is defined to be the set of all unit vectors; i.e., all points a unit distance from the origin. Thus, a point represented by the vector $\mathbf{x} = (x, y, z)$ lies on the unit sphere if and only if

$$x^2 + y^2 + z^2 = 1 . \quad (1.1)$$

You are probably familiar with this as the equation for the unit sphere. But before Descartes, geometry and algebra were very different subjects, and the idea of describing a geometric object in terms of an algebraic equation was unknown. It revolutionized mathematics.

1.1.3 Parameterization

Writing down the equation for the unit sphere is only a first step towards solving many problems involving spheres, such as, for example, computing the surface area of the unit sphere. Often the second step is to *solve the equation*. Now, for an equation like $x^2 = 1$, we can specify the set of all solutions by writing it out: $\{-1, 1\}$. But for $x^2 + y^2 + z^2 = 1$, there are clearly infinitely many solutions, and we cannot possibly write them all down.

What we can do, however, is to *parameterize* the solution set. Let us go through an example before formalizing this fundamental notion. Better yet, let us start with something even simpler: Consider the equation

$$x^2 + y^2 = 1 \quad (1.2)$$

in the x,y plane. (The x,y plane is the set of points (x, y, z) with $z = 0$.) You recognize (1.2) as the equation for the unit circle in the x,y plane. Recall the trigonometric identity

$$\cos^2 \theta + \sin^2 \theta = 1 . \quad (1.3)$$

Thus for all θ , the points

$$(x, y) = (\cos \theta, \sin \theta)$$

solve the equation (1.2).

Conversely, consider any solution (x, y) of (1.2). From the equation, $-1 \leq x \leq 1$, and hence we may define

$$\theta := \begin{cases} \arccos(x) & y \geq 0 \\ -\arccos(x) & y < 0 \end{cases} \quad (1.4)$$

By the definition of the arccos function, $-\pi < \theta \leq \pi$. Since $\cos \theta$ is an even function of θ , it follows that $x = \cos \theta$, and then one easily sees that $y = \sin \theta$.

Thus, we have a *one-to-one* correspondence between the points in the interval $(-\pi, \pi]$ and the set of solutions of (1.2). The correspondence is given by the function

$$\theta \mapsto (\cos \theta, \sin \theta)$$

from $(-\pi, \pi]$ onto the unit circle. This is an example of a *parameterization*: As the *parameter* θ varies over $(-\pi, \pi]$, $(\cos \theta, \sin \theta)$ varies over the unit circle, covering each point for exactly one value of the parameter θ .

Since the function in (1.4) is one-to-one and onto, it is invertible. The inverse is simply the map

$$(x, y) \mapsto \theta \quad (1.5)$$

where for x and y solving $x^2 + y^2 = 1$, θ is given by (1.4). The function in (1.5) is called the *angular coordinate function* on the unit circle. As you see in this example, finding a parameterization of the solution set of some equation and finding a system of coordinates on the solutions set are two aspects of the same thing. We will make formal definitions later; for now let us continue with examples.

For $r > 0$,

$$x^2 + y^2 = r^2$$

is the equation of the centered circle of radius r in the x, y plane. Since

$$x^2 + y^2 = r^2 \iff \left(\frac{x}{r}\right)^2 + \left(\frac{y}{r}\right)^2 = 1,$$

we can easily transform our parameterization of the unit circle into a parameterization of the circle of radius r : The parameterization is given by

$$\theta \mapsto (r \cos \theta, r \sin \theta) \quad (1.6)$$

while its inverse, the coordinate function, is given by

$$(x, y) \mapsto \theta := \begin{cases} \arccos\left(\frac{x}{\sqrt{x^2 + y^2}}\right) & y \geq 0 \\ -\arccos\left(\frac{x}{\sqrt{x^2 + y^2}}\right) & y < 0 \end{cases} \quad (1.7)$$

which specifies the angular coordinate as a function of x and y . Since $x^2 + y^2 = r^2 > 0$, we never divide by zero in this formula.

For our next example, let us parameterize the unit sphere; i.e., the solution set of (1.1). Note that $x^2 + y^2 + z^2 = 1$ implies that $-1 \leq z \leq 1$. Recalling (1.3) once more, we define

$$\phi = \arccos(z) , \quad (1.8)$$

so that $0 \leq \phi \leq \pi$, and $z = \cos \phi$.

It follows from (1.1) and (1.3) that $x^2 + y^2 = \sin^2 \phi$, and then, since $\sin \phi \geq 0$ for $0 \leq \phi \leq \pi$,

$$\sin \phi = \sqrt{x^2 + y^2} .$$

Evidently, for x and y not both zero, (x, y) lies on the circle of radius $\sin \phi$. We already know how to parameterize this: Setting $r = \sin \phi$ in (1.6), the parameterization function is

$$\theta \mapsto (\sin \phi \cos \theta, \sin \phi \sin \theta) = (x, y) .$$

Since (1.8) gives us $z = \cos \phi$, we combine results to obtain the parameterization

$$(\theta, \phi) \mapsto (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi) = (x, y, z) .$$

Define the three functions

$$x(\theta, \phi) := \sin \phi \cos \theta , \quad y(\theta, \phi) := \sin \phi \sin \theta \quad \text{and} \quad z(\theta, \phi) := \cos \phi . \quad (1.9)$$

Every solution (x, y, z) of the equation $x^2 + y^2 + z^2 = 1$ is of the form $(x(\theta, \phi), y(\theta, \phi), z(\theta, \phi))$ for some $(\theta, \phi) \in [-\pi, \pi] \times [0, \pi]$. Conversely, for all $(\theta, \phi) \in [-\pi, \pi] \times [0, \pi]$, $(x(\theta, \phi), y(\theta, \phi), z(\theta, \phi))$ solves the equation $x^2 + y^2 + z^2 = 1$.

Let S^2 denote the set of solutions to the equation $x^2 + y^2 + z^2 = 1$. That is, S^2 is the unit sphere. (This is a standard mathematical notation; the 2 in the exponent indicated that we are referring to the two dimensional sphere, a two dimensional surface in 3 dimensional space. In the same spirit, S^1 is a standard notation for the unit circle, a one dimensional curve in the plane. Later we will encounter S^n for arbitrary natural numbers n , and we will also give a mathematical definition of the term “dimension”, but the intuitively clear meaning in dimensions up to 3 will suffice for now. Note that just as we required only one parameter for S^1 , we require 2 parameters for S^2 .)

Define a function \mathbf{X} from $[-\pi, \pi] \times [0, \pi]$ to S^2 using the formula (1.9) by

$$\mathbf{X}(\theta, \phi) = (x(\theta, \phi), y(\theta, \phi), z(\theta, \phi)) . \quad (1.10)$$

Then \mathbf{X} maps $[-\pi, \pi] \times [0, \pi]$ onto S^2 . The function \mathbf{X} is not one-to-one, but only because at $(0, 0, 1)$, the “North pole”, and at $(0, 0, -1)$, the “South pole”, the value of θ is irrelevant. That is, when $\phi = 0$ or when $\phi = \pi$, the dependence of $\mathbf{X}(\theta, \phi)$ on θ drops out. However, the restriction of \mathbf{X} to the smaller domain $(-\pi, \pi] \times (0, \pi)$ is one-to-one and onto the “punctured sphere” that has the North Pole and South Pole removed. This restricted function is therefore invertible, and we already have formulas for the inverse: If $(x, y, z) \in S^2$, then θ is given by (1.7) and ϕ is given by (1.8). These formulas may be regarded as specifying the *coordinates* (θ, ϕ) of a point (x, y, z) of the punctured sphere.

We have now arrived at an important question: Are θ and ϕ coordinates, or are they parameters, or are they variables, and what is the difference? *The answer depends on the context.*

The function \mathbf{X} defined in (1.10) is the *parameterization function* from $(-\pi, \pi] \times (0, \pi)$ to the punctured sphere. Considered as *variables* in the domain of this function, θ and ϕ are *parameters*, and the function \mathbf{X} gives a description of the punctured sphere as a parametrized surface.

On the other hand, the variables θ and ϕ lie in the *range* of the inverse function \mathbf{X}^{-1} defined on the punctured sphere. In this context, $\theta(x, y, z)$ and $\phi(x, y, z)$ are the *coordinates* of a point $(x, y, z) \in S^2$, and \mathbf{X}^{-1} is the *coordinate function* on S^2 associated to the parameterization of S^2 that we have given. (There are many other ways to parameterize the sphere, or, what is effectively the same thing, to introduce coordinates on the sphere. We will encounter others later on.)

We have not yet given a formal definition of the terms *coordinate function* or *parameterization function*, only an example. But in all cases, like this one, the two functions are inverse to one another, and the coordinate functions map variables into a simple, usually “flat” set, here $(-\pi, \pi] \times (0, \pi)$, in a one-to-one manner onto an “interesting, more complicated” set, here the punctured sphere.

Now that we have a parameterization of the (punctured) sphere at hand, we can turn to the problem of finding a precise mathematical description of the tangent plane at a given point S^2 . We will proceed relying somewhat on geometric intuition and things you may know about equations for planes – namely that every plane in three dimensional space is the set of points whose coordinates (x, y, z) satisfy the equation

$$ax + by + cz = d$$

for some fixed set of numbers a, b, c and d . We also refrain from giving a precise definition of the term “tangent plane” at this point, but it is the natural generalization of the notion of a tangent line as the line that gives the “best fit” to the graph of a differentiable curve at a given point.

The North Pole is the point with coordinates $(0, 0, 1)$. The set of all point (x, y, z) such that $z = 1$ is a plane: It is the “horizontal plane” passing through $(0, 0, 1)$, and this is the tangent plane to S^2 at the North Pole; among all planes, it clearly gives the “best fit” to S^2 at the North Pole. Therefore, this is the equation of the tangent plane to S^2 at the North Pole. Likewise, the equation for the tangent plane at the South Pole is $z = -1$.

Now consider any other point (x_0, y_0, z_0) on S^2 . This naturally lies on the punctured sphere. Let (θ_0, ϕ_0) be the coordinates of this point so that with $\mathbf{X}(\theta, \phi)$ given by (1.9) and (1.10),

$$\mathbf{X}(\theta_0, \phi_0) = (x_0, y_0, z_0).$$

To be concrete, let us consider the case in which $\theta_0 = \phi_0 = \pi/4$, so that

$$(x_0, y_0, z_0) = (1/2, 1/2, 1/\sqrt{2}).$$

Let us consider values of the parameters θ and ϕ that are very close to θ_0 and ϕ_0 respectively. To write these conveniently, we introduce variables s and t by

$$s := \theta - \theta_0 \quad \text{and} \quad t := \phi - \phi_0.$$

Then $\theta = \theta_0 + s$ and $\phi = \phi_0 + t$. For small s and t , we have

$$\sin(\phi_0 + s) \approx \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}s \quad \text{and} \quad \sin(\theta_0 + t) \approx \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}t,$$

and we have

$$\cos(\phi_0 + s) \approx \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}s \quad \text{and} \quad \cos(\theta_0 + t) \approx \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}t.$$

These are simply the “tangent line” approximations to sin and cos at the relevant points, which are simply the first-order Taylor approximations.

Substitute these approximations into the formula (1.9) and discard all quadratic terms in s and t – remember, we are thinking of s and t as being very small, so that st , say, is negligibly small compared to either s or t . The result is:

$$\begin{aligned} x(\theta_0 + s, \phi_0 + t) &\approx \frac{1}{2} + \frac{1}{2}s - \frac{1}{2}t =: x(s, t) \\ y(\theta_0 + s, \phi_0 + t) &\approx \frac{1}{2} + \frac{1}{2}s + \frac{1}{2}t =: y(s, t) \\ z(\theta_0 + s, \phi_0 + t) &\approx \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}s =: z(s, t). \end{aligned}$$

The function

$$(s, t) \mapsto \left(\frac{1}{2} + \frac{1}{2}s - \frac{1}{2}t, \frac{1}{2} + \frac{1}{2}s + \frac{1}{2}t, \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}s \right) = (x(s, t), y(s, t), z(s, t))$$

is the parameterization of the plane in three dimensional space that “best fits” S^2 at the point $(x_0, y_0, z_0) = (1/2, 1/2, 1/\sqrt{2})$; it is the tangent plane at this point. Parameterization of planes is discussed thoroughly in the rest of this chapter, and we will see how to pass from a parameterization of a plane to an equation for the plane. For now, we simply give the answer: As you can check, for all s, t ,

$$\frac{1}{2}x(s, t) + \frac{1}{2}y(s, t) + \frac{1}{\sqrt{2}}z(s, t) = 1.$$

Therefore, the equation of the tangent plane can be written in the form $ax + by + cz = d$ with $a = b = 1/2$, $c = 1/\sqrt{2}$ and $d = 1$; it is

$$\frac{1}{2}x + \frac{1}{2}y + \frac{1}{\sqrt{2}}z = 1.$$

We have just completed our first tangent plane calculation, finding both a parameterization of the plane and an equation for it. Later, we will have more efficient methods to expedite such calculations, but there are several observations we can make now that will give a useful perspective on where we are headed:

- (1) Much of the work went into finding a parameterization of the sphere, and for this we used algebra and geometry, and not calculus *per se*.
- (2) The only calculus *per se* that we used was single variable calculus. We were ably to apply single variable methods “one variable at a time”.
- (3) Once we had a parameterization of the sphere, we readily obtained a parameterization of the tangent plane at (x_0, y_0, z_0) , but to get an equation for the tangent plane, there was still more geometry and algebra to deal with.

The short take-away from all of this is that we need some powerful algebraic and geometric tools to proceed efficiently. In the rest of this chapter, we develop those tools from the beginning. Before

proceeding, we close this section with a brief discussion of why computing equations for tangent planes is a useful thing to do.

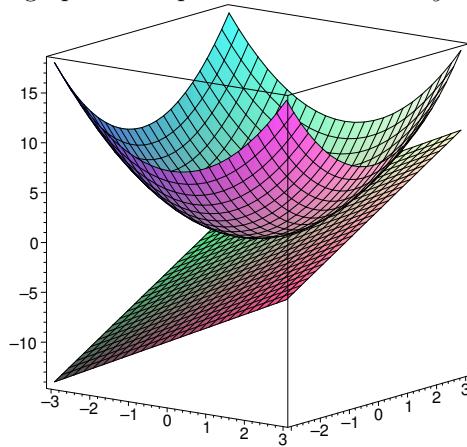
Geometry is very helpful in finding minima and maxima for functions of several variables. You learned in single variable calculus that if the tangent line to the graph of $y = f(x)$ is *not* horizontal at some point x_0 in the interior (a, b) of the interval $[a, b]$, then x_0 *cannot possibly* minimize or maximize the function f , even locally: Since the graph has a non-zero slope, you can move to higher or lower values by moving a little bit to either the left or to the right. Hence the only candidates for interior maxima and minima are the *critical points*; that is, points at which the tangent line to the graph is horizontal.

Now consider a very simple real valued function $f(x, y) = x^2 + y^2$. The graph of this function is the set of points (x, y, z) for which $z = f(x, y)$; i.e., $z = x^2 + y^2$. This graph is a parabolic surface in three dimensional space. It is parameterized by the function

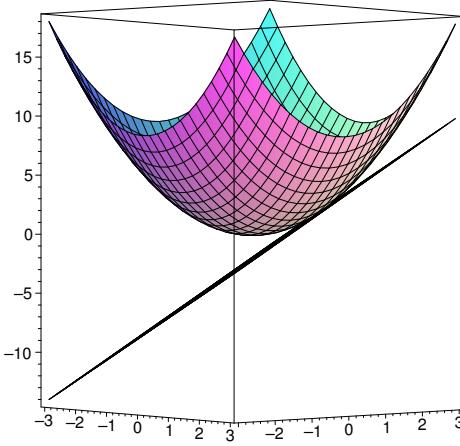
$$\mathbf{X}(x, y) = (x, y, x^2 + y^2).$$

At each point on the surface there is a *tangent plane*, which is the plane that “best fits” the graph at the point in a sense quite analogous to the sense in which that tangent line provides the “best fit” to the graph of a single variable differentiable function at a given point.

Here is a picture showing the portion of the graph of $z = x^2 + y^2$ for $-2 \leq x, y \leq 2$, together with the tangent plane to this graph at the point with $x = 1$ and $y = 1$.



Here is another picture of the same thing from a different vantage point, giving a better view of the point of contact:



As you can see, the tangent plane is tilted, so there are both uphill and downhill directions at this point, and so $(x, y) = (1, 1)$ cannot possibly minimize or maximize $f(x, y) = x^2 + y^2$ over any set U in the x, y plane that contains not only $(1, 1)$, but also all points sufficiently close to $(1, 1)$ – our reasoning requires some “wriggle-room” and does not apply if $(1, 1)$ is on the boundary of U .

Of course, for such a simple function, there are many ways to see this. However, for more interesting functions, this sort of reasoning in terms of tangent planes will be very useful, and it will lead to the following conclusion:

Let $f(x, y)$ be a function of x and y such for each x_0 and y_0 , there is a well-defined tangent plane to the graph $z = f(x, y)$ at (x_0, y_0) . Then if (x_0, y_0) maximizes or minimizes f for all (x, y) in some set U containing all points sufficiently close to (x_0, y_0) , then it is necessarily the case that the tangent plane to the graph $z = f(x, y)$ at (x_0, y_0) is horizontal, and we shall develop systematic methods for determining when this is the case.

Better yet, the same reasoning applies to functions $f(x_1, \dots, x_n)$ of arbitrarily many variables, except then we can no longer visualize the corresponding graphs and need a more algebraic or analytic interpretation of “horizontal”.

To make use of this sort of reasoning, we first need effective means of working with lines and planes that will allow us to express the words in the paragraph above in terms of equations which we can then solve. As indicated above, we not only need the two or three dimensional version of this, but a version that works for any number of dimensions – though in more than two variables it will be a “tangent hyperplane” that we will be computing.

There are many other subjects we shall study involving the calculus – both integral and differential – of functions that take vectors as input, or return them as output, or even both.

This concludes our necessarily somewhat vague look ahead in which we have balanced the goal of giving some perspective on where we are headed against the burden of providing too many technical definitions up front, which would make things precise, but then the trees would hide the forest.

We now shift gears. In the rest of this chapter, we proceed from the beginning with well-defined terms as we begin developing the algebraic and geometric tools that we shall use throughout the course.

1.1.4 The vector space \mathbb{R}^n

Definition 1 (Vectors in \mathbb{R}^n). *A vector is an ordered list of n numbers x_j , $j = 1, 2, \dots, n$, for some positive integer n , which is called the dimension of the vector. The integers $j = 1, 2, \dots, n$ that order the list are called the indices, and the corresponding numbers x_j are called the entries. That is, for each $j = 1, 2, \dots, n$, x_j is the j th entry on the list. The set of all n dimensional vectors is denoted by \mathbb{R}^n .*

Bold face is used to denote vectors, and $\mathbf{x} \in \mathbb{R}^n$ says that \mathbf{x} is a vector in \mathbb{R}^n . To write \mathbf{x} out in terms of its entries, we list the entries in a row, ordered left to right. The generic vector $\mathbf{x} \in \mathbb{R}^3$ is

$$\mathbf{x} = (x_1, x_2, x_3),$$

where x_1 , x_2 and x_3 are real numbers. When n is 2 or 3, it is often simpler to dispense with the subscripts, and distinguish the entries by using different letters. In this way, writing (x, y) to denote a generic vector in \mathbb{R}^2 or (x, y, z) to denote a generic vector in \mathbb{R}^3 . We use $\mathbf{0}$ to denote the vector in \mathbb{R}^n with 0 in every entry.

Finally, we shall often consider sets of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ in \mathbb{R}^n where the different vectors are distinguished by subscripts. A subscript on a boldface variable such as \mathbf{x}_j *always* indicates the j th vector in a list of vectors and not the j th entry of a vector \mathbf{x} . When we need to refer to the k th entry of \mathbf{x}_j , we shall write $(\mathbf{x}_j)_k$.

Definition 1 may not be what you expected. You may have seen two and three dimensional vectors defined as “quantities with length and direction”. When the term *vector* was coined, people had in mind the description of the position and motion of points in three dimensional physical space. For such vectors, the length and the direction have a clear geometric meaning.

But what about vectors like $(2, 1, 3, -1, 0, 2) \in \mathbb{R}^6$? What would we mean by the length of such a vector, and what would we mean by the angle between two vector in \mathbb{R}^6 ?

Perhaps surprisingly, there is a useful notion of length and direction in any number of dimensions.* But until we define direction and magnitude, we cannot use these notions to define vectors themselves! Therefore, the starting point is the definition of vectors in \mathbb{R}^n as ordered lists of n real numbers.

The *vector space* \mathbb{R}^n is more than just the set of all of the vectors in \mathbb{R}^n ; it is, by definition, this set *further equipped with a simple algebraic structure*, consisting of two algebraic operations: *scalar multiplication* and *vector addition*. As we have already stated, Descartes’ idea had such an enormous impact because it brought together what had been two quite separate branches of mathematics – algebra and geometry. Our plan for the rest of this section is to develop the algebraic aspects of Descartes’ idea, and then show how the algebra may be leveraged to apply our geometric intuition about three dimensional vectors to vectors of *any* dimension.

Definition 2 (Scalar Multiplication). *Given a number $a \in \mathbb{R}$ and a vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$, define the product of a and \mathbf{x} , denoted $a\mathbf{x}$, is defined by*

$$a\mathbf{x} = (ax_1, ax_2, \dots, ax_n).$$

*By “useful”, we mean useful for solving equations, among other things. In other words, useful in a practical sense, even in, say, eight dimensions.

For any vector \mathbf{x} , $-\mathbf{x}$ denotes the product of -1 and \mathbf{x} .

Example 1 (Multiplying numbers and vectors). Here are several examples:

$$2(-1, 0, 1) = (-2, 0, 2)$$

$$\pi(-1/2, 1/2) = (-\pi/2, \pi/2) = -(\pi/2, -\pi/2)$$

$$0(a, b, c) = (0, 0, 0) = \mathbf{0} .$$

Definition 3 (Vector Addition). Given two vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^n for some n , define their vector sum, $\mathbf{x} + \mathbf{y}$, by summing the corresponding entries:

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n) .$$

We define the vector difference of \mathbf{x} and \mathbf{y} , $\mathbf{x} - \mathbf{y}$ by $\mathbf{x} - \mathbf{y} = \mathbf{x} + (-\mathbf{y})$.

Note that vector addition does not mix up the entries of the vectors involved at all: For each j ,

$$(\mathbf{x} + \mathbf{y})_j = x_j + y_j .$$

The third entry, say, of the sum depends only on the third entries of the summands.

- For this reason, vector addition inherits the commutative and associative properties of addition in the real numbers. It is just the addition of real numbers “done in parallel”.

That is: vector addition is *commutative*, meaning that $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ and *associative*, meaning that $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$. In the same way, one sees that scalar multiplication *distributes* over vector addition:

$$a(\mathbf{x} + \mathbf{y}) = (a\mathbf{x}) + (a\mathbf{y}) \quad \text{and} \quad (a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x} .$$

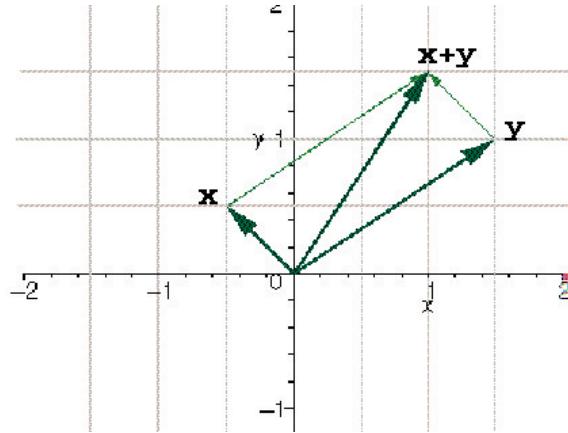
Example 2 (Vector addition).

$$(-3, 2, 5) + (1, 1, 1) = (-2, 3, 6)$$

$$(8, -2, 4, -12) + (0, 0, 0, 0) = (8, -2, 4, -12)$$

$$(8, -2, 4, -12) + (-8, 2, -4, 12) = (0, 0, 0, 0) = \mathbf{0} .$$

There is a geometric way to think about vector addition in \mathbb{R}^2 . Identify the vector $(x, y) \in \mathbb{R}^2$ with the point the Euclidean plane having these Cartesian coordinates. We can then represent this vector geometrically by drawing an arrow with its tail at the origin and its head at (x, y) . The following diagram shows three vectors represented this way: $\mathbf{x} = (-1/2, 1/2)$, $\mathbf{y} = (3/2, 1)$ and their sum, $\mathbf{x} + \mathbf{y} = (1, 3/2)$.



The vectors \mathbf{x} , \mathbf{y} and $\mathbf{x} + \mathbf{y}$ themselves are drawn in bold. There are also two arrows drawn more lightly: one is a parallel copy of \mathbf{x} “transported” so its tail is at the head of \mathbf{y} . The other is a parallel copy of \mathbf{y} “transported” so its tail is at the head of \mathbf{x} . These four arrows run along the sides of the parallelogram whose vertices are the origin, and the points corresponding to \mathbf{x} , \mathbf{y} and $\mathbf{x} + \mathbf{y}$. As you see, the arrow representing $\mathbf{x} + \mathbf{y}$ is the diagonal of this parallelogram that has its “tail end” at the origin.

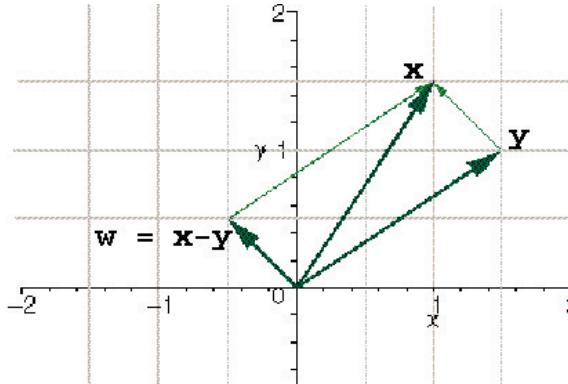
A similar diagram could be drawn for any pair of vectors and their sum, and you see that we can think of vector addition in the plane as corresponding to the following operation:

- Represent the vectors by arrows as in the diagram. Transport one arrow without turning it – that is, in a parallel motion – to bring its tail to the other arrow’s head. The head of the transported arrow is now at the point corresponding to the sum of the vectors.

Example 3 (Subtraction of vectors). Let \mathbf{x} and \mathbf{y} be two vectors in the plane \mathbb{R}^2 , and let $\mathbf{w} = \mathbf{x} - \mathbf{y}$. Then, using the associative and commutative properties of vector addition,

$$\mathbf{x} = \mathbf{x} + (\mathbf{y} - \mathbf{y}) = (\mathbf{x} - \mathbf{y}) + \mathbf{y} = \mathbf{y} + \mathbf{w} .$$

Using the same diagram, with the arrows labeled a bit differently, we see that $\mathbf{w} = \mathbf{x} - \mathbf{y}$ is the arrow running from the head of \mathbf{y} to the head of \mathbf{x} , “parallel transported” so that its tail is at the origin.



Definition 4 (Linear combination and Span). Let $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ be any set of m vectors in \mathbb{R}^n , for

any finite m . A linear combination of these vectors is any expression of the form

$$\sum_{j=1}^m t_j \mathbf{x}_j$$

where t_1, \dots, t_m are real numbers. (By the associative property of vector addition, this expression has an unambiguous meaning without any parentheses.)

Let $V \subset \mathbb{R}^n$ be any set of vectors in \mathbb{R}^n , finite or not. The span of V is the set of all possible linear combinations $\sum_{j=1}^m t_j \mathbf{x}_j$ formed using finite subsets $\{\mathbf{v}_1, \dots, \mathbf{v}_m\} \subset V$. The span of V is denoted by $\text{Span}(V)$.

Remark 1. The the subset symbol as used here simply mean the set on the left is a subset, not necessarily proper, of the set on the right. That is $A \subset B$ does not exclude $A = B$. Sometimes a more elaborate notation is used to distinguish proper and non-proper subsets, but that would be cumbersome here.

Let $V \subset \mathbb{R}^n$. For any $\mathbf{v} \in V$, we can write \mathbf{v} as a linear combination in the trivial way, taking $m = 1$ and $t_1 = 1$ so that $\mathbf{v} = 1\mathbf{v}$. This shows that $V \subset \text{Span}(V)$.

In general, given two subsets V and W of \mathbb{R}^n with $V \subset W$, every linear combination that one can form using vectors in V can also be formed using vectors in W since it includes V . That is

$$V \subset W \Rightarrow \text{Span}(V) \subset \text{Span}(W). \quad (1.11)$$

However, it can happen that when V is strictly contained in W , then $\text{Span}(V) = \text{Span}(W)$. For example, consider the case in which V is the singleton set $V = \{(1, 1, 1)\}$ in \mathbb{R}^3 . Then the

$$\text{Span}(V) = \{(t, t, t) : t \in \mathbb{R}\}.$$

which has infinitely many elements. Define $W = \text{Span}(V)$, and then certainly it is the case that V is a proper subset of W . However, by definition $\text{Span}(V) = W$, and it is also true that $\text{Span}(W) = W$. Any vector in $\text{Span}(W)$ is a linear combination of multiples of $(1, 1, 1)$, and therefore it is a multiple of $(1, 1, 1)$, and hence belongs to W . That is, W is its own span, and in this case $\text{Span}(V) = \text{Span}(W)$. The same reasoning leads to the following general conclusion:

Theorem 1. Let V be any subset of \mathbb{R}^n , and let $W = \text{Span}(V)$. Then $\text{Span}(W) = W$

Proof. By what we have noted above, $V \subset W$, and then by (1.11), $W = \text{Span}(V) \subset \text{Span}(W)$. To complete the proof, we need to show that $\text{Span}(W) \subset \text{Span}(V) = W$

Given two elements \mathbf{w}_1 and \mathbf{w}_2 of W , there are $m_1, m_2 \in \mathbb{N}$, $\mathbf{v}_1, \dots, \mathbf{v}_{m_1+m_2} \in V$ and $t_1, \dots, t_{m_1+m_2} \in \mathbb{R}$ so that $\mathbf{w}_1 = \sum_{j=1}^{m_1} t_j \mathbf{v}_j$ and $\mathbf{w}_2 = \sum_{j=m_1+1}^{m_1+m_2} t_j \mathbf{v}_j$. Therefore, for any $s_1, s_2 \in \mathbb{R}$,

$$s_1 \mathbf{w}_1 + s_2 \mathbf{w}_2 = \sum_{j=1}^{m_1} (s_1 t_j) \mathbf{v}_j + \sum_{j=m_1+1}^{m_1+m_2} (s_2 t_j) \mathbf{v}_j,$$

which is a linear combination of elements of V , and therefore belongs to W . This shows that any linear combination of two elements of W is again an element of W .

The general case follows by induction: Suppose that for $m \in \mathbb{N}$, any linear combination of m or fewer elements of W belong to W . Then the general linear combination of $m+1$ elements of W has the form

$$\sum_{j=1}^{m+1} s_j \mathbf{w}_j = 1 \left(\sum_{j=1}^m s_j \mathbf{w}_j \right) + s_{m+1} \mathbf{w}_{m+1} .$$

By the inductive hypothesis, $\sum_{j=1}^m s_j \mathbf{w}_j \in W$, and hence this $m+1$ term linear combination is also a linear combination of 2 elements of W , and hence it belongs to W .

This furnishes the inductive step and since we proved the hypothesis was valid for $m = 2$, this shows that for all $m \in \mathbb{N}$, any linear combination of m or fewer elements of W belong to W . \square

Example 4 (The span of 2 vectors in \mathbb{R}^3). Let $\mathbf{v}_1 = (1, 2, -3)$ and $\mathbf{v}_2 = (1, -2, 1)$. We have given a verbal definition of $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$. Can we find an equation that is satisfied by $\mathbf{v} = (x, y, z)$ if and only if $\mathbf{x} \in \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$? Can we parameterize the solutions set of this equation? Yes:

By definition a given vector \mathbf{v} satisfies $(x, y, z) \in \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$ if and only if for some numbers s and t ,

$$(x, y, z) = s\mathbf{v}_1 + t\mathbf{v}_2 = s(1, 2, -3) + t(1, -2, 1) = (s+t, 2s-2t, -3s+t) .$$

This single vector equation is equivalent to a system of 3 scalar equations for the unknowns s and t . (Recall that x , y , and z are given.)

$$x = s+t , \quad y = 2s-2t , \quad z = -3s+t ,$$

which give a parameterization of the plane spanned by \mathbf{v}_1 and \mathbf{v}_2 . Using the first two equations, $2x+y = (2s+2t)+(2s-2t) = 4s$, and hence $s = (2x+y)/4$. Likewise, using the same equations, $2x-y = 4t$, and hence $t = (2x-y)/4$. The first two equations have determined s and t , and then the third equation says that $z = t - 3s = \frac{2x-y}{4} - \frac{6x+3y}{4} = -x-y$. We have found our equation:

A vector $(x, y, z) \in \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$ if and only if

$$x + y + z = 0 . \tag{1.12}$$

Moreover, in this case $(x, y, z) = s\mathbf{v}_1 + t\mathbf{v}_2$ for a unique $(s, t) \in \mathbb{R}^2$, namely

$$(s, t) = \frac{1}{4}(2x+y, 2x-y) . \tag{1.13}$$

You may recognize the equation $z = -x - y$ as the equation of a plane in \mathbb{R}^3 passing through the origin, $\{\mathbf{0}\}$. In the next few sections we shall carefully study lines and planes in \mathbb{R}^3 . The fact that for each point (x, y, z) in this plane, there is exactly one choice of (s, t) so that $(x, y, z) = s\mathbf{v}_1 + t\mathbf{v}_2$ means that the function

$$\mathbf{x}(s, t) = s\mathbf{v}_1 + t\mathbf{v}_2 = (s+t, 2s-2t, -3s+t) ,$$

is a one-to-one function from \mathbb{R}^2 onto the solution set of the equation (1.12), which is the same as $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$. Thus, the function sending (s, t) to $\mathbf{x}(s, t) = s\mathbf{v}_1 + t\mathbf{v}_2$ is a parameterization of $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$.

Example 5 (The span of 3 vectors in \mathbb{R}^3). Let \mathbf{v}_1 and \mathbf{v}_2 be the vectors considered in the previous example, and now consider also a third vector $\mathbf{v}_3 = (-2, 1, 1)$. That is,

$$\mathbf{v}_1 = (1, 2, -3), \quad \mathbf{v}_2 = (1, -2, 1), \quad \mathbf{v}_3 = (-2, 1, 1).$$

What is $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\})$?

We now have more vectors to take linear combinations of than before, so certainly

$$\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}) \subset \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}). \quad (1.14)$$

However, it turns out that \mathbf{v}_3 is not really a “new” vector: $\mathbf{v}_3 \in \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$, to see this, notice that $(-2, 1, 1)$ satisfies the equation $x + y + z = 0$, and $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\})$ is the solution set of this equation, as we saw in the previous example. Moreover, by (1.13),

$$\mathbf{v}_3 = s\mathbf{v}_1 + t\mathbf{v}_2 \quad \text{where } (s, t) = -\frac{1}{4}(3, 5).$$

The general element in $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\})$ has the form $s\mathbf{v}_1 + t\mathbf{v}_2 + u\mathbf{v}_3$, but since $\mathbf{v}_3 = -\frac{3}{4}\mathbf{v}_1 - \frac{5}{4}\mathbf{v}_2$,

$$s\mathbf{v}_1 + t\mathbf{v}_2 + u\mathbf{v}_3 = (s - \frac{3}{4}u)\mathbf{v}_1 + (t - \frac{5}{4}u)\mathbf{v}_2 \in \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}).$$

This shows that for this choice of \mathbf{v}_1 , \mathbf{v}_2 and \mathbf{v}_3 ,

$$\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}) \subset \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}). \quad (1.15)$$

Together with (1.14), this proves that $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}) = \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$. (Note that the relation (1.14) is general, and would be valid for any set of 3 vectors, but the validity of the relation (1.15) depended on a special choice for \mathbf{v}_3 .

Another way to phrase all of this is that for given x, y, z , the system of three equations in the unknowns s, t, u

$$\begin{aligned} x &= s + t - 2u \\ y &= 2s - 2t + u \\ z &= -3s + t + u \end{aligned}$$

has a solution if and only if $x + y + z = 0$. We can say more: When the equations $x + y + z = 0$, there will be infinitely many solutions, one for each choice of the u .

Indeed, when $x + y + z = 0$, and $\mathbf{x} = (x, y, z)$, both \mathbf{x} and \mathbf{v}_3 belong to $\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\})$. By Theorem 1, for any $u \in \mathbb{R}$, so does the linear combination $\mathbf{x} - u\mathbf{v}_3$. Therefore, for some uniquely determined s and t , $\mathbf{x} - u\mathbf{v}_3 = s\mathbf{v}_1 + t\mathbf{v}_2$. Hence there is exactly one solution for each choice of the “free variable” u .

Now let us begin to connect the algebra we have developed in this subsection with Descartes’ ideas. The key is the introduction of the *standard basis for \mathbb{R}^n* :

Definition 5 (Standard basis for \mathbb{R}^n). For $j = 1, \dots, n$, let \mathbf{e}_j denote the vector in \mathbb{R}^n whose j th entry is 1, and all of whose remaining entries are 0. The ordered set $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is the standard basis for \mathbb{R}^n .

For example, if $n = 3$, we have

$$\mathbf{e}_1 = (1, 0, 0) \quad \mathbf{e}_2 = (0, 1, 0) \quad \text{and} \quad \mathbf{e}_3 = (0, 0, 1).$$

In only three dimensions, subscripts are often more of a hinderance than a help and a standard notation is $\mathbf{i} = \mathbf{e}_1$, $\mathbf{j} = \mathbf{e}_2$ and $\mathbf{k} = \mathbf{e}_3$.

Theorem 2 (Fundamental property of the standard basis). *Let $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ be the standard basis in \mathbb{R}^n . $\text{Span}(\{\mathbf{e}_1, \dots, \mathbf{e}_n\}) = \mathbb{R}^n$, and for each $\mathbf{x} \in \mathbb{R}^n$, there is exactly one vector of coefficients (t_1, \dots, t_n) such that*

$$\mathbf{x} = \sum_{j=1}^n t_j \mathbf{e}_j,$$

namely $(t_1, \dots, t_n) = (x_1, \dots, x_n)$.

Proof. By definition, $\sum_{j=1}^n x_j \mathbf{e}_j = x_1(1, 0, \dots, 0) + \dots + x_n(0, 0, \dots, 1) = (x_1, x_2, \dots, x_n)$.

Thus, any vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$ can be written as a linear combination of the standard basis vectors: Next, by the computation we just made, $\sum_{j=1}^n t_j \mathbf{e}_j = (t_1, t_2, \dots, t_n)$. Thus,

$$\mathbf{x} = \sum_{j=1}^n t_j \mathbf{e}_j \iff t_j = x_j \quad \text{for each } j = 1, \dots, n,$$

and hence the coordinates are uniquely determined. \square

The fact that every vector in \mathbb{R}^n can be expressed as a unique linear combination of the standard basis vectors is a special property of this set. As we have seen in Example 5, there are sets $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of 3 vectors in \mathbb{R}^3 such that the span of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is not all of \mathbb{R}^3 , and moreover, when \mathbf{x} is the span, so that it is possible to write $\mathbf{x} = t_1 \mathbf{v}_1 + t_2 \mathbf{v}_2 + t_3 \mathbf{v}_3$ for some t_1, t_2, t_3 , there are *infinitely many* choices of the coefficients t_1, t_2, t_3 .

The standard basis vectors thus provide the analog of a Cartesian frame for \mathbb{R}^n , in that one can get to any vector in \mathbb{R}^n by adding up multiples of the standard basis vectors, just as one can get to any point by moving along the directions of the vectors in the frame. However, frames were defined in terms of orthogonality, and so far we have no notion of geometry in \mathbb{R}^n , only algebra. The next subsection brings in the geometry.

1.1.5 Geometry and the dot product

So far, we have considered \mathbb{R}^n in purely algebraic terms. Indeed, the modern notion of an *abstract vector space* is a purely algebraic construct generalizing the algebraic structure on \mathbb{R}^n that has been the subject of the last subsection.

Additional structure is required to make contact with the notions of length and direction that have been traditionally associated to vectors. Let us begin with length. We shall identify the length of a vector with its “distance” from the origin. We introduce the notion of a *metric* which provides a

measure of the “distance” between two points, in some set. (The concept of a metric is very general, and in making the following definition, we do not assume that the set X is a set of vectors. It can be *any* set.)

Definition 6. Let X be a set. A function ϱ on the Cartesian product $X \times X$ with values in $[0, \infty)$ is a metric on X in case:

- (1) $\varrho(x, y) = 0$ if and only if $x = y$.
- (2) For all $x, y \in X$, $\varrho(x, y) = \varrho(y, x)$.
- (3) For all $x, y, z \in X$,

$$\varrho(x, z) \leq \varrho(x, y) + \varrho(y, z) .$$

When ϱ is a metric on X , the pair (X, ϱ) is called a metric space.

At an intuitive level, $\varrho(x, y)$ is supposed to represent the “distance” between x and y , or even more intuitively, the “length of the shortest path connecting x and y ”. This intuitive picture motivates the three items in the definition: Two points are the same if and only if there is no distance between them. This leads to (1). The distance from x to y is the same as the distance from y to x ; just “go back” on the same path. This leads to (2). Finally, if you insist on stopping by y on your way from x to z , the detour can only increase the total distance traveled. This leads to (3).

You might be able to think of some more requirements you would like to impose on the concept of distance. However, the mathematical value of a definition lies in the applicability of the theorems one can prove using it. It turns out that the definition of metric that we have just given provides a framework in which one can prove a great many very useful theorems. It is a very fruitful *abstraction* of the notion of distance in physical space. Our first example is the *Euclidean metric* on \mathbb{R}^n .

Definition 7 (Euclidean distance). The Euclidean length of a vector $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ is denoted by $\|\mathbf{x}\|$, and is defined by

$$\|\mathbf{x}\| = \left(\sum_{j=1}^n x_j^2 \right)^{1/2} .$$

The distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is defined by $\|\mathbf{x} - \mathbf{y}\|$. A vector $\mathbf{x} \in \mathbb{R}^n$ is called a unit vector in case $\|\mathbf{x}\| = 1$; i.e., in case \mathbf{x} has unit length.

We sometimes think of unit vectors as reprinting “pure directions”. Given any non-zero vector \mathbf{x} , we can define the unit vector $\mathbf{u} = \frac{1}{\|\mathbf{x}\|} \mathbf{x}$, which is called the *normalization* of \mathbf{x} , and then we have

$$\mathbf{x} = \|\mathbf{x}\| \left(\frac{1}{\|\mathbf{x}\|} \mathbf{x} \right) = \|\mathbf{x}\| \mathbf{u} .$$

This way of writing \mathbf{x} uses scalar multiplication to express it as the product of its length and direction.

As you can check from the definition, for any $\mathbf{x} \in \mathbb{R}^n$ and any $t \in \mathbb{R}$,

$$\|t\mathbf{x}\| = |t| \|\mathbf{x}\| .$$

As we shall soon see, the function ϱ_E defined by

$$\varrho_E(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\| \quad (1.16)$$

is a metric on \mathbb{R}^n , and is called the *Euclidean metric*. Indeed, with this definition of ϱ_E on \mathbb{R}^2 , the distance between two vectors (x, y) and (u, v) is $\sqrt{(x-u)^2 + (y-v)^2}$, which is of course the usual formula derived from the Pythagorean Theorem.

It is easy to see that the function ϱ_E defined in (1.16) satisfies requirements (1) and (2) in the definition of a metric. The fact that it also satisfies (3) is less transparent, but fundamentally important.

The first step towards this is to write $\|\mathbf{x}\|$ in terms of the dot product, which we now define:

Definition 8 (Dot product). *The dot product of two vectors $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ in \mathbb{R}^n , $\mathbf{a} \cdot \mathbf{b}$, is given by*

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n .$$

Note that the dot product is commutative, meaning that $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. This follows directly from the definition and the commutativity of multiplication in \mathbb{R} . In the same way one sees that the dot product *distributes* in the sense that for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, and all $s, t \in \mathbb{R}$,

$$(s\mathbf{a} + t\mathbf{b}) \cdot \mathbf{c} = s(\mathbf{a} \cdot \mathbf{c}) + t(\mathbf{b} \cdot \mathbf{c}) .$$

However, except when $n = 1$, it does not make sense to talk about the associativity of the dot product: For $n > 1$ and $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, $\mathbf{a} \cdot \mathbf{b} \notin \mathbb{R}^n$, so $(\mathbf{a} \cdot \mathbf{b}) \cdot \mathbf{c}$ is not defined.

From the definitions, we have $\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x}$. Therefore, using the distributive and commutative properties of the dot product,

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|^2 &= (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) \\ &= \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2\mathbf{x} \cdot \mathbf{y} , \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x} \cdot \mathbf{y} . \end{aligned} \quad (1.17)$$

The formula (1.17) has an interpretation in terms of the lengths of the vectors \mathbf{x} and \mathbf{y} , and the angle between these vectors. The key to this is the law of cosines. Recall that if the lengths of the three sides of a triangle in a Euclidean plane are A, B and C , and the angle between the sides with lengths A and B is θ , then $C^2 = A^2 + B^2 - 2AB \cos \theta$.

Now let \mathbf{x} and \mathbf{y} be any vectors in the plane \mathbb{R}^2 . Consider the triangle whose vertices are $\mathbf{0}$, \mathbf{x} and \mathbf{y} . Define the angle between \mathbf{x} and \mathbf{y} to be the angle between the two sides of the triangle issuing from the vertex at $\mathbf{0}$. Since the length of the side of the triangle opposite this vertex is $\|\mathbf{x} - \mathbf{y}\|$, by the law of cosines,

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos \theta .$$

Comparing this with (1.17), we conclude that $\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos \theta$. Therefore, in two dimensions, we have proved that the angle θ between two non-zero vectors \mathbf{x} and \mathbf{y} , considered as sides of a triangle in the plane \mathbb{R}^2 , is given by the formula

$$\theta = \arccos \left(\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \right) , \quad (1.18)$$

where the arccosine function is defined on $[-1, 1]$ with values in $[0, \pi]$.

The same sort of reasoning applies to vectors in \mathbb{R}^3 since any two non-collinear vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^3 lie in the plane determined by the three points $\mathbf{0}$, \mathbf{x} and \mathbf{y} , and then the law of cosines may be applied in this plane. Thus, the formula (1.18) is valid in \mathbb{R}^3 as well.

Why not go on from here? It may seem intuitively clear that just as in \mathbb{R}^3 , again in \mathbb{R}^n for any $n \geq 3$, any two non-collinear vectors \mathbf{x} and \mathbf{y} lie in a two dimensional plane in which we can apply the law of cosines, just as we did in \mathbb{R}^2 and \mathbb{R}^3 . This suggests that we use (1.18) to *define* the angle between two vectors in \mathbb{R}^n :

Definition 9. Let \mathbf{x} and \mathbf{y} be two non-zero vectors in \mathbb{R}^n . Then the angle θ between \mathbf{x} and \mathbf{y} is defined to be

$$\theta = \arccos \left(\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \right), \quad (1.19)$$

Two vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^n are orthogonal in case $\mathbf{x} \cdot \mathbf{y} = 0$.

There is an important matter to be checked before going forward: Does this definition make sense? The issue is that the arccos function is defined on $[-1, 1]$, so it had better be the case that

$$-1 \leq \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \leq 1$$

for all nonzero vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^n . This certainly is the case for $n = 2$ and $n = 3$, where we have proved the formula (1.18) is true with the classical definition of θ . but what about larger values of n ? The following theorem shows that there is no problem with using (1.19) to define θ no matter what the dimension n is. (It has many other uses as well!)

Theorem 3 (Cauchy–Schwarz inequality). For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\| \|\mathbf{b}\|. \quad (1.20)$$

There is equality in (1.20) if and only if $\|\mathbf{b}\| \mathbf{a} = \pm \|\mathbf{a}\| \mathbf{b}$

Proof. Clearly (1.20) is true, with equality, in case either of the vectors is the zero vector, and also in this case $\|\mathbf{b}\| \mathbf{a} = \|\mathbf{a}\| \mathbf{b} = 0$.

Hence we may assume that neither \mathbf{a} nor \mathbf{b} is the zero vector. Under this assumption, define $\mathbf{x} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$ and $\mathbf{y} = \frac{\mathbf{b}}{\|\mathbf{b}\|}$. Now let us compute $\|\mathbf{x} - \mathbf{y}\|^2$:

$$\|\mathbf{x} - \mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2\mathbf{x} \cdot \mathbf{y} = 2(1 - \mathbf{x} \cdot \mathbf{y}),$$

where we have used the fact that \mathbf{x} and \mathbf{y} are unit vectors. But since the left hand side is certainly non-negative, is must be the case that $\mathbf{x} \cdot \mathbf{y} \leq 1$.

Likewise, computing $\|\mathbf{x} + \mathbf{y}\|^2 = 2(1 + \mathbf{x} \cdot \mathbf{y})$, we see that $\mathbf{x} \cdot \mathbf{y} \geq -1$. Thus,

$$-1 \leq \mathbf{x} \cdot \mathbf{y} \leq 1,$$

which is equivalent to (1.20) by the definition of \mathbf{x} and \mathbf{y} .

As for the cases of equality, $|\mathbf{x} \cdot \mathbf{y}| = 1$ if and only if either $\|\mathbf{x} - \mathbf{y}\| = 0$ or $\|\mathbf{x} + \mathbf{y}\| = 0$, which means $\mathbf{x} = \pm \mathbf{y}$. From the definitions of \mathbf{x} and \mathbf{y} , this is the same as $\|\mathbf{b}\| \mathbf{a} = \pm \|\mathbf{a}\| \mathbf{b}$. \square

Theorem 4 (Triangle inequality). *For any three vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$,*

$$\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\| . \quad (1.21)$$

Proof: Let $\mathbf{a} = \mathbf{x} - \mathbf{y}$ and $\mathbf{b} = \mathbf{z} - \mathbf{y}$ so that $\mathbf{x} - \mathbf{z} = \mathbf{a} - \mathbf{b}$. Then by (1.17), and then the Cauchy-Schwarz inequality,

$$\begin{aligned} \|\mathbf{x} - \mathbf{z}\|^2 &= \|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a} \cdot \mathbf{b} \\ &\leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{a}\|\|\mathbf{b}\| \\ &= (\|\mathbf{a}\| + \|\mathbf{b}\|)^2 . \end{aligned}$$

Taking square roots of both sides, and recalling the definitions of \mathbf{a} and \mathbf{b} , we obtain (1.21). \square

Theorem 5. *The Euclidean distance function*

$$\varrho_E(\mathbf{x}, \mathbf{y}) := \sqrt{(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})} = \|\mathbf{x} - \mathbf{y}\|$$

is a metric on \mathbb{R}^n .

Proof. The function $\varrho_E(\mathbf{x}, \mathbf{y})$ is non-negative and clearly satisfies conditions (1) and (2) in the definition of a metric. By Theorem 4, it also satisfies (3). \square

1.1.6 Parallel and orthogonal components

Given a non-zero vector \mathbf{a} in \mathbb{R}^n , it is often useful to decompose other vectors $\mathbf{x} \in \mathbb{R}^n$ as a sum of a two vectors – one that is a multiple of \mathbf{a} , and another that is orthogonal to \mathbf{a} . There is exactly one way to do this. Let $\mathbf{a} \neq \mathbf{0}$ be given. For any $t \in \mathbb{R}$, we have

$$\mathbf{x} = t\mathbf{a} + (\mathbf{x} - t\mathbf{a}) \quad (1.22)$$

which is the general decomposition of \mathbf{x} into the sum of a multiple of \mathbf{a} , and some other vector. We now wish to choose t so that this second vector is orthogonal to \mathbf{a} . Taking the dot products of both sides of (1.22) with \mathbf{a} we find

$$\mathbf{x} \cdot \mathbf{a} = t\|\mathbf{a}\|^2 + (\mathbf{x} - t\mathbf{a}) \cdot \mathbf{a} ,$$

and thus $(\mathbf{x} - t\mathbf{a})$ is orthogonal to \mathbf{a} if and only if $t = \|\mathbf{a}\|^{-2}\mathbf{x} \cdot \mathbf{a}$. Using this value of t in (1.22), we obtain

$$\mathbf{x} = \frac{\mathbf{x} \cdot \mathbf{a}}{\|\mathbf{a}\|^2} \mathbf{a} + \left(\mathbf{x} - \frac{\mathbf{x} \cdot \mathbf{a}}{\|\mathbf{a}\|^2} \mathbf{a} \right) ,$$

which we can write more simply in terms of the unit vector $\mathbf{u} = \|\mathbf{a}\|^{-1}\mathbf{a}$:

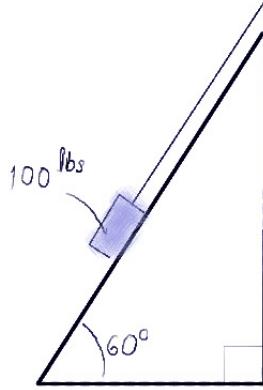
Definition 10 (Parallel and orthogonal components). *Given some non-zero vector $\mathbf{a} \in \mathbb{R}^n$, let $\mathbf{u} := \frac{1}{\|\mathbf{a}\|}\mathbf{a}$, which is the unit vector in the direction of \mathbf{a} . We can decompose any vector $\mathbf{x} \in \mathbb{R}^n$ into two pieces, \mathbf{x}_{\parallel} and \mathbf{x}_{\perp} where*

$$\mathbf{x}_{\parallel} := (\mathbf{x} \cdot \mathbf{u})\mathbf{u} \quad \text{and} \quad \mathbf{x}_{\perp} := \mathbf{x} - (\mathbf{x} \cdot \mathbf{u})\mathbf{u} .$$

These two vectors are called, respectively, the parallel and orthogonal components of \mathbf{x} with respect to \mathbf{a} , and $\mathbf{x} = \mathbf{x}_{\parallel} + \mathbf{x}_{\perp}$ is the decomposition of \mathbf{x} into its components parallel and perpendicular to \mathbf{a} .

The decomposition of vectors into parallel and orthogonal components is often useful. Here is a first example of this.

Example 6. A 100 pound weight sits on an slick (frictionless) incline making a 60 degree angle with the horizontal. It is held in place by a rope attached to the base of the weight and tied down at the top of the ramp. The tensile strength of the rope is such that it is only guaranteed not to break for tensions of no more than 80 pounds. Is this a dangerous situation?



To answer this we need to compute the tension in the rope. Let us use coordinates with the rope lying in the x, y plane, and the y axis being vertical. The gravitational force vector, measured in pounds, is

$$\mathbf{f} = (0, -100) .$$

The unit vector pointing down the slope is

$$\mathbf{u} = -(\cos(\pi/3), \sin(\pi/3)) = -\frac{1}{2}(1, \sqrt{3})$$

since 60 degrees is $\pi/3$ radians. The tension in the rope must balance the component of the gravitational force in the direction \mathbf{u} ; i.e., the direction of possible motion. That is, the magnitude of the tension will be $\|\mathbf{f}_\parallel\|$ where \mathbf{f}_\parallel is computed with respect to \mathbf{u} . Doing the computation we find

$$\mathbf{f}_\parallel = (\mathbf{f} \cdot \mathbf{u})\mathbf{u} = -25\sqrt{3}(1, \sqrt{3}) ,$$

and thus $\|\mathbf{f}_\parallel\| = 50\sqrt{3} \approx 86.6$. Look out below!

Here is another way to think about the computation in the previous example. The gravitational force vector \mathbf{f} has a simple expression in standard x, y coordinates, but these coordinates are not well adapted to the problem at hand since neither coordinate axis corresponds to a possible direction of possible motion. The direction of possible motion is given by \mathbf{u} .

Let us consider a coordinate system built around the direction of \mathbf{u} . We then take \mathbf{v} to be one of the two unit vectors in \mathbb{R}^2 that is orthogonal to \mathbf{u} , which are $\pm\frac{1}{2}(-\sqrt{3}, 1)$. We (arbitrarily) choose $\mathbf{v} = \frac{1}{2}(-\sqrt{3}, 1)$, and then as you can easily check, $\{\mathbf{u}, \mathbf{v}\}$ is an orthogonal pair of unit vectors.

Let us write the force vector \mathbf{f} in coordinates based on the $\{\mathbf{u}, \mathbf{v}\}$ frame of reference. That is,

$$\mathbf{f} = u\mathbf{u} + v\mathbf{v} \tag{1.23}$$

for some numbers u and v , which are the coordinates of \mathbf{f} with respect to this frame of reference. The u, v coordinates of \mathbf{f} are directly relevant to our problem. In particular u is the magnitude of the force in the direction u , and is what the tension in the rope must balance. Hence in these coordinates, our question becomes: *Is $|u| > 80$?*

To compute u (and v), we can take advantage of the orthogonality of \mathbf{u} and \mathbf{v} : Taking the dot product of both sides of $\mathbf{f} = u\mathbf{u} + v\mathbf{v}$ with \mathbf{u} we find

$$\mathbf{f} \cdot \mathbf{u} = (u\mathbf{u} + v\mathbf{v}) \cdot \mathbf{u} = u\mathbf{u} \cdot \mathbf{u} + v\mathbf{v} \cdot \mathbf{u} = u$$

where we have used the distributive property of the dot product and the orthogonality of \mathbf{u} and \mathbf{v} . In the same way, we find $\mathbf{f} \cdot \mathbf{v} = v$. Thus, we can re-write (1.23) as

$$\mathbf{f} = (\mathbf{f} \cdot \mathbf{u})\mathbf{u} + (\mathbf{f} \cdot \mathbf{v})\mathbf{v},$$

or in other words, $u = \mathbf{f} \cdot \mathbf{u}$ and $v = \mathbf{f} \cdot \mathbf{v}$. Now simple computations show that $|u| = |\mathbf{f} \cdot \mathbf{u}| > 80$.

- *The first step in solving many problems is to introduce a system of coordinates that is adapted to the problem, and in particular is “built out of” directions given in the problem.*

The most broadly useful and convenient coordinate systems in \mathbb{R}^n are those constructed using a set of n mutually orthogonal unit vectors, such as the set $\{\mathbf{u}, \mathbf{v}\}$ of orthogonal unit vectors in \mathbb{R}^2 that we have just used to build coordinates for our inclined plane problem. The next subsection develops this idea in general.

1.1.7 Orthonormal subsets of \mathbb{R}^n

Definition 11 (Orthonormal vectors in \mathbb{R}^n). *A set $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ of m vectors in \mathbb{R}^n is orthonormal in case for all $1 \leq i, j \leq m$*

$$\mathbf{u}_i \cdot \mathbf{u}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \quad (1.24)$$

Example 7. *The set of standard basis vectors $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is one very simple example of an orthonormal set. Also, any non-empty subset of an orthonormal set is easily seen to be orthonormal, so we can get other examples by taking non-empty subsets of $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. These are important examples, but not the only ones. Here is a more interesting example: Let*

$$\mathbf{u}_1 = \frac{1}{3}(1, 2, -2) \quad \mathbf{u}_2 = \frac{1}{3}(2, 1, 2) \quad \text{and} \quad \mathbf{u}_3 = \frac{1}{3}(2, -2, -1). \quad (1.25)$$

Then you can easily check that (1.24) is satisfied, so that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an orthonormal set in \mathbb{R}^3 .

The main theorem concerning orthonormal sets in \mathbb{R}^n is the following.

Theorem 6 (Fundamental Theorem on Orthonormal Sets in \mathbb{R}^n). *Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be any orthonormal set in \mathbb{R}^n consisting of exactly n vectors. Then every vector $\mathbf{x} \in \mathbb{R}^n$ can be written as a linear combination of the the vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ in exactly one way, namely*

$$\mathbf{x} = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j. \quad (1.26)$$

Moreover, the squared length of \mathbf{x} is the sum of the squares of the coefficients in this expansion:

$$\|\mathbf{x}\|^2 = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j)^2 .$$

The standard basis of \mathbb{R}^n , $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is a set of n orthonormal vectors in \mathbb{R}^n , and so the theorem says that

$$\mathbf{x} = \sum_{j=1}^n x_j \mathbf{e}_j = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{e}_j) \mathbf{e}_j$$

is the unique way to express any \mathbf{x} in \mathbb{R}^n as a linear combination of the standard basis vectors. This is the content of Theorem 2. Theorem 6 generalizes this to arbitrary sets of n orthonormal vectors in \mathbb{R}^n . It allows us to take *any* set of n orthonormal vectors in \mathbb{R}^n as the basis of a coordinate system in \mathbb{R}^n . This will prove to be *very useful* in practice. It will allow us to use coordinates that are especially adapted to whatever computation we are trying to make. The next definitions pave the way for this.

Definition 12 (Orthonormal basis). *An orthonormal basis in \mathbb{R}^n is any set of n orthonormal vectors in \mathbb{R}^n .*

The standard basis is one example of many. We have seen in Example 7 that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ with the vectors specified by (1.25) is an orthonormal basis for \mathbb{R}^3 .

Definition 13 (Coordinates with respect to an orthonormal basis). *Consider an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n , and a vector \mathbf{x} in \mathbb{R}^n . Then the numbers $\mathbf{x} \cdot \mathbf{u}_j$, $1 \leq j \leq n$, are called the coordinates of \mathbf{x} with respect to $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$.*

The coordinates of a vector $\mathbf{x} \in \mathbb{R}^n$ with respect to an orthonormal basis behave just like Cartesian coordinates in that they tell you how to get from $\mathbf{0}$ to \mathbf{x} : If the coordinates are y_1, \dots, y_n , start at $\mathbf{0}$, move y_1 units in the \mathbf{u}_1 direction, then move y_2 units in the \mathbf{u}_2 direction and so forth.

The hard part of the proof of Theorem 6 is contained in the following lemma:

Lemma 1 (No $n+1$ orthonormal vectors in \mathbb{R}^n). *There does not exist any set of $n+1$ orthonormal vectors in \mathbb{R}^n .*

The proof of this simple statement is very instructive, and very important, but somewhat involved. We give it in full in the next subsection. For now, let us take it on faith, and see how we may use it to prove Theorem 6. You will probably agree that the lemma is “geometrically obvious” in \mathbb{R}^2 , or even \mathbb{R}^3 , where you can easily visualize things.

Proof of Theorem 6: Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be any set of n orthonormal vectors in \mathbb{R}^n , and let \mathbf{x} be any non-zero vector in \mathbb{R}^n . Define the vector \mathbf{z} by

$$\mathbf{z} := \mathbf{x} - \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j .$$

for each $i = 1, \dots, n$, we have

$$\mathbf{z} \cdot \mathbf{u}_i = \left(\mathbf{x} - \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j \right) \cdot \mathbf{u}_i = \mathbf{x} \cdot \mathbf{u}_i - \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j \cdot \mathbf{u}_i = \mathbf{x} \cdot \mathbf{u}_i - \mathbf{x} \cdot \mathbf{u}_i = 0 .$$

Thus, \mathbf{z} is orthogonal to each \mathbf{u}_i .

Now suppose that $\mathbf{z} \neq \mathbf{0}$. Then we may define a unit vector \mathbf{u}_{n+1} by $\mathbf{u}_{n+1} = \frac{1}{\|\mathbf{z}\|}\mathbf{z}$. Since this unit vector is orthogonal to each unit vector in the orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, the augmented set $\{\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{u}_{n+1}\}$ would be a set of $n+1$ orthonormal vectors in \mathbb{R}^n . By Lemma 1, this is impossible. Therefore, $\mathbf{z} = \mathbf{0}$. By the definition of \mathbf{z} , this means that $\mathbf{x} = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j$, which is (1.26). Thus, every vector $\mathbf{x} \in \mathbb{R}^n$ can be written as a linear combination of the vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$. Moreover, there is only one way to do this, since if $\mathbf{x} = \sum_{j=1}^n y_j \mathbf{u}_j$, taking the dot product of both sides with \mathbf{u}_i yields

$$\mathbf{x} \cdot \mathbf{u}_i = \left(\sum_{j=1}^n y_j \mathbf{u}_j \right) \cdot \mathbf{u}_i = \sum_{j=1}^n y_j (\mathbf{u}_j \cdot \mathbf{u}_i) = y_i .$$

That is, each y_i must equal $\mathbf{x} \cdot \mathbf{u}_i$.

Finally, going back to (1.26), we compute

$$\begin{aligned} \|\mathbf{x}\|^2 &= \mathbf{x} \cdot \mathbf{x} &= \left(\sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j \right) \cdot \left(\sum_{k=1}^n (\mathbf{x} \cdot \mathbf{u}_k) \mathbf{u}_k \right) \\ &= \sum_{j,k=1}^n (\mathbf{x} \cdot \mathbf{u}_j)(\mathbf{x} \cdot \mathbf{u}_k) \mathbf{u}_j \cdot \mathbf{u}_k = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j)^2 . \end{aligned}$$

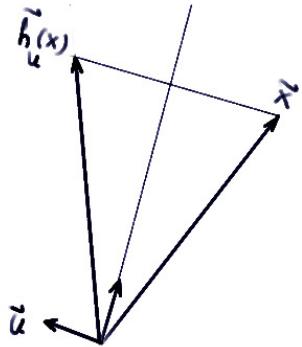
□

1.1.8 Householder reflections and orthonormal bases

In this subsection we shall prove Lemma 1. The proof is very interesting, and introduces many techniques and ideas that will be important later on.

We begin by introducing an extremely useful class of functions \mathbf{f} from \mathbb{R}^n to \mathbb{R}^n : the *Householder reflections*.

First, for $n = 2$, fix a unit vector $\mathbf{u} \in \mathbb{R}^2$ and consider the line $\ell_{\mathbf{u}}$ through the origin that is *orthogonal* to \mathbf{u} . Then, for any $\mathbf{x} \in \mathbb{R}^2$, define $\mathbf{h}_{\mathbf{u}}(\mathbf{x})$ to be the *mirror image* of \mathbf{x} across the line $\ell_{\mathbf{u}}$. That is, $\mathbf{h}_{\mathbf{u}}(\mathbf{x})$ is the *reflection* of \mathbf{x} across the line $\ell_{\mathbf{u}}$. Here is a picture illustrating the transformation from \mathbf{x} to $\mathbf{h}_{\mathbf{u}}(\mathbf{x})$:



The transformation from \mathbf{x} to $\mathbf{h}_\mathbf{u}(\mathbf{x})$ is geometrically well defined, and you could easily plot the output point $\mathbf{h}_\mathbf{u}(\mathbf{x})$ for any given input point \mathbf{x} . But to do computations, we need a formula. Let us derive a formula.

The key thing to realize, which you can see in the picture, is that both \mathbf{x} and $\mathbf{h}_\mathbf{u}(\mathbf{x})$ have the *same* component orthogonal to \mathbf{u} (that is, along the line $\ell_\mathbf{u}$) and have *opposite* components parallel to \mathbf{u} . In formulas, with respect to the direction \mathbf{u} ,

$$(\mathbf{h}_\mathbf{u}(\mathbf{x}))_\perp = \mathbf{x}_\perp \quad \text{and} \quad (\mathbf{h}_\mathbf{u}(\mathbf{x}))_\parallel = -\mathbf{x}_\parallel .$$

Therefore, since $\mathbf{h}_\mathbf{u}(\mathbf{x}) = (\mathbf{h}_\mathbf{u}(\mathbf{x}))_\perp + (\mathbf{h}_\mathbf{u}(\mathbf{x}))_\parallel$, we have the formula

$$\mathbf{h}_\mathbf{u}(\mathbf{x}) = \mathbf{x}_\perp - \mathbf{x}_\parallel . \quad (1.27)$$

Then since $\mathbf{x}_\perp = \mathbf{x} - (\mathbf{x} \cdot \mathbf{u})\mathbf{u}$ and $\mathbf{x}_\parallel = (\mathbf{x} \cdot \mathbf{u})\mathbf{u}$, we deduce the more explicit formula

$$\mathbf{h}_\mathbf{u}(\mathbf{x}) = \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{u})\mathbf{u} . \quad (1.28)$$

We have derived the formula (1.28) for $n = 2$. However, the formula does not explicitly involve the dimension and makes sense in any dimension. Now, given any unit vector $\mathbf{u} \in \mathbb{R}^n$, for any positive integer n , we use this formula to *define* the transformation $\mathbf{h}_\mathbf{u}$ from \mathbb{R}^n to \mathbb{R}^n that we shall call the *Householder reflection on \mathbb{R}^n in the direction \mathbf{u}* :

Definition 14 (Householder reflection in the direction \mathbf{u}). *For any unit vector $\mathbf{u} \in \mathbb{R}^n$, the function $\mathbf{h}_\mathbf{u}$ from \mathbb{R}^n to \mathbb{R}^n is defined by (1.28).*

Example 8. Let $\mathbf{u} = \frac{1}{\sqrt{3}}(-1, 1, -1)$. As you can check, this is a unit vector. Let $\mathbf{x} = (x, y, z)$ denote the generic vector in \mathbb{R}^3 . Let us compute $\mathbf{h}_\mathbf{u}(\mathbf{x})$. First, we find

$$\mathbf{x} \cdot \mathbf{u} = \frac{y - x - z}{\sqrt{3}} \quad \text{and hence} \quad (\mathbf{x} \cdot \mathbf{u})\mathbf{u} = \frac{y - x - z}{3}(-1, 1, -1) .$$

Notice that the square roots have (conveniently) gone away! Now, from the definition of $\mathbf{h}_\mathbf{u}$,

$$\mathbf{h}_\mathbf{u}(x, y, z) = (x, y, z) - \frac{2y - 2x - 2z}{3}(-1, 1, -1) = \frac{1}{3}(x + 2y - 2z, 2x + y + 2z, -2x + 2y + z) .$$

This is an example of a function, or, what is the same thing, transformation from \mathbb{R}^3 to \mathbb{R}^3 . If the input vector is (x, y, z) , the output vector is

$$\mathbf{h}_\mathbf{u}(x, y, z) = \left(\frac{x + 2y - 2z}{3}, \frac{2x + y + 2z}{3}, \frac{-2x + 2y + z}{3} \right) . \quad (1.29)$$

To conclude the example, let us evaluate the transformation at a particular vector. We choose, more or less at random, $\mathbf{x} = (1, 2, 3)$. Plugging this choice into our formula (1.29) we find

$$\mathbf{h}_\mathbf{u}(1, 2, 3) = \frac{1}{3}(-1, 10, 5) .$$

Let us compute the length of $\mathbf{h}_\mathbf{u}(\mathbf{x})$: $\|\mathbf{h}_\mathbf{u}(\mathbf{x})\| = \frac{1}{3}\|(-1, 10, 5)\| = \frac{1}{3}\sqrt{1 + 100 + 25} = \sqrt{14}$. Notice that $\|\mathbf{x}\| = \sqrt{1 + 4 + 9} = \sqrt{14}$, so $\|\mathbf{h}_\mathbf{u}(\mathbf{x})\| = \|\mathbf{x}\|$. This will always be the case, as we explain next – and as should be the case: reflection preserves the lengths of vectors – and more.

Lemma 2 (Householder reflections preserve dot products). *For any two vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^n , and any unit vector \mathbf{u} in \mathbb{R}^n ,*

$$(\mathbf{h}_\mathbf{u}(\mathbf{x})) \cdot (\mathbf{h}_\mathbf{u}(\mathbf{y})) = \mathbf{x} \cdot \mathbf{y} . \quad (1.30)$$

In particular,

$$\|\mathbf{h}_\mathbf{u}(\mathbf{x})\| = \|\mathbf{x}\| . \quad (1.31)$$

Proof. We use (1.27) and the fact that \mathbf{x}_\perp is orthogonal to \mathbf{y}_\parallel and that \mathbf{x}_\parallel is orthogonal to \mathbf{y}_\perp to compute:

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

and

$$\begin{aligned} \mathbf{x} \cdot \mathbf{y} &= (\mathbf{x}_\perp - \mathbf{x}_\parallel) \cdot (\mathbf{y}_\perp + \mathbf{y}_\parallel) \\ &= \mathbf{x}_\perp \cdot \mathbf{y}_\perp + \mathbf{x}_\parallel \cdot \mathbf{y}_\parallel + \mathbf{x}_\perp \cdot \mathbf{y}_\parallel + \mathbf{x}_\parallel \cdot \mathbf{y}_\perp \\ &= \mathbf{x}_\perp \cdot \mathbf{y}_\perp + \mathbf{x}_\parallel \cdot \mathbf{y}_\parallel , \end{aligned}$$

Comparing the two computations proves (1.30). Then (1.31) follows by considering the special case $\mathbf{y} = \mathbf{x}$. \square

In fact, since angles between vectors as well as the lengths of vectors are defined in terms of the dot product, Householder reflections preserve angles between vectors as well as the lengths of vectors, as you would expect from the diagram. In particular, *Householder reflections preserve orthogonality*.

Householder reflections are invertible transformations. *In fact, they are their own inverses:* For all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{h}_\mathbf{u}(\mathbf{h}_\mathbf{u}(\mathbf{x})) = \mathbf{x}$, That is, $\mathbf{h}_\mathbf{u} \circ \mathbf{h}_\mathbf{u}$ is the identity function.

To see this from the defining formula, we compute

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

That is, reflecting a vector twice (about the same direction) leaves you with the vector you started with.

Since reflection does not alter the length of a vector, if we are given vectors \mathbf{x} and \mathbf{y} with $\|\mathbf{x}\| \neq \|\mathbf{y}\|$, then we cannot possibly find a unit vector \mathbf{u} such that $\mathbf{h}_\mathbf{u}(\mathbf{x}) \neq \mathbf{y}$. However, if $\|\mathbf{x}\| = \|\mathbf{y}\|$, but $\mathbf{x} \neq \mathbf{y}$, then there is always a “standard” Householder reflection $\mathbf{h}_\mathbf{u}$ such that $\mathbf{h}_\mathbf{u}(\mathbf{x}) = \mathbf{y}$:

Lemma 3. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $n \geq 2$, satisfy $\|\mathbf{x}\| = \|\mathbf{y}\|$, but $\mathbf{x} \neq \mathbf{y}$. Then there is a unit vector $\mathbf{u} \in \mathbb{R}^n$ so that for the corresponding Householder reflection $\mathbf{h}_\mathbf{u}$,*

$$\mathbf{h}_\mathbf{u}(\mathbf{x}) = \mathbf{y} .$$

In particular, one may always choose

$$\mathbf{u} = \frac{1}{\|\mathbf{x} - \mathbf{y}\|} (\mathbf{x} - \mathbf{y}) . \quad (1.32)$$

Moreover, with this choice of \mathbf{u} , for any $\mathbf{z} \in \mathbb{R}^n$ that is orthogonal to both \mathbf{x} and \mathbf{y}

$$\mathbf{h}_{\mathbf{u}}(\mathbf{z}) = \mathbf{z} . \quad (1.33)$$

Proof. Define \mathbf{u} by (1.32), and compute

$$2(\mathbf{x} \cdot \mathbf{u})\mathbf{u} = \frac{2\mathbf{x} \cdot (\mathbf{x} - \mathbf{y})}{\|\mathbf{x} - \mathbf{y}\|^2}(\mathbf{x} - \mathbf{y}) . \quad (1.34)$$

Since $\|\mathbf{x}\| = \|\mathbf{y}\|$,

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x} \cdot \mathbf{y} = 2(\|\mathbf{x}\|^2 - \mathbf{x} \cdot \mathbf{y}) = 2\mathbf{x} \cdot (\mathbf{x} - \mathbf{y}) .$$

Therefore, from (1.34) we have $2(\mathbf{x} \cdot \mathbf{u})\mathbf{u} = \mathbf{x} - \mathbf{y}$, and so $\mathbf{h}_{\mathbf{u}}(\mathbf{x}) = \mathbf{x} - (\mathbf{x} - \mathbf{y}) = \mathbf{y}$.

The final part is simple: If \mathbf{z} is orthogonal to both \mathbf{x} and \mathbf{y} , then it is orthogonal to \mathbf{u} , and then (1.33) follows from the definition of $\mathbf{h}_{\mathbf{u}}$. \square

Example 9. Let $\mathbf{x} = \frac{1}{3}(1, 2, -2)$ and $\mathbf{y} = \mathbf{e}_1 = (1, 0, 0)$. These are both unit vectors, and hence $\|\mathbf{x}\| = \|\mathbf{y}\|$, so there is a unit vector \mathbf{u} such that $\mathbf{h}_{\mathbf{u}}(\mathbf{x}) = \mathbf{y}$, and \mathbf{u} is given by (1.32). We compute \mathbf{u} :

$$\mathbf{x} - \mathbf{y} = \frac{1}{3}(1, 2, -2) - (1, 0, 0) = \frac{1}{3}[(1, 2, -2) - (3, 0, 0)] = \frac{2}{3}(-1, 1, -1) .$$

Normalizing, we find

$$\mathbf{u} = \frac{1}{\sqrt{3}}(-1, 1, -1) .$$

Now simple computations verify that, as claimed, $\mathbf{h}_{\mathbf{u}}(\mathbf{x}) = \mathbf{y}$.

We are now ready to prove Lemma 1, which says that there does not exist any orthonormal set of $n + 1$ vectors in \mathbb{R}^n .

Proof of Lemma 1. First observe that for $n = 1$, there are exactly two unit vectors, namely (1) and (-1) . Since these vectors are not mutually orthogonal, there are exactly two orthonormal sets in \mathbb{R}^1 , namely $\{(1)\}$ and $\{(-1)\}$, and each consists of exactly one vector. This proves the Lemma for $n = 1$.

We now proceed by induction. For any $n \geq 2$ we suppose it is proved that there does not exist any orthonormal set of n vectors in \mathbb{R}^{n-1} . We shall show that then there does not exist any orthonormal set of $n + 1$ vectors in \mathbb{R}^n .

Suppose on the contrary that $\{\mathbf{u}_1, \dots, \mathbf{u}_{n+1}\}$ is an orthonormal set of vectors in \mathbb{R}^n . Then there exists an orthonormal set $\{\mathbf{v}_1, \dots, \mathbf{v}_{n+1}\}$ of vectors in \mathbb{R}^n such that $\mathbf{v}_{n+1} = \mathbf{e}_n$. To see this, note that if $\mathbf{u}_{n+1} = \mathbf{e}_n$, we already have the desired orthonormal set. Otherwise, by Lemma 3 there exists a unit vector in \mathbb{R}^n such that $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_{n+1}) = \mathbf{e}_n$. Then, since Householder reflections preserve lengths and orthogonality, if we define $\mathbf{v}_j = \mathbf{h}_{\mathbf{u}}(\mathbf{u}_j)$, $j = 1, \dots, n + 1$ $\{\mathbf{v}_1, \dots, \mathbf{v}_{n+1}\}$ is also an orthonormal set in \mathbb{R}^n , and by construction, $\mathbf{v}_{n+1} = \mathbf{e}_n$.

Therefore, for each $j = 1, \dots, n$,

$$0 = \mathbf{v}_j \cdot \mathbf{v}_{n+1} = \mathbf{v}_j \cdot \mathbf{e}_n .$$

Since $\mathbf{v}_j \cdot \mathbf{e}_n$ is simply the final entry of \mathbf{v}_j , this means that for each $j = 1, \dots, n$, \mathbf{v}_j has the form

$$\mathbf{v}_j = (\mathbf{w}_j, 0)$$

where \mathbf{w}_j is a unit vector in \mathbb{R}^{n-1} .

Since $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is orthonormal in \mathbb{R}^n , $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ is orthonormal in \mathbb{R}^{n-1} , since the final zero coordinate simply “goes along for the ride”. However, this is impossible, since we know that there does not exist any orthonormal set of n vectors in \mathbb{R}^{n-1} . We arrived at this contradiction by assuming that there existed an orthonormal set of $n+1$ vectors in \mathbb{R}^n . Hence this must be false. \square

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

In this section we shall study the geometry of lines and planes in \mathbb{R}^3 . We shall see that if we use *coordinates based on a well-chosen chosen orthonormal basis*, it is very easy to compute many geometric quantities such as, for example, the distance between two lines in \mathbb{R}^3 . Of course, to do this, we need a systematic method for constructing orthonormal bases. In \mathbb{R}^3 , the *cross product* provides such a method, and has many other uses as well. In the next subsection, we introduce the cross product, starting from a question about area that the cross product is designed to answer.

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

Let \mathbf{a} and \mathbf{b} be two vectors in \mathbb{R}^3 , neither a multiple of the other, and consider the triangle with vertices at $\mathbf{0}, \mathbf{a}, \mathbf{b}$, which naturally lies in the plane through these three points. The cross product gives the answer to the following question, and a number of other geometric questions as well:

- How can we express the area of this triangle in terms of the Cartesian coordinates of \mathbf{a} and \mathbf{b} ?

The classical formula for the area of a triangle in a plane is that it is one half the length of the base times the height. Let us take the side running from $\mathbf{0}$ to \mathbf{a} as the base, so that the length of the base is $\|\mathbf{a}\|$. Then, using θ to denote the angle between \mathbf{a} and \mathbf{b} , the height is $\|\mathbf{b}\| \sin \theta$. Thus, the area A of the triangle is

$$A := \frac{1}{2} \|\mathbf{a}\| \|\mathbf{b}\| \sin \theta .$$

(Note that since, by definition, $\theta \in [0, \pi]$ $\sin \theta \geq 0$.)

Using the identity $\sin^2 \theta + \cos^2 \theta = 1$, we can write this as

$$4A^2 = \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \sin^2 \theta = \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 (1 - \cos^2 \theta) = \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2 .$$

Now calculate the right hand side, taking $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$.

We find, after a bit of algebra,

$$\|\mathbf{a}\|^2 \|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2 = (a_2 b_3 - a_3 b_2)^2 + (a_3 b_1 - a_1 b_3)^2 + (a_1 b_2 - a_2 b_1)^2 .$$

The square root of the right hand side is the twice area in question. Notice that the right hand side is also square of the length of a vector in \mathbb{R}^3 , namely, the vector $\mathbf{a} \times \mathbf{b}$, defined as follows:

Definition 15 (Cross product in \mathbb{R}^3). Let \mathbf{a} and \mathbf{b} be two vectors in \mathbb{R}^3 . Then the cross product of \mathbf{a} and \mathbf{b} is the vector $\mathbf{a} \times \mathbf{b}$ where

$$\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2) \mathbf{e}_1 + (a_3 b_1 - a_1 b_3) \mathbf{e}_2 + (a_1 b_2 - a_2 b_1) \mathbf{e}_3 . \quad (1.35)$$

Example 10. Computing from the definition, we find

$$\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3 \quad \mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1 \quad \text{and} \quad \mathbf{e}_3 \times \mathbf{e}_1 = \mathbf{e}_2 . \quad (1.36)$$

By the computations that led to the definition, we have that

$$\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\| \|\mathbf{b}\| \sin \theta .$$

This tells us the magnitude of $\mathbf{a} \times \mathbf{b}$. What is its direction? Before dealing with this geometric question, it will help to first establish a few algebraic properties of the cross product.

Notice from the defining formula (1.35) that

$$\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a} .$$

Thus the cross product is not commutative; instead, it is *anti-commutative*. In particular, for any $\mathbf{a} \in \mathbb{R}^3$,

$$\mathbf{a} \times \mathbf{a} = -\mathbf{a} \times \mathbf{a} = \mathbf{0} . \quad (1.37)$$

Also, introducing a third vector $\mathbf{c} = (c_1, c_2, c_3)$, we have from the definition that

$$\begin{aligned} & \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \\ = & [a_2(b_3 + c_3) - a_3(b_2 + c_2)] \mathbf{e}_1 + [a_3(b_1 + c_1) - a_1(b_3 + c_3)] \mathbf{e}_2 + [a_1(b_2 + c_2) - a_2(b_1 + c_1)] \mathbf{e}_3 \\ = & (a_2 b_3 - a_3 b_2) \mathbf{e}_1 + (a_3 b_1 - a_1 b_3) \mathbf{e}_2 + (a_1 b_2 - a_2 b_1) \mathbf{e}_3 \\ & + (a_2 c_3 - a_3 c_2) \mathbf{e}_1 + (a_3 c_1 - a_1 c_3) \mathbf{e}_2 + (a_1 c_2 - a_2 c_1) \mathbf{e}_3 \\ = & \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} . \end{aligned}$$

Thus, $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}$, which means that the cross product *distributes* over vector addition. From this identity and the anti-commutivity, we see that $(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}$; i.e., the distributivity holds on both sides of the product.

Finally, a similar but simpler proof shows that for any number t , $(t\mathbf{a}) \times \mathbf{b} = t(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (t\mathbf{b})$. We summarize our conclusions in a theorem:

Theorem 7 (Algebraic properties of the cross product). Let \mathbf{a} , \mathbf{b} and \mathbf{c} be any three vectors in \mathbb{R}^3 , and let t be any number. Then

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= -\mathbf{b} \times \mathbf{a} \\ \mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \\ (t\mathbf{a}) \times \mathbf{b} &= t(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (t\mathbf{b}) . \end{aligned}$$

The following identity relates the cross product and the dot product:

Theorem 8 (Triple product identity). *Let $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$. Then*

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = (\mathbf{b} \times \mathbf{c}) \cdot \mathbf{a} . \quad (1.38)$$

In other words, the triple product $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$ is unchanged when the vectors in it are cyclicly permuted.

Proof: We compute:

$$\begin{aligned} (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} &= (a_2 b_3 - a_3 b_2)c_1 + (a_3 b_1 - a_1 b_3)c_2 + (a_1 b_2 - a_2 b_1)c_3 \\ &= (b_2 c_2 - b_3 c_2)a_1 + (b_3 c_1 - b_1 c_3)a_2 + (b_1 c_2 - b_2 c_1)a_3 = (\mathbf{b} \times \mathbf{c}) \cdot \mathbf{a} . \end{aligned}$$

□

Since the dot product is commutative (1.38) is equivalent to

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

where the order of the vectors is kept the same, but the positions of the dot and cross products are switched. Therefore, by (1.37) and Theorem 8, for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$, $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{b} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{b}) = \mathbf{0}$. Likewise,

$$\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = (\mathbf{a} \times \mathbf{a}) \cdot \mathbf{b} = \mathbf{0} .$$

We have proved:

Theorem 9. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$. Then $\mathbf{a} \times \mathbf{b}$ is orthogonal to both \mathbf{a} and \mathbf{b} .*

Let \mathbf{v}_1 and \mathbf{v}_2 be two vectors such that neither is a multiple of the other. Then by Theorem 9, $\mathbf{v}_1 \times \mathbf{v}_2$ is a non-zero vector orthogonal to every vector of the form

$$s\mathbf{v}_1 + t\mathbf{v}_2 \quad s, t \in \mathbb{R} ,$$

which is to say that $\mathbf{a} := \mathbf{v}_1 \times \mathbf{v}_2$ is orthogonal to every vector in the plane in \mathbb{R}^3 determined by (passing through) the 3 points $\mathbf{0}$, \mathbf{v}_1 and \mathbf{v}_2 . In other words:

- The cross product $\mathbf{v}_1 \times \mathbf{v}_2$ gives the direction of the normal line to the plane through $\mathbf{0}$, \mathbf{v}_1 and \mathbf{v}_2 , provided these are non-collinear, so that they do determine a plane.

Theorem 9 says when $\mathbf{a} \times \mathbf{b} \neq \mathbf{0}$, the unit vector $\frac{1}{\|\mathbf{a} \times \mathbf{b}\|} \mathbf{a} \times \mathbf{b}$ is one of the two unit vectors orthogonal to the plane through $\mathbf{0}$, \mathbf{a} and \mathbf{b} . Which one is it?

Definition 16. An orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of \mathbb{R}^3 is a right-handed orthonormal basis in case $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3$ and is a left-handed orthonormal basis in case

$$\mathbf{u}_1 \times \mathbf{u}_2 = -\mathbf{u}_3 .$$

Note that every orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of \mathbb{R}^3 is either left-handed or right-handed, since $\mathbf{u}_1 \times \mathbf{u}_2$ must be a unit vector orthogonal to both \mathbf{u}_1 and \mathbf{u}_2 , so that $\pm \mathbf{u}_3$ are the only possibilities. Also note that the standard basis of \mathbb{R}^3 is right-handed by (1.36).

Theorem 10. Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be any orthonormal basis of \mathbb{R}^3 . Then

$$\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3 \iff \mathbf{u}_2 \times \mathbf{u}_3 = \mathbf{u}_1 \iff \mathbf{u}_3 \times \mathbf{u}_1 = \mathbf{u}_2 . \quad (1.39)$$

In particular, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is right handed if and only if any one of the identities in (1.39) is valid, and in that case, all of them are valid.

Proof: Suppose that $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3$. Then by Theorem 6,

$$\mathbf{u}_2 \times \mathbf{u}_3 = (\mathbf{u}_2 \times \mathbf{u}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{u}_2 \times \mathbf{u}_3 \cdot \mathbf{u}_2)\mathbf{u}_2 + (\mathbf{u}_2 \times \mathbf{u}_3 \cdot \mathbf{u}_3)\mathbf{u}_3 .$$

Since $\mathbf{u}_2 \times \mathbf{u}_3$ orthogonal to \mathbf{u}_2 and \mathbf{u}_3 , this reduces to

$$\mathbf{u}_2 \times \mathbf{u}_3 = (\mathbf{u}_2 \times \mathbf{u}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 . \quad (1.40)$$

By Theorem 8 and the hypothesis that $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3$, $\mathbf{u}_2 \times \mathbf{u}_3 \cdot \mathbf{u}_1 = \mathbf{u}_1 \times \mathbf{u}_2 \cdot \mathbf{u}_3 = \mathbf{u}_3 \cdot \mathbf{u}_3 = 1$. Then (1.40) becomes $\mathbf{u}_2 \times \mathbf{u}_3 = \mathbf{u}_1$. Summarizing, $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3 \Rightarrow \mathbf{u}_2 \times \mathbf{u}_3 = \mathbf{u}_1$. The same sort of computation also shows that $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3 \Rightarrow \mathbf{u}_3 \times \mathbf{u}_1 = \mathbf{u}_2$.

Thus, the first of the identities in (1.39) implies the other two. The same sort of computations, which are left to the reader, show each of them implies the other two, so that they are all equivalent. \square

Why is the distinction between right and left handed orthonormal bases useful? One consequence of Theorem 10 is that one can use a formula just like (1.35) to compute the Cartesian components of $\mathbf{a} \times \mathbf{b}$ in terms of the Cartesian components of \mathbf{a} and \mathbf{b} for *any* coordinate system based on a *any* right handed orthonormal basis, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, and not only the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

Theorem 11 (Invariance under change of basis). Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be any right handed orthonormal basis in \mathbb{R}^3 . For any vectors $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2, y_3)$, define $\mathbf{a} = x_1\mathbf{u}_1 + x_2\mathbf{u}_2 + x_3\mathbf{u}_3$ and $\mathbf{b} = y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + y_3\mathbf{u}_3$, so that \mathbf{x} and \mathbf{y} are the coordinate vectors of \mathbf{a} and \mathbf{b} with respect to the coordinate system based on the orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. Then

$$\mathbf{a} \times \mathbf{b} = (x_2y_3 - x_3y_2)\mathbf{u}_1 + (x_3y_1 - x_1y_3)\mathbf{u}_2 + (x_1y_2 - x_2y_1)\mathbf{u}_3 . \quad (1.41)$$

In other words, one can compute the cross product of \mathbf{a} and \mathbf{b} by computing the cross product of their coordinate vectors for any right handed orthonormal basis.

Proof. Simply expand $\mathbf{a} \times \mathbf{b} = (x_1\mathbf{u}_1 + x_2\mathbf{u}_2 + x_3\mathbf{u}_3) \times (y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + y_3\mathbf{u}_3)$ using Theorem 10, and the result is (1.41). \square

The next identity, for the cross product of three vectors, has many uses. For example, we shall use it later on to deduce Kepler's Laws from Newton's Universal Theory of Gravitation. It was for exactly this purpose that Lagrange proved the identity, though he stated it in a different form.

Theorem 12 (Lagrange's Identity). Let $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$. Then

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c} . \quad (1.42)$$

Proof. Assume that none of \mathbf{a} , \mathbf{b} or \mathbf{c} is the zero vector, since then the identity is trivial and both sides are $\mathbf{0}$. There is another useful reduction to make: Let \mathbf{b}_{\parallel} and \mathbf{b}_{\perp} be the components of \mathbf{b} that are parallel and orthogonal to \mathbf{c} respectively. Then since $\mathbf{b}_{\parallel} \times \mathbf{c} = \mathbf{0}$

$$\begin{aligned}\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \mathbf{a} \times ((\mathbf{b}_{\perp} + \mathbf{b}_{\parallel}) \times \mathbf{c}) \\ &= \mathbf{a} \times (\mathbf{b}_{\perp} \times \mathbf{c}) + \mathbf{a} \times (\mathbf{b}_{\parallel} \times \mathbf{c}) = \mathbf{a} \times (\mathbf{b}_{\perp} \times \mathbf{c}) .\end{aligned}\quad (1.43)$$

Likewise,

$$(\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c} = [(\mathbf{a} \cdot \mathbf{c})\mathbf{b}_{\perp} - (\mathbf{a} \cdot \mathbf{b}_{\perp})\mathbf{c}] + [(\mathbf{a} \cdot \mathbf{c})\mathbf{b}_{\parallel} - (\mathbf{a} \cdot \mathbf{b}_{\parallel})\mathbf{c}] ,$$

and since $\mathbf{b}_{\parallel} = t\mathbf{c}$ for some t , $[(\mathbf{a} \cdot \mathbf{c})\mathbf{b}_{\parallel} - (\mathbf{a} \cdot \mathbf{b}_{\parallel})\mathbf{c}] = t[(\mathbf{a} \cdot \mathbf{c})\mathbf{c} - (\mathbf{a} \cdot \mathbf{c})\mathbf{c}] = 0$. Therefore,

$$(\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c} = (\mathbf{a} \cdot \mathbf{c})\mathbf{b}_{\perp} - (\mathbf{a} \cdot \mathbf{b}_{\perp})\mathbf{c} .$$

Combining this with (1.43), we see that \mathbf{b}_{\parallel} makes no contribution to either side, and that Lagrange's identity is valid for \mathbf{a} , \mathbf{b} and \mathbf{c} if and only if it is valid for \mathbf{a} , \mathbf{b}_{\perp} and \mathbf{c} . Therefore, we may assume without loss of generality that \mathbf{b} is orthogonal to \mathbf{c} .

Proceeding under this assumption (and the original assumption that $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{c} \neq \mathbf{0}$), we define

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{b}\|}\mathbf{b} , \quad \mathbf{u}_2 = \frac{1}{\|\mathbf{c}\|}\mathbf{c} \quad \text{and} \quad \mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2 .$$

By the properties of the cross product, and the orthogonality of \mathbf{b} and \mathbf{c} , $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right handed orthonormal basis for \mathbb{R}^3 . We now compute

$$\mathbf{b} \times \mathbf{c} = (\|\mathbf{b}\|\mathbf{u}_1) \times (\|\mathbf{c}\|\mathbf{u}_2) = \|\mathbf{b}\|\|\mathbf{c}\|\mathbf{u}_1 \times \mathbf{u}_2 = \|\mathbf{b}\|\|\mathbf{c}\|\mathbf{u}_3 ,$$

where we have used Theorem 10 in the final step. Next, using Theorem 6, we write

$$\mathbf{a} = (\mathbf{a} \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{a} \cdot \mathbf{u}_2)\mathbf{u}_2 + (\mathbf{a} \cdot \mathbf{u}_3)\mathbf{u}_3 ,$$

and compute $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$. By our computation of $\mathbf{b} \times \mathbf{c}$ above, this gives

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \|\mathbf{b}\|\|\mathbf{c}\|[(\mathbf{a} \cdot \mathbf{u}_1)\mathbf{u}_1 \times \mathbf{u}_3 + (\mathbf{a} \cdot \mathbf{u}_2)\mathbf{u}_2 \times \mathbf{u}_3 + (\mathbf{a} \cdot \mathbf{u}_3)\mathbf{u}_3 \times \mathbf{u}_3] ,$$

and by Theorem 10 once more, this gives

$$\begin{aligned}\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \|\mathbf{b}\|\|\mathbf{c}\|[(\mathbf{a} \cdot \mathbf{u}_1)\mathbf{u}_2 + (\mathbf{a} \cdot \mathbf{u}_2)\mathbf{u}_1] \\ &= (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}\end{aligned}$$

which is Lagrange's identity. \square

The identity we get from Theorem 12 in the special case $\mathbf{a} = \mathbf{b}$ is often useful:

Corollary 1. *Let \mathbf{u} be any unit vector in \mathbb{R}^3 . Then for all $\mathbf{c} \in \mathbb{R}^3$*

$$\mathbf{c}_{\perp} = -\mathbf{u} \times (\mathbf{u} \times \mathbf{c}) \quad (1.44)$$

where \mathbf{c}_{\perp} is the component of \mathbf{c} orthogonal to \mathbf{u} .

One can rewrite (1.44) as $\mathbf{c}_\perp = (\mathbf{u} \times \mathbf{c}) \times \mathbf{u}$ which resembles the identity $\mathbf{c}_\parallel = (\mathbf{u} \cdot \mathbf{c})\mathbf{u}$.

Proof of Corollary 1. Applying (1.42) with $\mathbf{a} = \mathbf{b} = \mathbf{u}$, we obtain

$$\mathbf{u} \times (\mathbf{u} \times \mathbf{c}) = (\mathbf{c} \cdot \mathbf{u})\mathbf{u} - (\mathbf{u} \cdot \mathbf{u})\mathbf{c} = -(\mathbf{c} - (\mathbf{c} \cdot \mathbf{u})\mathbf{u}) = -\mathbf{c}_\perp .$$

Now using the anti-commutivity of the cross product, we obtain (1.44). \square

Thus, one can readily compute orthogonal components by computing cross products. The magnitude of \mathbf{c}_\perp is even simpler to compute: Since $\mathbf{u} \times \mathbf{c} = \mathbf{u} \times \mathbf{c}_\perp$, and since $\|\mathbf{u} \times \mathbf{c}_\perp\| = \|\mathbf{u}\|\|\mathbf{c}_\perp\|$,

$$\|\mathbf{c}_\perp\| = \|\mathbf{u} \times \mathbf{c}\| .$$

The cross product is not only useful for checking whether a given orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is right handed or not; it is useful for *constructing* such bases. The next example concerns a problem that often arises when working with lines and planes in \mathbb{R}^3 .

Example 11 (Constructing a right-handed orthonormal basis containing a given direction). *Given a nonzero vector $\mathbf{v} \in \mathbb{R}^3$, how can we find an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ in which \mathbf{u}_3 is a positive multiple of \mathbf{v} ?*

Here is one way to do it using the cross product. First, $\mathbf{u}_3 = \|\mathbf{v}\|^{-1}\mathbf{v}$ is the only unit vector that is a positive multiple of \mathbf{v} . Let us next choose \mathbf{u}_1 . This has to be some unit vector that is orthogonal to \mathbf{v} .

If $\mathbf{v} = (v_1, v_2, v_3)$ and $v_j = 0$ for some j , then $\mathbf{v} \cdot \mathbf{e}_j = v_j = 0$, and we may take $\mathbf{u}_1 = \mathbf{e}_j$ for this j . On the other hand, if each v_j is non-zero, define $\mathbf{w} := \mathbf{e}_3 \times \mathbf{v} = (-v_2, v_1, 0)$. This is orthogonal to \mathbf{v} by Theorem 9, and $\|\mathbf{w}\| \neq 0$. Now define $\mathbf{u}_1 = \frac{1}{\|\mathbf{w}\|}\mathbf{w}$. Finally, define $\mathbf{u}_2 = \mathbf{u}_3 \times \mathbf{u}_1$, so that \mathbf{u}_2 is orthogonal to both \mathbf{u}_3 and \mathbf{u}_1 , and is a unit vector. Thus, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an orthonormal basis, and since $\mathbf{u}_3 \times \mathbf{u}_1 = \mathbf{u}_2$, Theorem 10 says that this basis is right handed.

Now we do this for specific vectors. Let $\mathbf{v} = (2, 1, 2)$. We set $\mathbf{u}_3 = \frac{1}{\|\mathbf{v}\|}\mathbf{v} = \frac{1}{3}(2, 1, 2)$. Next, since none of the entries of \mathbf{v} is zero, we define $\mathbf{w} := \mathbf{e}_3 \times \mathbf{v} = (-1, 2, 0)$, and then

$$\mathbf{u}_1 := \frac{1}{\|\mathbf{w}\|}\mathbf{w} = \frac{1}{\sqrt{5}}(-1, 2, 0) ,$$

Finally we define

$$\mathbf{u}_2 := \mathbf{u}_3 \times \mathbf{u}_1 = \frac{1}{3\sqrt{5}}(-4, -2, 5) .$$

In the next section we shall see many uses of such constructions.

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

Let \mathbf{a} be a non-zero vector in \mathbb{R}^3 , and let \mathbf{x}_0 be *any* vector in \mathbb{R}^3 . Consider the two equations in the vector variable \mathbf{x} :

$$\mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) = 0 \tag{1.45}$$

$$\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0} \tag{1.46}$$

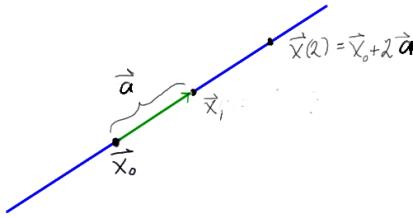
The solution sets of these equations are planes and lines, respectively, in \mathbb{R}^3 .

Let us start with (1.46). Note that $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ is true if and only if $\mathbf{x} - \mathbf{x}_0$ is a multiple of \mathbf{a} . But $\mathbf{x} - \mathbf{x}_0 = t\mathbf{a}$ can be written as $\mathbf{x} = \mathbf{x}_0 + t\mathbf{a}$. As t ranges over the real line, the function $\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{a}$ traces out a line in \mathbb{R}^3 , passing through \mathbf{x}_0 at time $t = 0$, and moving in the direction given by \mathbf{a} .

A line in \mathbb{R}^3 is determined by two distinct points. Given two distinct points \mathbf{x}_0 and \mathbf{x}_1 in \mathbb{R}^3 , define $\mathbf{a} = \mathbf{x}_1 - \mathbf{x}_0$, and then for all $t \in \mathbb{R}$, define

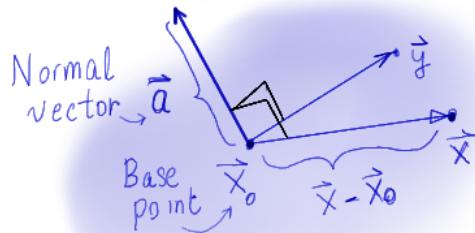
$$\mathbf{x}(t) := \mathbf{x}_0 + t\mathbf{a} :$$

As illustrated below, as t varies, $\mathbf{x}(t)$ varies over the set of points that one can reach starting from \mathbf{x}_0 , and moving only in the direction of \mathbf{a} (or its opposite).



Since it is clear that $\mathbf{a} \times (\mathbf{x}(t) - \mathbf{x}_0) = t\mathbf{a} \times \mathbf{a} = \mathbf{0}$ for all t , and since every solution of (1.46) equals $\mathbf{x}(t)$ for some uniquely determined value of t , the function sending t to $\mathbf{x}(t)$ is a parameterization of this line; i.e., the solution set of (1.46). We have just that if one has any one of: (1) a two-point description of a line, (2) a parameterization of a line, or (3), an equation specifying a line, it is very easy to find the other two.

Now let us turn to (1.45). A point $\mathbf{x} \in \mathbb{R}^3$ satisfies (1.45) if and only if $\mathbf{x} - \mathbf{x}_0$ is orthogonal to \mathbf{a} , and therefore to every multiple $t\mathbf{a}$ of \mathbf{a} . The line through $\mathbf{0}$ parameterized by $t\mathbf{a}$ is the *normal line* to the plane. The vector \mathbf{a} is sometimes called the *normal vector* to the plane, but keep in mind that \mathbf{a} can be replaced by any non-zero multiple of \mathbf{a} without changing the plane. Here is a diagram showing a plane with base point \mathbf{x}_0 , normal vector $t\mathbf{a}$, and two vectors \mathbf{x} and \mathbf{y} in the plane:



Example 12 (The equation of a plane specified in terms of a base point and normal vector). Consider the plane passing through the point $\mathbf{x}_0 = (3, 2, 1)$ that is orthogonal to the vector $\mathbf{a} = (1, 2, 3)$. Then $\mathbf{x} = (x, y, z)$ belongs to this plane if and only if $(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{a} = 0$. Doing the computations,

$$((x, y, z) - (3, 2, 1)) \cdot (1, 2, 3) = 0 \iff x + 2y + 3z = 10 .$$

The equation $x + 2y + 3z = 10$ is an equation specifying this plane, but written without using the dot product. It may look simpler, but the geometry is not so explicit in this form.

As seen in the last example, an equation specifying a plane can also be written in the form $\mathbf{a} \cdot \mathbf{x} = d$. Indeed, defining $d := \mathbf{a} \cdot \mathbf{x}_0$,

$$(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{a} = 0 \iff \mathbf{a} \cdot \mathbf{x} = d .$$

Example 13 (Parameterizing a plane specified by an equation). Consider the equation $\mathbf{a} \cdot \mathbf{x} = d$ where $\mathbf{a} = (1, 2, 1)$, and $d = 10$. Writing $\mathbf{x} = (x, y, z)$, the equation becomes $x + 2y + z = 10$.

Parameterizing solution sets means solving equations. Solving equations means eliminating variables. Let us eliminate z :

$$z = 10 - 2y - x .$$

Thus, $\mathbf{x} = (x, y, z)$ belong to the plane if and only if

$$\mathbf{x} = (x, y, 10 - 2y - x) = (0, 0, 10) + x(1, 0, -1) + y(0, 1, -2) .$$

We have expanded the left hand side, and collected terms into a constant vector, a second constant vector times x and a third constant vector times y . The point of this is that defining

$$\mathbf{x}_0 := (0, 0, 10) , \quad \mathbf{v}_1 := (1, 0, -1) \quad \text{and} \quad \mathbf{v}_2 := (0, 1, -2) ,$$

a vector \mathbf{x} belongs to the plane if and only if it can be written in the form $\mathbf{x}_0 + x\mathbf{v}_1 + y\mathbf{v}_2$ for some numbers x and y . Moreover, whenever \mathbf{x} can be written in the form $\mathbf{x}_0 + x\mathbf{v}_1 + y\mathbf{v}_2$, the numbers x and y are uniquely determined, since direct computation yields

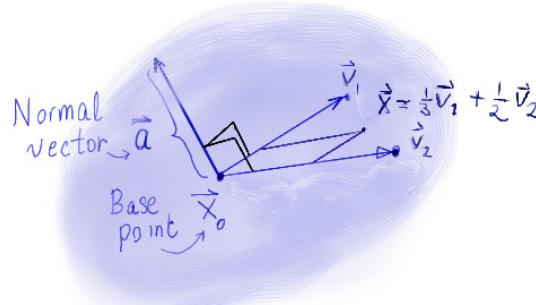
$$\mathbf{x}_0 + s\mathbf{v}_1 + t\mathbf{v}_2 = (s, t, 10 - 2s - t) .$$

If this is to equal (x, y, z) , we must have $s = x$ and $y = t$.

Thus there is a one-to-one correspondence between points \mathbf{x} in the plane and vectors $(s, t) \in \mathbb{R}^2$ given by

$$\mathbf{x}(s, t) = s\mathbf{v}_1 + t\mathbf{v}_2 .$$

As (s, t) varies over \mathbb{R}^2 , $\mathbf{x}(s, t)$ varies over the plane in question in a one-to one way. Thus, we have parameterized the plane.



Two distinct points determine a line. Any three points that are *not collinear*, i.e., do not all lie on one line, determine a plane: There is a unique plane in \mathbb{R}^3 that contains these three points.

Example 14 (When do four points belong to one plane?). Consider the points

$$\mathbf{p}_1 = (1, 2, 3) \quad \mathbf{p}_2 = (3, 2, 1) \quad \mathbf{p}_3 = (1, 3, 2) \quad \text{and} \quad \mathbf{p}_4 = (4, -1, 3) . \quad (1.47)$$

Do all of these points lie in the same plane?

It is easy to answer this question once we know the equation for the plane determined by the first three points: Simply plug the fourth point into this equation. If the equation is satisfied, the point is on the plane, and otherwise it is not.

To find the equation, choose \mathbf{p}_1 as the base point, and define

$$\mathbf{x}_0 := (1, 2, 3) \quad \mathbf{v}_1 := \mathbf{p}_2 - \mathbf{x}_0 = (2, 0, -2) \quad \text{and} \quad \mathbf{v}_2 := \mathbf{p}_3 - \mathbf{x}_0 = (0, 1, -1) .$$

Writing the equation of the plane in the form $\mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$, and plugging in $\mathbf{p}_2 = \mathbf{x}_0 + \mathbf{v}_1$ and $\mathbf{p}_3 = \mathbf{x}_0 + \mathbf{v}_2$, we have

$$\mathbf{a} \cdot \mathbf{v}_1 = 0 \quad \text{and} \quad \mathbf{a} \cdot \mathbf{v}_2 = 0 .$$

Thus \mathbf{a} must be orthogonal to \mathbf{v}_1 and \mathbf{v}_2 . We get such a vector by taking the cross product of \mathbf{v}_1 and \mathbf{v}_2 . Thus we define

$$\mathbf{a} := \mathbf{v}_1 \times \mathbf{v}_2 = (2, 0, -2) \times (0, -1, 1) = (2, 2, 2) .$$

We then compute $\mathbf{a} \cdot \mathbf{x} = 2x + 2y + 2z$ and $\mathbf{a} \cdot \mathbf{x}_0 = 12$ so the equation $\mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ written out in terms of x , y and z is $2x + 2y + 2z = 12$, or, what is the same thing,

$$x + y + z = 6 . \quad (1.48)$$

This is the equation for the plane passing through the first three points in the list (1.47). You should check that these points do satisfy the equation.

We can now easily decide whether \mathbf{p}_4 lies in the same plane as the first three points. With $x = 4$, $y = -1$ and $z = 3$, the equation (1.48) is satisfied, so it is in the plane. It is left as an exercise for the reader to prove that three points \mathbf{p}_1 , \mathbf{p}_2 and \mathbf{p}_3 are not collinear if and only if $(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1) \neq \mathbf{0}$, as in this example.

Consider again the general equation for a line: $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$, which can always be written as $\mathbf{a} \times \mathbf{x} = \mathbf{a} \times \mathbf{x}_0$. Since \mathbf{x}_0 and \mathbf{a} are both given, we can define $\mathbf{d} = \mathbf{a} \times \mathbf{x}_0$, and with this definition, (1.46) is equivalent to

$$\mathbf{a} \times \mathbf{x} = \mathbf{d} . \quad (1.49)$$

However, for general $\mathbf{d} \in \mathbb{R}^3$, the solution set of (1.49) need not be a line; in fact, it can be empty.

Example 15 (The equation $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ as a system of equations). Let $\mathbf{a} = (1, -2, 1)$ and $\mathbf{d} = (d_1, d_2, d_3)$. Computing $\mathbf{a} \times \mathbf{x}$ for $\mathbf{x} = (x, y, z)$, we find

$$(1, -2, 1) \times (x, y, z) = (-y - 2z, x - z, 2x + y) . \quad (1.50)$$

Hence, $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ is equivalent to the system of three equations

$$\begin{aligned} -y - 2z &= d_1 \\ x - z &= d_2 \\ 2x + y &= d_3 . \end{aligned} \quad (1.51)$$

We can also write (1.51), and hence (1.50), as

$$x\mathbf{v}_1 + y\mathbf{v}_2 + z\mathbf{v}_3 = \mathbf{d} \quad (1.52)$$

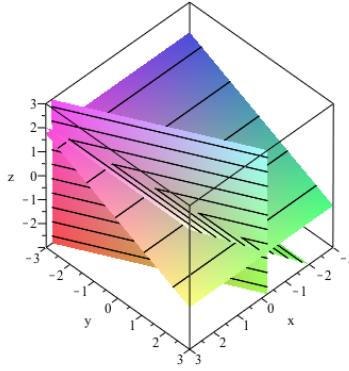
where $\mathbf{v}_1 = (0, 1, 2)$, $\mathbf{v}_2 = (-1, 0, 1)$ and $\mathbf{v}_3 = (-2, -1, 0)$.

Notice that $\mathbf{a} \cdot \mathbf{v}_j = 0$ for $j = 1, 2, 3$. From this it is easy to see, as in Example 4 and Example 5 that

$$\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}) = \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}) = \{(x, y, z) : x - 2y + z = 0\} .$$

That is, the span of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is precisely the set of vectors that are orthogonal to \mathbf{a} . Therefore, (1.52) has a solution if and only if $\mathbf{a} \cdot \mathbf{d} = 0$. In this case, the third equation in (1.51) is redundant, and the solution set of all three equations is the same as the solution set of the first two. That is, when the solution set of (1.51) is not empty, it is equal to the intersection of the two planes described by its first two equations. In this case, if \mathbf{x}_0 is *any* vector in the solution set, $\mathbf{a} \times \mathbf{x}_0 = \mathbf{d}$. Choosing any such vector, we can rewrite the equation $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ as $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$. As we have seen, the solution set of this equation is always a line, and therefore the intersection of the two planes discussed above is exactly this line.

Now take $\mathbf{d} = (-1/2, 1, 5/2)$, which is orthogonal to \mathbf{a} . Here is a plot showing the three planes intersecting in the line described by $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ for this choice of \mathbf{d} :



However, each pair of these planes intersects in the same line; to specify the line we only need *any* two of the equations. The third equation is redundant.

While for any non-zero $\mathbf{a} \in \mathbb{R}^3$ and any number d , $\mathbf{a} \cdot \mathbf{x} = d$ is *always* the equation of a plane, as we have seen in the last example, the analog of this is not true for $\mathbf{a} \times \mathbf{x} = \mathbf{d}$: The left hand side is orthogonal to \mathbf{a} for all \mathbf{x} , and so if the right hand side is not orthogonal to \mathbf{a} , there can be no solution. That is, when $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{d} \cdot \mathbf{a} \neq 0$, the solution set of $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ is the empty set.

On the other hand, when \mathbf{d} is orthogonal to \mathbf{a} , Lagrange's identity tells us

$$\mathbf{a} \times (\mathbf{a} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{d})\mathbf{a} - (\mathbf{a} \cdot \mathbf{a})\mathbf{d} = -\|\mathbf{a}\|^2\mathbf{d}.$$

Therefore, if we define $\mathbf{x}_0 = -\|\mathbf{a}\|^{-2}\mathbf{a} \times \mathbf{d}$, we have that $\mathbf{d} = \mathbf{a} \times \mathbf{x}_0$, and then $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ is equivalent to $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$, which is the equation of a line. Summarizing: *For all non-zero $\mathbf{a} \in \mathbb{R}^3$, the equation $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ specifies a line if and only if $\mathbf{d} \cdot \mathbf{a} = 0$, in which case the equation is equivalent to $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ where $\mathbf{x}_0 = -\|\mathbf{a}\|^{-2}\mathbf{a} \times \mathbf{d}$.*

Example 16 (Solving $\mathbf{a} \times \mathbf{x} = \mathbf{d}$). Let $\mathbf{a} = (1, -2, 1)$ and $\mathbf{d} = (-1/2, 1, 5/2)$. We check that $\mathbf{a} \cdot \mathbf{d} = 0$, and hence the solutions set of $\mathbf{a} \times \mathbf{x} = \mathbf{d}$ is a line. We then compute $\|\mathbf{a}\|^2 = 6$ and

$$\mathbf{a} \times \mathbf{d} = (-6, -3, 0).$$

Hence

$$\mathbf{x}_0 := -\frac{1}{\|\mathbf{a}\|^2}\mathbf{a} \times \mathbf{d} = -\frac{1}{6}(-6, -3, 0) = (1, 1/2, 0).$$

Thus, the solution set is the line parameterized by

$$\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{a} = (-1, 1/2, 0) + t(1, -2, 1) = (1+t, 1/2-2t, t).$$

It is easy to pass from the equation specifying a line in the form $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ to a system of equations

$$\begin{aligned} \mathbf{b}_1 \cdot \mathbf{x} &= c_1 \\ \mathbf{b}_2 \cdot \mathbf{x} &= c_2 \end{aligned} \tag{1.53}$$

that specify the line as an intersection of two planes. By the triple product identity, for any \mathbf{v} ,

$$\mathbf{v} \cdot (\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0)) = (\mathbf{v} \times \mathbf{a}) \cdot (\mathbf{x} - \mathbf{x}_0).$$

The vector $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0)$ is orthogonal to \mathbf{a} no matter what \mathbf{x} is. Therefore, let $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ be an orthonormal basis of \mathbb{R}^3 in which $\mathbf{v}_3 = \|\mathbf{a}\|^{-1}\mathbf{a}$. Then $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ if and only if $\mathbf{v}_j \cdot (\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0)) = 0$ for $j = 1, 2$ since this is automatically the case for $j = 3$.

Hence $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ is equivalent to the system of equations

$$\begin{aligned} (\mathbf{v}_1 \times \mathbf{a}) \cdot (\mathbf{x} - \mathbf{x}_0) &= 0 \\ (\mathbf{v}_2 \times \mathbf{a}) \cdot (\mathbf{x} - \mathbf{x}_0) &= 0 \end{aligned}$$

which is the same as the system (1.53) if we define $\mathbf{b}_j = \mathbf{v}_j \times \mathbf{a}$ and $c_j = \mathbf{b}_j \cdot \mathbf{x}_0$ for $j = 1, 2$.

Conversely, given a pair of equations such as (1.53) with $\mathbf{b}_1, \mathbf{b}_2$ non-zero, define $\mathbf{a} = \mathbf{b}_1 \times \mathbf{b}_2$. Suppose that there exists some solution \mathbf{x}_0 . This need not be the case: If the two planes are parallel and distinct, they do not intersect. However, as long as one solution \mathbf{x}_0 can be found, every point of the form $\mathbf{x}_0 + t\mathbf{a}$ satisfies both equations in (1.53). Moreover, if \mathbf{x} is any solution of both equations in (1.53), then $\mathbf{x} - \mathbf{x}_0$ is orthogonal to both \mathbf{b}_1 and \mathbf{b}_2 , and hence has the form $t\mathbf{a}$ for some $t \in \mathbb{R}$. Thus, the solution set of (1.53) is precisely the line of vectors of the form $\mathbf{x}_0 + t\mathbf{a}$, $t \in \mathbb{R}$, and this is the line described by the equation $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$.

It is also easy to pass from a system such as (1.53) to a parameterization of the line they describe – when the planes intersect in a line – without using the cross product.

Example 17 (Parameterizing a line given by a pair of equations for planes). *Consider*

$$\begin{aligned} x + 2y + 3z &= 2 \\ 3x + 2y + z &= 4 \end{aligned}$$

Use the first equation to eliminate some variable, say x : $x = 2 - 2y - 3z$. Substituting this into the second equation, it becomes

$$3(2 - 2y - 3z) + 2y + z = 4 \quad \text{which yields } y = \frac{1}{2} - 2z. \quad (1.54)$$

This expresses y as a function of z . Go back to $x = 2 - 2y - 3z$, and use (1.54) to eliminate y , thus to expressing x as a function of z alone. We obtain $x = 1 + z$, and we have $y = 1/2 - 2z$.

Substituting these into $\mathbf{x} = (x, y, z)$, we see that \mathbf{x} belongs to the line if and only if

$$\mathbf{x} = (1 + z, 1/2 - 2z, z) = (1, 1/2, 0) + z(1, -2, 1).$$

Defining $\mathbf{x}_0 = (1, 1/2, 0)$ and $\mathbf{a} = (1, -2, 1)$, $\mathbf{x}(t) := \mathbf{x}_0 + t\mathbf{a}$ is a parameterization of the line, and $\mathbf{a} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ is the equation for it.

There are infinitely many ways to parameterize a given line with this scheme: Any point on the line can serve as the base point \mathbf{x}_0 , and any non-zero vector that runs parallel to the line can serve as the direction vector. The same is true of planes; any point in the plane can serve as the base point.

1.2.3 Distance problems

Consider a line in \mathbb{R}^3 given in parametric form by $\mathbf{x}(u) = \mathbf{x}_0 + u\mathbf{a}$. Let \mathbf{p} be any point in \mathbb{R}^3 . It turns out that there is a unique point \mathbf{q} on the line that comes closer to \mathbf{p} than any other point on

the line. That is, there is some $u_0 \in \mathbb{R}$ such that

$$\|\mathbf{x}(u_0) - \mathbf{p}\| < \|\mathbf{x}(u) - \mathbf{p}\| \quad (1.55)$$

for all $u \neq u_0$. Then, by definition, the number $\|\mathbf{x}(u_0) - \mathbf{p}\|$ is the *distance from \mathbf{p} to the line*.

To see that there exists a $u_0 \in \mathbb{R}$ so that (1.55) is true, note that we may replace \mathbf{a} by $\mathbf{u} = \|\mathbf{a}\|^{-1}\mathbf{a}$ in $\mathbf{x}(u)$ without changing the line itself; this is just another parameterization of the same line. Hence we may assume without loss of generality that $\mathbf{x}(u) = \mathbf{x}_0 + u\mathbf{u}$ and that $\|\mathbf{u}\|$ is a unit vector. Now decompose $\mathbf{x}(u) - \mathbf{p}$ into its components parallel and orthogonal to \mathbf{u} : By the Pythagorean Theorem,

$$\begin{aligned} \|\mathbf{x}(u) - \mathbf{p}\|^2 &= \|(\mathbf{x}(u) - \mathbf{p})_{\parallel}\|^2 + \|(\mathbf{x}(u) - \mathbf{p})_{\perp}\|^2 \\ &= ((\mathbf{x}(u) - \mathbf{p}) \cdot \mathbf{u})^2 + \|(\mathbf{x}(u) - \mathbf{p}) \times \mathbf{u}\|^2. \end{aligned}$$

Now note that

$$(\mathbf{x}(u) - \mathbf{p}) \cdot \mathbf{u} = ((\mathbf{x}_0 - \mathbf{p}) + u\mathbf{u}) \cdot \mathbf{u} = (\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u} + u,$$

and

$$(\mathbf{x}(u) - \mathbf{p}) \times \mathbf{u} = ((\mathbf{x}_0 - \mathbf{p}) + u\mathbf{u}) \times \mathbf{u} = (\mathbf{x}_0 - \mathbf{p}) \times \mathbf{u},$$

which is *independent of u* . Therefore,

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

The second term on the right is independent of u , while the first is clearly minimized by choosing $u = u_0$ with $u_0 = -(\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}$. Therefore, $\mathbf{q} = \mathbf{a}(u_0) = \mathbf{x}_0 + ((\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u})\mathbf{u}$. Identifying the line as the solution set of the equation $\mathbf{u} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$, we have proved:

Theorem 13. *Let \mathbf{u} be a unit vector in \mathbb{R}^3 and let \mathbf{x}_0 and \mathbf{p} be any two points in \mathbb{R}^3 . There is a unique point \mathbf{q} in the line given by $\mathbf{u} \times (\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ such that $\|\mathbf{q} - \mathbf{p}\| < \|\mathbf{x} - \mathbf{p}\|$ for all other \mathbf{x} in the line. Moreover, $\|\mathbf{q} - \mathbf{p}\| = \|(\mathbf{x}_0 - \mathbf{p}) \times \mathbf{u}\|$ and $\mathbf{q} = \mathbf{x}_0 + ((\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u})\mathbf{u}$.*

Remark 2. *We do not need to compute a cross product to compute $\|\mathbf{q} - \mathbf{p}\|$: By the Pythagorean Theorem,*

$$\|(\mathbf{x}_0 - \mathbf{p}) \times \mathbf{u}\|^2 = \|\mathbf{x}_0 - \mathbf{p}\|^2 - \|(\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}\|^2. \quad (1.56)$$

Example 18. *Consider the line parameterized by $\mathbf{x}_0 + t\mathbf{v}$ with $\mathbf{x}_0 = (2, -2, 3)$ and $\mathbf{v} = (1, 2, 1)$. Let $\mathbf{p} = (1, 2, 3)$. What is the distance from \mathbf{p} to this line? We obtain \mathbf{u}_3 by normalizing \mathbf{v} , and then we apply Theorem 13 and (1.56), finding*

$$\|(\mathbf{p} - \mathbf{x}_0)_{\perp}\| = \left(\|(-1, 4, 0)\|^2 - \frac{1}{6}[(-1, 4, 0) \cdot (1, 2, 1)]^2 \right)^{1/2} = \left(17 - \frac{49}{6} \right)^{1/2}.$$

There is a similar result concerning the distance between a point and a plane: Let \mathbf{u} be any unit vector in \mathbb{R}^3 , and let \mathbf{x}_0 and \mathbf{p} be any vectors in \mathbb{R}^3 . Consider the plane that is the solution set of the equation $\mathbf{u} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$. Again, there is a unique point \mathbf{q} in the plane such that $\|\mathbf{q} - \mathbf{p}\| < \|\mathbf{x} - \mathbf{p}\|$ for all other \mathbf{x} in the plane. We call $\|\mathbf{q} - \mathbf{p}\|$ the distance from \mathbf{p} to the plane.

To see this, let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be a right-handed orthonormal basis of \mathbb{R}^3 such that $\mathbf{u}_3 = \mathbf{u}$. Then $\mathbf{x}(s, t) = \mathbf{x}_0 + s\mathbf{u}_1 + t\mathbf{u}_2$ is a parameterization of the plane. By the Pythagorean Theorem, and the definition $\mathbf{x}(s, t) = \mathbf{x}_0 + s\mathbf{u}_1 + t\mathbf{u}_2$,

$$\begin{aligned}\|\mathbf{x}(s, t) - \mathbf{p}\|^2 &= \sum_{j=1}^3 (\mathbf{x}(s, t) - \mathbf{p}) \cdot \mathbf{u}_j)^2 \\ &= \sum_{j=1}^3 ((\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_j + (s\mathbf{u}_1 + t\mathbf{u}_2) \cdot \mathbf{u}_j)^2 + \\ &= ((\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_1 + s)^2 + ((\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_2 + t)^2 + ((\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_3)^2.\end{aligned}$$

Evidently we minimize $\|\mathbf{x}(s, t) - \mathbf{p}\|^2$ by choosing $s = -(\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_1$ and $t = -(\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_2$. Therefore, $\mathbf{q} - \mathbf{p} = \mathbf{x}(s_0, t_0) - \mathbf{p} = ((\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_3)\mathbf{u}_3$. In other words, since $\mathbf{u}_3 = \mathbf{u}$, $\mathbf{q} - \mathbf{p}$ is the component of $\mathbf{x}_0 - \mathbf{p}$ parallel to \mathbf{u} , $(\mathbf{x}_0 - \mathbf{p})_{\parallel}$. A picture will probably convince you that this is the correct formula.

Theorem 14. *Let \mathbf{u} be a unit vector in \mathbb{R}^3 and let \mathbf{x}_0 and \mathbf{p} be any two points in \mathbb{R}^3 . There is a unique point \mathbf{q} in the plane given by $\mathbf{u} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ such that $\|\mathbf{p} - \mathbf{q}\| < \|\mathbf{p} - \mathbf{x}\|$ for all other \mathbf{x} in the plane and $\mathbf{q} = \mathbf{p} + (\mathbf{x}_0 - \mathbf{p})_{\parallel}$ where $(\mathbf{x}_0 - \mathbf{p})_{\parallel}$ is the component of $\mathbf{x}_0 - \mathbf{p}$ parallel to \mathbf{u} , and therefore*

$$\|\mathbf{q} - \mathbf{p}\| = \|(\mathbf{x}_0 - \mathbf{p})_{\parallel}\| = |(\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}|$$

is the distance from \mathbf{p} to the plane.

Example 19 (A point-plane distance problem). Consider the plane given by $2x + y + 2z = 1$, and let $\mathbf{p} = (1, 1, 1)$. what point \mathbf{q} in the plane comes closest to \mathbf{p} , and what is the distance from \mathbf{p} to the plane?

We first find a base point \mathbf{x}_0 and a unit vector \mathbf{u} such that $2x + y + 2z = 1$ has the same solution set as $\mathbf{u} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$, and then a unit vector \mathbf{u} such that $\mathbf{u} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ is the equation of this plane in geometric form.

To find a base point, we simply need any one solution of the equation. Choosing $x = z = 0$, the equation reduces to $y = 1$, and so we choose $\mathbf{x}_0 = (0, 1, 0)$ as our base point.

The normal vector is $\mathbf{a} = (2, 1, 2)$, and normalize this to obtain $\mathbf{u} = \frac{1}{3}(2, 1, 2)$. Now compute $(\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u}_3 = \frac{1}{3}(1, 0, 1) \cdot (2, 1, 2) = \frac{4}{3}$ and so the distance to the plane is $\|\mathbf{q} - \mathbf{p}\| = \frac{4}{3}$, and $\mathbf{q} = \mathbf{p} + \frac{4}{3}\mathbf{u}$. That is,

$$\mathbf{q} = \mathbf{p} + \frac{4}{3}\mathbf{u} = (1, 1, 1) + \frac{4}{9}(2, 1, 2) = \frac{1}{9}(1, 5, 1).$$

The third question of this type that arises in \mathbb{R}^3 concerns the distance between two lines. Consider two lines parameterized by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$ and $\mathbf{x}_2(t) = \mathbf{x}_2 + t\mathbf{v}_2$, respectively. Let us assume that \mathbf{v}_1 and \mathbf{v}_2 are not multiples of one another, so that the lines are not parallel. (The parallel case is easier; we shall come back to it.) What values of s and t minimize $\|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|$?

To answer this, let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be an orthonormal basis of \mathbb{R}^3 in which \mathbf{u}_1 is orthogonal to both \mathbf{v}_1 and \mathbf{v}_2 , and in which \mathbf{u}_2 is orthogonal to \mathbf{v}_1 (and of course to \mathbf{u}_1). To produce this basis, define

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{v}_1 \times \mathbf{v}_2\|} \mathbf{v}_1 \times \mathbf{v}_2 \quad \text{and then} \quad \mathbf{u}_2 = \frac{1}{\|\mathbf{v}_1\|} \mathbf{v}_1 \times \mathbf{u}_1,$$

By the properties of the cross product, \mathbf{u}_1 is a unit vector orthogonal to both \mathbf{v}_1 and \mathbf{v}_2 , and \mathbf{u}_2 is a unit vector orthogonal to both \mathbf{u}_1 and \mathbf{v}_1 . Finally, we define $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$, and this gives us the orthonormal basis we seek. Since \mathbf{v}_1 is orthogonal to both \mathbf{u}_1 and \mathbf{u}_2 , $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$ must be a non-zero multiple of \mathbf{v}_1 .

We compute $\|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|^2$ in terms of coordinates for this basis. To simplify the notation, define $\mathbf{b} := \mathbf{x}_1 - \mathbf{x}_2$. Then

$$\begin{aligned}\|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|^2 &= \|\mathbf{b} + s\mathbf{v}_1 - t\mathbf{v}_2\|^2 \\ &= [(\mathbf{b} + s\mathbf{v}_1 - t\mathbf{v}_2) \cdot \mathbf{u}_1]^2 + [(\mathbf{b} + s\mathbf{v}_1 - t\mathbf{v}_2) \cdot \mathbf{u}_2]^2 + [(\mathbf{b} + s\mathbf{v}_1 - t\mathbf{v}_2) \cdot \mathbf{u}_3]^2 \\ &= [\mathbf{b} \cdot \mathbf{u}_1]^2 + [\mathbf{b} \cdot \mathbf{u}_2 - t(\mathbf{v}_2 \cdot \mathbf{u}_2)]^2 + [\mathbf{b} \cdot \mathbf{u}_3 + s(\mathbf{v}_1 \cdot \mathbf{u}_3) - t(\mathbf{v}_2 \cdot \mathbf{u}_3)]^2\end{aligned}\tag{1.57}$$

This is a sum of three squares. The first does not depend on s or t . The second depends only on t , and provided that $\mathbf{v}_2 \cdot \mathbf{u}_2 \neq 0$, we can make it zero by choosing

$$t = t_0 := \frac{\mathbf{b} \cdot \mathbf{u}_2}{\mathbf{v}_2 \cdot \mathbf{u}_2}\tag{1.58}$$

Then with this choice of t , provided that $\mathbf{v}_1 \cdot \mathbf{u}_3 \neq 0$, we can then make the third term zero by choosing

$$s = s_0 := \frac{t_0(\mathbf{v}_2 \cdot \mathbf{u}_3) - \mathbf{b} \cdot \mathbf{u}_3}{\mathbf{v}_1 \cdot \mathbf{u}_3}.\tag{1.59}$$

This then leaves only the first term, which is then the square of the minimal distance. However, we have to first verify that we are not dividing by zero in (1.58) and (1.59).

First, since \mathbf{u}_3 is a non-zero multiple of \mathbf{v}_1 , $\mathbf{v}_1 \cdot \mathbf{u}_3 \neq 0$. To show that $\mathbf{v}_2 \cdot \mathbf{u}_2 \neq 0$, we use (for the first time) our assumption that the lines are not parallel. Since \mathbf{v}_2 is orthogonal to \mathbf{u}_1 , we have the expansion $\mathbf{v}_2 = (\mathbf{v}_2 \cdot \mathbf{u}_2)\mathbf{u}_2 + (\mathbf{v}_2 \cdot \mathbf{u}_3)\mathbf{u}_3$. If $\mathbf{v}_2 \cdot \mathbf{u}_2$ were zero, this would reduce to $\mathbf{v}_2 = (\mathbf{v}_2 \cdot \mathbf{u}_3)\mathbf{u}_3$, but then since \mathbf{u}_3 is a multiple of \mathbf{v}_1 , the two lines would be parallel.

Thus, for any choices of s and t ,

$$\|\mathbf{x}_1(s) - \mathbf{x}_2(t)\| \geq \|\mathbf{x}_1(s_0) - \mathbf{x}_2(t_0)\| = |\mathbf{b} \cdot \mathbf{u}_1|,\tag{1.60}$$

and there is equality on the left if and only if $s = s_0$ and $t = t_0$. Thus, (1.60) gives the distance between the two lines. We then define the distance between the two lines to be the distance between these two closest points, and then (1.58) and (1.59) determine $\mathbf{x}_1(s_0)$ and $\mathbf{x}_2(t_0)$ the two closest points.

Now, what about the case of parallel lines? It is left to the reader to show that if the lines are parallel, the distance from any point on the first line to the second line is independent of the choice of the point on the first line. Thus, this problem reduces to the problem of computing the distance between a point and a line. Altogether, we have proved:

Theorem 15 (The distance between two lines). *The distance between two non-parallel lines in \mathbb{R}^3 parameterized by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$ and $\mathbf{x}_2(s) = \mathbf{x}_2 + t\mathbf{v}_2$ is*

$$\frac{|(\mathbf{x}_1 - \mathbf{x}_2) \cdot (\mathbf{v}_1 \times \mathbf{v}_2)|}{\|\mathbf{v}_1 \times \mathbf{v}_2\|}.$$

If the two lines are parallel, so that $\mathbf{v}_1 \times \mathbf{v}_2 = 0$, the distance is the distance from \mathbf{x}_1 to the second line; i.e., $\|(\mathbf{x}_1 - \mathbf{x}_2)_{\perp}\|$ where $(\mathbf{x}_1 - \mathbf{x}_2)_{\perp}$ is the component of $\mathbf{x}_1 - \mathbf{x}_2$ orthogonal to \mathbf{v}_1 , or, what is the same thing, orthogonal to \mathbf{v}_2 .

Notice that the last two distance problems solved in this subsection involved minimizing a squared distance over two parameters s and t , and that in each case, this minimization problem was rendered transparent by an appropriate choice of an orthonormal basis.

Example 20 (The distance between two lines in \mathbb{R}^3). Consider the two lines parameterized by

$$\mathbf{x}_1 + s\mathbf{v}_1 = (1, 2, 3) + s(1, 4, 5) \quad \text{and} \quad \mathbf{x}_2 + t\mathbf{v}_2 = (2, -1, 1) + t(-2, -1, 2).$$

We compute $\mathbf{v}_1 \times \mathbf{v}_2 = (1, 4, 5) \times (-2, -1, 2) = (13, -12, 7)$, and then $\|\mathbf{v}_1 \times \mathbf{v}_2\| = \|(13, -12, 7)\| = \sqrt{362}$, $\mathbf{u}_1 = \frac{1}{\sqrt{362}}(13, -12, 7)$. Next, $\mathbf{x}_2 - \mathbf{x}_1 = (2, -1, 1) - (1, 2, 3) = (-1, 3, 2)$. Finally we compute

$$\frac{(\mathbf{x}_1 - \mathbf{x}_2) \cdot (\mathbf{v}_1 \times \mathbf{v}_2)}{\|\mathbf{v}_1 \times \mathbf{v}_2\|} = \frac{(-1, 3, 2) \cdot (13, -12, 7)}{\sqrt{362}} = \frac{-13 - 36 + 14}{\sqrt{362}} = -\frac{35}{\sqrt{362}}.$$

Thus, the distance between the two lines is $35/\sqrt{362}$. If we had wanted to find the point on the first line that comes closest to the second, and the point of the second line that comes closest to the first, we would compute the right-handed orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ in which $\mathbf{u}_1 = (362)^{-1/2}(13, -12, 7)$, and then used (1.58) and (1.59).

We actually proved more than we have recorded in Theorem 15: We found formulas, (1.58) and (1.59), for the parameter values s_0 and t_0 that give the two closest points. There is another way to look at the problem that provides an easy proof of Theorem 15: Determining the distance between two lines is actually the problem of determining the distance between a point and a plane in disguise.

Going back to (1.57), notice that it gives us

$$\|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|^2 = \|(\mathbf{x}_2 - \mathbf{x}_1) + s\mathbf{v}_1 - t\mathbf{v}_2\|^2.$$

Define $\mathbf{x}_0 = \mathbf{x}_2 - \mathbf{x}_1$, $\mathbf{w}_1 = \mathbf{v}_1$ and $\mathbf{w}_2 = -\mathbf{v}_2$. Then

$$\|\mathbf{x}_1(s) - \mathbf{x}_2(t)\|^2 = \|(\mathbf{x}_0 + s\mathbf{w}_1 + t\mathbf{w}_2) - \mathbf{0}\|^2.$$

Thus, the distance we are seeking is the same as the distance from $\mathbf{0}$ to the plane parameterized by $\mathbf{x}_0 + s\mathbf{w}_1 + t\mathbf{w}_2$. But the equation of this plane is $\mathbf{u} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ where $\mathbf{u} = \|\mathbf{w}_1 \times \mathbf{w}_2\|^{-1}\mathbf{w}_1 \times \mathbf{w}_2$. Hence our formula for the distance from a point to a plane gives the distance as $|\mathbf{u} \cdot \mathbf{x}_0|$, which you can see is the same as the value that is given by the formula in Theorem 15. Moreover, Theorem 14 says that the point on this plane that is closest to $\mathbf{0}$ is $(\mathbf{x}_0)_{\parallel} = (\mathbf{x}_2 - \mathbf{x}_1)_{\parallel}$. Therefore, s_0 and t_0 , the minimizing values of s and t , are given by $(\mathbf{x}_2 - \mathbf{x}_1)_{\parallel} = (\mathbf{x}_2 - \mathbf{x}_1) + s_0\mathbf{v}_1 - t_0\mathbf{v}_2$, which reduces to

$$s_0\mathbf{v}_1 - t_0\mathbf{v}_2 = -(\mathbf{x}_2 - \mathbf{x}_1)_{\perp}.$$

Solving this equation yields the closest points. The previous approach did not require us to solve any equations; it avoided this by introducing a judiciously chosen orthonormal basis.

1.3 The Gram-Schmidt Orthonormalization Algorithm

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

We have seen that the cross product is an efficient device for constructing orthonormal bases of \mathbb{R}^3 that are suited to solving natural geometric problems, such as computing the distance between two lines in \mathbb{R}^3 . However, the cross product is special to \mathbb{R}^3 . We shall encounter problems in which we need to construct “custom made” orthonormal bases in \mathbb{R}^n for $n > 3$. How shall we construct them?

There is a very useful procedure for “extracting” a maximal orthonormal set from any collection of m vectors in \mathbb{R}^n for arbitrary m and n . This is the Gram-Schmidt Orthonormalization Algorithm, which we now describe.

When $n = 3$ and $m = 2$, this is already familiar: Consider $\{\mathbf{v}_1, \mathbf{v}_2\} \subset \mathbb{R}^3$, and suppose that neither of these vectors is the zero vector. Define \mathbf{u}_1 to be the direction vector (unit vector) corresponding to \mathbf{v}_1 :

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{v}_1\|} \mathbf{v}_1 . \quad (1.61)$$

Next, define $\mathbf{w}_2 = \mathbf{v}_2^\perp$, the component of \mathbf{v}_2 that is perpendicular to \mathbf{u}_1 , and therefore also to \mathbf{v}_1 . The familiar formula for \mathbf{w}_2 is

$$\mathbf{w}_2 = \mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{u}_1) \mathbf{u}_1 . \quad (1.62)$$

Provided $\mathbf{w}_2 \neq \mathbf{0}$, we can normalize it to obtain a unit vector \mathbf{u}_2 that is orthogonal to \mathbf{u}_1 :

$$\mathbf{u}_2 = \frac{1}{\|\mathbf{w}_2\|} \mathbf{w}_2 . \quad (1.63)$$

Example 21. Consider $\{\mathbf{v}_1, \mathbf{v}_2\} := \{(1, 1, 0), (1, 0, 1)\}$. Define \mathbf{u}_1 by normalizing \mathbf{v}_1 , as in (1.61):

$$\mathbf{u}_1 := \frac{1}{\|\mathbf{v}_1\|} \mathbf{v}_1 = \frac{1}{\sqrt{2}} (1, 1, 0) .$$

Next define \mathbf{w}_2 by subtracting off from \mathbf{v}_2 its component that is parallel to \mathbf{u}_1 , as in (1.62): That is, $\mathbf{w}_2 := \mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{u}_1) \mathbf{u}_1$. Computing, we find $\mathbf{v}_2 \cdot \mathbf{u}_1 = (1, 0, 1) \cdot \frac{1}{\sqrt{2}} ((1, 1, 0)) = \frac{1}{\sqrt{2}}$, and hence

$$\mathbf{w}_2 = (1, 0, 1) - \frac{1}{2} (1, 1, 0) = \frac{1}{2} (1, -1, 2) .$$

We normalize \mathbf{w}_2 to define \mathbf{u}_2 : In fact, we may as well ignore the $1/2$, and normalize $(1, -1, 2)$ since it is the direction of \mathbf{w}_2 that concerns us. We compute $\|(1, -1, 2)\| = \sqrt{6}$, and so

$$\mathbf{u}_2 := \frac{1}{\sqrt{6}} (1, -1, 2) .$$

Since \mathbf{u}_2 is a multiple of \mathbf{w}_2 , and since \mathbf{w}_2 is a linear combination of \mathbf{v}_1 and \mathbf{v}_2 , so is \mathbf{u}_2 . This gives us our orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$.

This procedure for constructing the orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$ out of the original set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a special case of the Gram-Schmidt Algorithm. We now explain the geometric content of the construction in the case that $\mathbf{w}_2 \neq \mathbf{0}$: *The two sets of vectors, $\{\mathbf{u}_1, \mathbf{u}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2\}$, span the same plane in \mathbb{R}^3 .*

To see this, combine (1.61), (1.62) and (1.63) to obtain

$$\begin{aligned}\mathbf{u}_2 &= \frac{1}{\|\mathbf{w}_2\|} \left(\mathbf{v}_2 - \frac{\mathbf{v}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 \right) \\ &= \frac{1}{\|\mathbf{w}_2\|} \mathbf{v}_2 - \frac{\mathbf{v}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2 \|\mathbf{w}_2\|} \mathbf{v}_1 = a\mathbf{v}_1 + b\mathbf{v}_2\end{aligned}$$

for some a and b that can be read off from the middle line of (1.64). Therefore, (1.64) shows that \mathbf{u}_2 is a linear combination of \mathbf{v}_1 and \mathbf{v}_2 . Even more simply, (1.61) displays \mathbf{u}_1 as a linear combination of \mathbf{v}_1 and \mathbf{v}_2 , but of a very simple kind: it is a multiple of \mathbf{v}_1 alone.

Thus, the procedure described above yields an orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$ such that \mathbf{u}_1 is a multiple of \mathbf{v}_1 , and \mathbf{u}_2 is a linear combination of \mathbf{v}_1 and \mathbf{v}_2 .

There is also a converse: obviously \mathbf{v}_1 is a multiple of \mathbf{u}_1 , and (1.62) can be rewritten as

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

so that \mathbf{v}_2 is a linear combination of \mathbf{u}_1 and \mathbf{u}_2 . We therefore have

$$\{\mathbf{u}_1, \mathbf{u}_2\} \subset \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}) \quad \text{and} \quad \{\mathbf{v}_1, \mathbf{v}_2\} \subset \text{Span}(\{\mathbf{u}_1, \mathbf{u}_2\}) .$$

By Theorem 1, this means that whenever \mathbf{u}_2 is defined,

$$\text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}) = \text{Span}(\{\mathbf{u}_1, \mathbf{u}_2\}) .$$

So far, we have assumed that $\mathbf{w}_2 \neq \mathbf{0}$. However, If $\mathbf{w}_2 = \mathbf{0}$, then (1.62) says that

$$\mathbf{v}_2 = (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1 = \frac{\mathbf{v}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 ,$$

so that \mathbf{v}_2 is a multiple of \mathbf{v}_1 . Conversely, suppose that \mathbf{v}_2 is a multiple of \mathbf{v}_1 , and hence of \mathbf{u}_1 . Then for some $a \in \mathbb{R}$, $\mathbf{v}_2 = a\mathbf{u}_1$, and $\mathbf{v}_2 \cdot \mathbf{u}_1 = a$. Then $(\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1 = a\mathbf{u}_1 = \mathbf{v}_2$, and by (1.62), $\mathbf{w}_2 = \mathbf{0}$. Thus: *The procedure fails to produce a second unit vector exactly when \mathbf{v}_1 and \mathbf{v}_2 are multiples of one another.* In this case, it is clear that the span of $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a line, and we cannot hope to extract an orthonormal set of two or more vectors form it.

Thus, given any set $\{\mathbf{v}_1, \mathbf{v}_2\}$ of two non-zero vectors in \mathbb{R}^3 , the procedure described above detects whether the set of all vector of the form $s_1\mathbf{v}_1 + s_2\mathbf{v}_2$, $s_1, s_2 \in \mathbb{R}$, is a plane or a line. Moreover, in case the set is a plane, the procedure yields an orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$ such that

$$(t_1, t_2) \mapsto t_1\mathbf{u}_1 + t_2\mathbf{u}_2$$

is a parameterization of this plane.

Supposing that \mathbf{u}_2 is defined, we now “flesh out” the orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$ into an orthonormal basis of \mathbb{R}^3 without using the cross product. Let \mathbf{x} be any vector in \mathbb{R}^3 that is *not* in the plane spanned by \mathbf{v}_1 and \mathbf{v}_2 , or equivalently \mathbf{u}_1 and \mathbf{u}_2 . Define

$$\mathbf{w} := \mathbf{x} - (\mathbf{x} \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{x} \cdot \mathbf{u}_2)\mathbf{u}_2 . \tag{1.64}$$

Since $\mathbf{x} \notin \text{Span}(\{\mathbf{u}_1, \mathbf{u}_2\})$, $\mathbf{w} \neq \mathbf{0}$, A familiar computation gives $\mathbf{w} \cdot \mathbf{u}_1 = \mathbf{w} \cdot \mathbf{u}_2 = 0$, and therefore \mathbf{w} is a non-zero vector orthogonal to both \mathbf{u}_1 and \mathbf{u}_2 . Define $\mathbf{u}_3 = \|\mathbf{w}\|^{-1}\mathbf{w}$. Then $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an orthonormal basis of \mathbb{R}^3 .

Example 22. Consider $\{\mathbf{v}_1, \mathbf{v}_2\} := \{(1, 1, 0), (1, 0, 1)\}$ as in Example 21. The procedure we have described above yields, as seen in Example 21, the orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$ where

$$\mathbf{u}_1 := \frac{1}{\|\mathbf{v}_1\|} \mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, 1, 0) \quad \text{and} \quad \mathbf{u}_2 := \frac{1}{\sqrt{6}}(1, -1, 2) .$$

It is not hard to see that $\mathbf{e}_1 \notin \text{span}(\{\mathbf{v}_1, \mathbf{v}_2\})$, but one way to check this is to make this choice for \mathbf{x} , and then see if the procedure yields \mathbf{u}_3 or not. If not, make another choice. Applying (1.64) with $\mathbf{x} = \mathbf{e}_1$, we compute $\mathbf{x} \cdot \mathbf{u}_1 = 1/\sqrt{2}$ and $\mathbf{x} \cdot \mathbf{u}_2 = 1/\sqrt{6}$. Hence,

$$\mathbf{w} = \mathbf{e}_1 - (\mathbf{e}_1 \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{e}_1 \cdot \mathbf{u}_2)\mathbf{u}_2 = (1, 0, 0) - \frac{1}{2}(1, 1, 0) - \frac{1}{6}(1, -1, 2) = \frac{1}{6}(2, -2, -2) .$$

This is not the zero vector, confirming that $\mathbf{e}_1 \notin \text{span}(\{\mathbf{v}_1, \mathbf{v}_2\})$. Normalizing, we find

$$\mathbf{u}_3 = \frac{1}{\sqrt{3}}(1, -1, -1) .$$

It is now easy to compute that $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3$ so that our dot product construction has produced the same result that we would have obtained using the cross product. However in general, this may not happen: The procedure we have described will always produce an orthonormal basis in \mathbb{R}^3 , but it might be a left handed basis or it might be a right handed basis. For now, let us focus on constructing orthonormal bases and put aside the issue of whether they are right handed or left handed – a concept that we have only defined so far in \mathbb{R}^3 .

1.3.2 The Gram-Schmidt Algorithm in general

Definition 17 (The Gram-Schmidt Algorithm). Let $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ be any ordered list of m vectors in \mathbb{R}^n such that at least one of these vectors is not the zero vector.

(1) Let p_1 be the least value of j such that $\mathbf{v}_j \neq \mathbf{0}$. Define

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{v}_{p_1}\|} \mathbf{v}_{p_1} .$$

The vector \mathbf{v}_{p_1} is called the first pivotal vector. If $p_1 = m$, the procedure is terminates, and produces the set $\{\mathbf{u}_1\}$.

(2) Otherwise, starting with $j = p_1 + 1$, compute

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

If this is zero for all $j > p_1 + 1$. the procedure is terminates, and produces the set $\{\mathbf{u}_1\}$. Otherwise, let p_2 be the least value of $j > p_1$ such that $\mathbf{w}_2 \neq \mathbf{0}$, and define

$$\mathbf{u}_2 = \frac{1}{\|\mathbf{w}_2\|} \mathbf{w}_2 .$$

The vector \mathbf{v}_{p_2} is called the second pivotal vector. If $p_2 = m$, the procedure is terminates, and produces the set $\{\mathbf{u}_1, \mathbf{u}_2\}$.

(3) Now suppose that $\{p_1, \dots, p_k\}$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ have been defined, and $p_k < m$. Starting with $j = p_k + 1$,

$$\mathbf{w}_j = \mathbf{v}_j - \sum_{\ell=1}^k (\mathbf{v}_j \cdot \mathbf{u}_\ell) \mathbf{u}_\ell .$$

If this is zero for all $j > p_k + 1$, the procedure is terminates, and produces the set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$. otherwise, let p_{k+1} be the least value of $j > p_k$ such that $\mathbf{w}_{p_{k+1}} \neq \mathbf{0}$, and define

$$\mathbf{u}_{k+1} = \frac{1}{\|\mathbf{w}_{p_{k+1}}\|} \mathbf{w}_{p_{k+1}}.$$

The vector $\mathbf{v}_{p_{k+1}}$ is called the $(k+1)$ st pivotal vector. If $p_{k+1} = m$, the procedure is terminates, and produces the set $\{\mathbf{u}_1, \dots, \mathbf{u}_{k+1}\}$. Otherwise, repeat the procedure until it terminates, which it must do after at most m steps.

Example 23 (Using the Gram-Schmidt Algorithm). Let $\mathbf{v}_1 = (1, 2, -3)$, $\mathbf{v}_2 = (1, -2, 1)$, $\mathbf{v}_3 = (-2, 1, 1)$ and $\mathbf{v}_4 = (0, 1, 1)$. Applying the Gram-Schmidt Algorithm to $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$, we see that since $\mathbf{v}_1 \neq \mathbf{0}$, $p_1 = 1$ and

$$\mathbf{u}_1 = \frac{1}{\sqrt{14}}(1, 2, -3).$$

We then compute

$$\mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{u}_1)\mathbf{u}_1 = (1, -2, 1) + \frac{6}{14}(1, 2, -3) = \frac{1}{7}(10, -8, -2).$$

Since this is not the zero vector, we renormalize it to obtain $p_2 = 2$, and

$$\mathbf{u}_2 = \frac{1}{\sqrt{42}}(5, -4, -1).$$

There are more vectors left in our list, so we go on to compute

$$\begin{aligned} \mathbf{v}_3 - (\mathbf{v}_3 \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{v}_3 \cdot \mathbf{u}_2)\mathbf{u}_2 &= (-2, 1, 1) + \frac{3}{14}(1, 2, -3) + \frac{15}{42}(5, -4, -1) \\ &= (0, 0, 0) \end{aligned}$$

Since we got the zero vector, we cannot normalize, so we go on to try the same thing for \mathbf{v}_4 :

$$\begin{aligned} \mathbf{v}_4 - (\mathbf{v}_4 \cdot \mathbf{u}_1)\mathbf{u}_1 - (\mathbf{v}_4 \cdot \mathbf{u}_2)\mathbf{u}_2 &= (0, 1, 1) + \frac{1}{14}(1, 2, -3) + \frac{5}{42}(5, -4, -1) \\ &= \frac{2}{3}(1, 1, 1) \end{aligned}$$

Since this is not the zero vector, $p_3 = 4$. Normalizing, we find $\mathbf{u}_3 = \frac{1}{\sqrt{3}}(1, 1, 1)$.

We have now reached the end of our list of vectors, and the procedure terminates, providing the set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$.

The set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ that the Gram-Schmidt algorithm produces from $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ is orthonormal, and it has the same span as the original set: We now prove this and more:

Theorem 16 (Properties of the output of the Gram-Schmidt algorithm). Let $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ be a set of m vectors in \mathbb{R}^n , not all of which are the zero vector. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ be the set produced from it by the Gram-Schmidt Algorithm. Then:

- (1) The set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is orthonormal, and $r \leq \min\{m, n\}$,
- (2) There are r pivotal vectors $\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\}$ and

$$\text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\}) = \text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_m\}) = \text{Span}(\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\}). \quad (1.65)$$

- (3) For each $j = 1 \dots, r$, $\mathbf{u}_j \cdot \mathbf{v}_{p_j} > 0$.

Proof. By definition, each pivotal vector yields one vector in the set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$, and so there are exactly r pivotal vectors $\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\}$. Also by definition, for each $1 < j \leq r$, the j th pivotal vector \mathbf{v}_{p_j} and $\{\mathbf{u}_1, \dots, \mathbf{u}_j\}$ are related by

$$\mathbf{u}_j := \|\mathbf{w}_{p_j}\|^{-1} \mathbf{w}_{p_j} \quad \text{where} \quad \mathbf{w}_{p_j} := \mathbf{v}_{p_j} - \sum_{i=1}^{j-1} (\mathbf{v}_{p_j} \cdot \mathbf{u}_i) \mathbf{u}_i . \quad (1.66)$$

For all $\ell < j$, one readily checks that $\mathbf{w}_{p_j} \cdot \mathbf{u}_\ell = 0$, and hence \mathbf{u}_j is a unit vector that is orthogonal to \mathbf{u}_ℓ for all $\ell < j$. Thus, $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is orthonormal. By Lemma 1, $r \leq n$, and since at most m of the vectors in $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ can be pivotal, $r \leq m$. This proves (1).

We may rearrange (1.66) to obtain $\mathbf{v}_{p_j} = \|\mathbf{w}_{p_j}\| \mathbf{u}_j + \sum_{i=1}^{j-1} (\mathbf{v}_{p_j} \cdot \mathbf{u}_i) \mathbf{u}_i$, which shows that each \mathbf{v}_{p_j} belongs to $\text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$. That is,

$$\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\} \subset \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\}) \quad (1.67)$$

Consider the non-pivotal vectors, if any. Any such vector belongs to $\text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$: If $\mathbf{v}_j = \mathbf{0}$, this is trivially true – the zero vector is in the span of any non-empty set of vectors. If \mathbf{v}_j is non-zero and non-pivotal, this means, by definition, that with $k(j)$ being the number of pivotal vectors with indices less than j , $\mathbf{w}_j = \mathbf{0}$ where $\mathbf{w}_j := \mathbf{v}_j - \sum_{\ell=1}^k (\mathbf{v}_j \cdot \mathbf{u}_\ell) \mathbf{u}_\ell$. Hence $\mathbf{v}_j = \sum_{\ell=1}^k (\mathbf{v}_j \cdot \mathbf{u}_\ell) \mathbf{u}_\ell$, which means that \mathbf{v}_j belongs to $\text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$. Combing this with (1.67), each \mathbf{v}_j , *pivotal or not*, belongs to $\text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$, so that

$$\{\mathbf{v}_1, \dots, \mathbf{v}_m\} \subset \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\}) . \quad (1.68)$$

We now show that for each $k = 1, \dots, r$, $\mathbf{u}_k \in \text{Span}(\{\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_k}\})$. For $k = 1$ this is clear since by definition, $\mathbf{u}_1 = \|\mathbf{v}_{j_1}\|^{-1} \mathbf{v}_{j_1}$. Now suppose we know that for some $k = 2, \dots, r$, $\mathbf{u}_\ell \in \text{Span}(\{\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_\ell}\})$ for all $1 \leq \ell \leq k-1$. Then (1.66) says that \mathbf{w}_{p_k} is a linear combination of vectors in $\text{Span}(\{\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_k}\})$, and by Theorem 1, then \mathbf{w}_{p_k} itself belongs to $\text{Span}(\{\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_k}\})$. But since \mathbf{u}_k is a multiple of \mathbf{w}_{p_k} , this means that $\mathbf{u}_k \in \text{Span}(\{\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_k}\})$. This is the inductive step, and thus for each $k = 1, \dots, r$, $\mathbf{u}_k \in \text{Span}(\{\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_k}\})$. That is,

$$\{\mathbf{u}_1, \dots, \mathbf{u}_r\} \subset \text{Span}(\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\}) . \quad (1.69)$$

Theorem 1 says that for any set $W \subset \mathbb{R}^n$, $\text{Span}(\text{Span}(W)) = \text{Span}(W)$, and hence for any $V \subset \mathbb{R}^n$ with $V \subset \text{Span}(W)$, $\text{Span}(V) \subset \text{Span}(W)$. Applying this to (1.67), (1.68) and (1.69), and the obvious fact that $\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\} \subset \text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_m\})$

$$\text{Span}(\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\}) \subset \text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_m\}) \subset \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\}) \subset \text{Span}(\{\mathbf{v}_{p_1}, \dots, \mathbf{v}_{p_r}\}) .$$

Hence all of these sets are equal to one another. This proves (2).

Finally, by (1) and (1.66), $\mathbf{u}_j \cdot \mathbf{v}_{p_j} = \mathbf{u}_j \cdot \mathbf{w}_{p_j} = \|\mathbf{w}_{p_j}\| > 0$. □

Corollary 2 (Corollary of Theorem 16). *Let $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ be a set of m vectors in \mathbb{R}^n , not all of which are the zero vector. A non-zero vector \mathbf{v}_ℓ is non-pivotal for the Gram-Schmidt Algorithm if and only if it is a linear combination of the vectors in $\{\mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}\}$.*

Proof. Apply the Gram-Schmidt Algorithm to the set $\{\mathbf{v}_1, \dots, \mathbf{v}_\ell\}$, which contains at least one non-zero vector, namely \mathbf{v}_ℓ . Let $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ be the resulting orthonormal set.

Suppose that \mathbf{v}_ℓ is not pivotal. Then one gets the same set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ applying the Gram-Schmidt Algorithm to the set $\{\mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}\}$ – the additional vector \mathbf{v}_ℓ adds nothing new. By (2) of Theorem 16, applied twice

$$\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}\}) = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\}) = \text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_\ell\}).$$

Therefore, $\mathbf{v}_\ell \in \text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}\})$.

Conversely, if $\mathbf{v}_\ell \in \text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}\})$, and $\{\mathbf{u}_1, \dots, \mathbf{u}_s\}$ is the orthonormal set one gets applying the Gram-Schmidt Algorithm to the set $\{\mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}\}$, then by (2) of Theorem 16, $\mathbf{v}_\ell \in \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_s\})$, and $\mathbf{v}_\ell - \sum_{j=1}^s (\mathbf{v}_\ell \cdot \mathbf{u}_j) \mathbf{u}_j = \mathbf{0}$. Therefore, \mathbf{v}_ℓ is non-pivotal. \square

1.3.3 Subspaces of \mathbb{R}^n

For $n > 3$, subsets of \mathbb{R}^n of the form $\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_m\})$ are the natural higher dimensional analogs of lines and planes through the origin in \mathbb{R}^3 . These special subsets of \mathbb{R}^n are called *subspaces* of \mathbb{R}^n . The Gram-Schmidt Algorithm provides the means to answer all sorts of geometric questions concerning subspaces.

We begin with some preliminary observations: Let $V \subset \mathbb{R}^n$ be non-empty. Suppose V has the properties that:

- (1) V is *closed under scalar multiplication*, meaning that if $t \in \mathbb{R}$ and $\mathbf{v} \in V$, then $t\mathbf{v} \in V$.
- (2) V is *closed under vector addition*, meaning that if $\mathbf{v}_1, \mathbf{v}_2 \in V$, then $\mathbf{v}_1 + \mathbf{v}_2 \in V$.

Then for every $\mathbf{v}_1, \mathbf{v}_2 \in V$ and $t_1, t_2 \in \mathbb{R}$, $t_1\mathbf{v}_1 + t_2\mathbf{v}_2 \in V$. Moreover, for any additional $t_3 \in \mathbb{R}$, and $\mathbf{v}_3 \in V$,

$$t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + t_3\mathbf{v}_3 = (t_1\mathbf{v}_1 + t_2\mathbf{v}_2) + t_3\mathbf{v}_3 \in V,$$

and by induction we see that for all m and all $t_1, \dots, t_m \in \mathbb{R}$ and all $\{\mathbf{v}_1, \dots, \mathbf{v}_m\} \subset V$, $\sum_{k=1}^m t_k \mathbf{v}_k \in V$.

This shows that $\text{Span}(V) \subset V$, and since we always have $V \subset \text{Span}(V)$, our hypotheses on V imply that $V = \text{Span}(V)$. Conversely, if $V = \text{Span}(V)$, V is closed under scalar multiplication and vector addition since every linear combination of vectors in V belongs to V .

Definition 18 (Subspaces of \mathbb{R}^n). A non-empty subset $V \subset \mathbb{R}^n$ is a *subspace* of \mathbb{R}^n in case V is closed under scalar multiplication and vector addition, or, equivalently, in case $V = \text{Span}(V)$.

We have seen that every subset of \mathbb{R}^n having the form $\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_m\})$ is a subspace of \mathbb{R}^n . By Theorem 16, provided at least one of the vectors in non-zero, $\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_m\})$ is equal to $\text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$ for some orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ with $r \leq m$.

In fact, *every* subspace V of \mathbb{R}^n that contains something other than the zero vector has the form

$$V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$$

for some orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\} \subset V$. (The set $V = \{\mathbf{0}\}$, is a subspace, called the *zero subspace*, but there is not much to say about it.) To see this, use the Gram-Schmidt Algorithm as follows:

First, select a nonzero vector $\mathbf{v}_1 \in V$, and normalize it to obtain \mathbf{u}_1 . If $\text{Span}(\{\mathbf{u}_1\}) = V$, we are done. Otherwise, there exists some $\mathbf{v}_2 \in V$ with $\mathbf{v}_2 \notin \text{Span}(\{\mathbf{u}_1\})$. Apply the Gram-Schmidt Algorithm to $\{\mathbf{u}_1, \mathbf{v}_2\}$ to obtain $\{\mathbf{u}_1, \mathbf{u}_2\}$, noting that \mathbf{v}_2 is necessarily pivotal since it is not in $\text{Span}(\{\mathbf{u}_1\})$. (See the Corollary to Theorem 16). Finally, note that by (1.65), $\mathbf{u}_2 \in \text{Span}(\{\mathbf{u}_1, \mathbf{v}_2\}) \subset \text{Span}(V) = V$, the last equality being true precisely because V is a subspace.

If $\text{Span}(\{\mathbf{u}_1, \mathbf{u}_2\}) = V$, we are done. Otherwise, there exists some $\mathbf{v}_3 \in V$ with $\mathbf{v}_3 \notin \text{Span}(\{\mathbf{u}_1, \mathbf{u}_2\})$. Apply the Gram-Schmidt Algorithm to $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_3\}$ to obtain $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, noting that \mathbf{v}_3 is necessarily pivotal since it is not in $\text{Span}(\{\mathbf{u}_1, \mathbf{u}_2\})$. By the same reasoning as above, again using the fact that V is a subspace, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \subset V$.

Continue in this way until the procedure terminates. This must happen by the n th step at most since \mathbb{R}^n does not contain any orthonormal subset consisting of more than n vectors. When the procedure terminates, we have an *orthonormal subset* $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ of V such that $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$.

Although the procedure we described involved a largely arbitrary choice of the set of vectors to which we applied the Gram-Schmidt Algorithm, the number r does not depend on any of the choices we made. The number r is characteristic of the non-zero subspace V , and is called its *dimension*.

To prove this important fact, we use the natural parameterization and coordinate functions associated to $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$:

Let V be a non-zero subspace of \mathbb{R}^n , and let $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ be any orthonormal subset of V such that $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$. Define the *parameterization function* P mapping \mathbb{R}^r into V by

$$P((t_1, \dots, t_r)) = \sum_{j=1}^r t_j \mathbf{u}_j .$$

and define the *coordinate function* C mapping V into \mathbb{R}^r by

$$C(\mathbf{v}) = (\mathbf{v} \cdot \mathbf{u}_1, \dots, \mathbf{v} \cdot \mathbf{u}_r) .$$

both maps are one-to one and onto, so that both are invertible, and in fact, they are inverse to each other.

The function P maps \mathbb{R}^r onto V since $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$. The function P is one-to-one since if $\mathbf{t}, \mathbf{s} \in \mathbb{R}^r$, and $P(\mathbf{t}) = P(\mathbf{s})$, then for each $j = 1, \dots, r$, $P(\mathbf{t}) \cdot \mathbf{u}_j = P(\mathbf{s}) \cdot \mathbf{u}_j$, which is the same as $t_j = s_j$.

This shows that P is invertible, and then $t_k = (\sum_{j=1}^r t_j \mathbf{u}_j) \cdot \mathbf{u}_k = (P((t_1, \dots, t_r))) \cdot \mathbf{u}_k$ shows that C is the inverse of P .

The functions P and C are *isometries*. That is, the both preserves distances and angles between vectors. To see this, let \mathbf{s} and \mathbf{t} be two vectors in \mathbb{R}^r . We compute

$$P(\mathbf{s}) \cdot P(\mathbf{t}) = \left(\sum_{j=1}^r s_j \mathbf{u}_j \right) \cdot \left(\sum_{k=1}^r t_k \mathbf{u}_k \right) = \sum_{j,k=1}^r s_j t_k \mathbf{u}_j \cdot \mathbf{u}_k = \sum_{j=1}^r s_j t_j = \mathbf{s} \cdot \mathbf{t} .$$

In short, for all $\mathbf{s}, \mathbf{t} \in \mathbb{R}^r$,

$$P(\mathbf{s}) \cdot P(\mathbf{t}) = \mathbf{s} \cdot \mathbf{t} . \quad (1.70)$$

Now let $\mathbf{v}, \mathbf{w} \in V$. Define $\mathbf{s} := C(\mathbf{v})$ and $\mathbf{t} = C(\mathbf{w})$. $\mathbf{v} = P(\mathbf{s})$ and $\mathbf{w} = P(\mathbf{t})$. Then we can rewrite (1.70) as

$$\mathbf{v} \cdot \mathbf{w} = C(\mathbf{v}) \cdot C(\mathbf{w}), \quad (1.71)$$

for all $\mathbf{v}, \mathbf{w} \in V$.

The formulas (1.70) and (1.71) have the following consequence: *A set $\{\mathbf{v}_1, \dots, \mathbf{v}_m\} \subset V$ is orthonormal if and only if $\{C(\mathbf{v}_1), \dots, C(\mathbf{v}_m)\} \subset \mathbb{R}^r$ is orthonormal. Likewise, as set $\{\mathbf{s}_1, \dots, \mathbf{s}_m\} \subset \mathbb{R}^r$ is orthonormal if and only if $\{P(\mathbf{s}_1), \dots, P(\mathbf{s}_m)\} \subset V$ is orthonormal.*

Now let V be a non-zero subspace of \mathbb{R}^n . Suppose that $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ and $\{\mathbf{z}_1, \dots, \mathbf{z}_s\}$ are two orthonormal subsets of V and that both span V . We want to show that $s = r$. By symmetry, it suffices to show that $s \leq r$.

To do this, use $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ to define a coordinate function C from V to \mathbb{R}^r , but use $\{\mathbf{z}_1, \dots, \mathbf{z}_s\}$ to define a parameterization function P from \mathbb{R}^s to V . The composition, $C \circ P$, is a function from \mathbb{R}^s to \mathbb{R}^r . Since it is the compositions of functions that take orthonormal sets into orthonormal sets, $C \circ P$ also has this property. Hence if $\{\mathbf{e}_1, \dots, \mathbf{e}_s\}$ is the standard basis for \mathbb{R}^s , $\{C(P(\mathbf{e}_1)), \dots, C(P(\mathbf{e}_s))\}$ is orthonormal in \mathbb{R}^r . By Lemma 1, $s \leq r$. We have proved:

Theorem 17. *Let V be a subspace of \mathbb{R}^n . Then there exists an orthonormal subset $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ of V whose span is V , and all such spanning orthonormal subsets of V have the same cardinality r . In particular, if $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is an orthonormal set that spans \mathbb{R}^n , then $r = n$.*

Definition 19 (Dimension). *The dimension of a subspace V of \mathbb{R}^n , $\dim(V)$, is the cardinality r of any orthonormal subset $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ of V that spans V . An orthonormal set of $\dim(V)$ vectors in V is called an orthonormal basis for V .*

If V is a subspace of \mathbb{R}^n , $n > 3$, we say that V is a *line through the origin* in case $\dim(V) = 1$, and a *plane through the origin* in case $\dim(V) = 2$. If $\dim(V) = n - 1$, we say that V is a *hyperplane through the origin of \mathbb{R}^n* . Subspaces of other dimensions do not have nick-names.

Dimension is often used in proofs that two subspaces are the same:

Theorem 18. *Let V and W be subspaces of \mathbb{R}^n . If $V \subset W$ and $\dim(V) = \dim(W)$, then $V = W$.*

Proof. Let $d = \dim(V)$, and let $\{\mathbf{u}_1, \dots, \mathbf{u}_d\}$ be an orthonormal set of d vectors in V , which exists by Theorem 17. Suppose there exists some $\mathbf{w} \in W$ such that $\mathbf{w} \notin V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_d\})$. Then we can apply the Gram-Schmidt Algorithm to $\{\mathbf{u}_1, \dots, \mathbf{u}_d, \mathbf{z}\}$ to produce an orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_d, \mathbf{u}_{d+1}\}$ in W . But this is impossible if $\dim(W) = d$. \square

1.3.4 Orthogonal complements

We have seen that in \mathbb{R}^n , $n > 3$, subspaces of \mathbb{R}^n are the natural higher dimensional analogs of lines and planes through the origin in \mathbb{R}^3 . Lines and planes in \mathbb{R}^3 are the sets of solutions to certain equations in \mathbb{R}^3 that can be written in terms of dot products. Moreover corresponding to any plane through the origin in \mathbb{R}^3 is its *normal line*, the line through the origin consisting of the vectors that are orthogonal to the plane. There is an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of \mathbb{R}^3 such that $\mathbf{u}_3 \cdot \mathbf{x} = 0$ is the equation of the plane, and $(\mathbf{u}_1 \cdot \mathbf{x}, \mathbf{u}_2 \cdot \mathbf{x}) = (0, 0)$ is the equation of the line. We have seen that

this way of writing the equations in terms of an orthonormal basis, is very useful in solving distance problems , for example,

The extension of all of this to arbitrary dimension involves the useful notion of *orthogonal complements* which we now define:

Definition 20 (Orthogonal complement). *Let S be any subset of \mathbb{R}^n , finite or infinite. The orthogonal complement of S , S^\perp , is the subset of vectors $\mathbf{x} \in \mathbb{R}^n$ that are orthogonal to each $\mathbf{v} \in S$.*

Note that if $S_1 \subset S_2$, then a vector \mathbf{y} has to satisfy more conditions of the form $\mathbf{x} \cdot \mathbf{y} = 0$ to belong to S_2^\perp than to S_1^\perp . Hence:

$$S_1 \subset S_2 \Rightarrow S_2^\perp \subset S_1^\perp . \quad (1.72)$$

Example 24 (Orthogonal complements). *Let $\{\mathbf{a}\}$ be a non-zero vector in \mathbb{R}^3 , and let $S = \{\mathbf{a}\}$. Then S^\perp is the set of all vectors orthogonal to \mathbf{a} , i.e., the set of all $\mathbf{x} \in \mathbb{R}^3$ such that*

$$\mathbf{a} \cdot \mathbf{x} = 0 .$$

This is the equation of the plane through the origin that is normal to \mathbf{a} .

Likewise, let $S = \{\mathbf{a}_1, \mathbf{a}_2\}$ be a set of two non-zero vectors in \mathbb{R}^3 that are not multiples of one another. Then S^\perp is the line specified by

$$\begin{aligned} \mathbf{a}_1 \cdot \mathbf{x} &= 0 \\ \mathbf{a}_2 \cdot \mathbf{x} &= 0 . \end{aligned}$$

In both of these examples, S^\perp is a subspace. This is always the case:

Theorem 19. *Let S be any subset of \mathbb{R}^n . Then S^\perp is a subspace. For any subspace V of \mathbb{R}^n ,*

$$\dim(V) + \dim(V^\perp) = n \quad (1.73)$$

and

$$(V^\perp)^\perp = V . \quad (1.74)$$

Finally, with $k := \dim(V)$, there exists an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{u}_{k+1}, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n such that $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is an orthonormal basis of V , and $\{\mathbf{u}_{k+1}, \dots, \mathbf{u}_n\}$ is an orthonormal basis of V^\perp .

Proof. Let $\mathbf{x}, \mathbf{y} \in S^\perp$, and let a, b be any two numbers. For any $\mathbf{v} \in S$,

$$(a\mathbf{x} + b\mathbf{y}) \cdot \mathbf{v} = a(\mathbf{x} \cdot \mathbf{v}) + b(\mathbf{y} \cdot \mathbf{v}) = a0 + b0 = 0 .$$

Thus $a\mathbf{x} + b\mathbf{y} \in S^\perp$. This shows that S^\perp is a subspace.

Now let V be a non-trivial subspace of \mathbb{R}^n , and let $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ be an orthonormal basis of V . Since V^\perp is a subspace of \mathbb{R}^n , it too has an orthonormal basis $\{\mathbf{z}_1, \dots, \mathbf{z}_\ell\}$. Since every vector in V^\perp is orthogonal to every vector in V , if define $\mathbf{u}_{k+j} = \mathbf{z}_j$ for $j = 1, \dots, \ell$,

$$\{\mathbf{v}_1, \dots, \mathbf{v}_k, \mathbf{v}_{k+1}, \dots, \mathbf{v}_{k+\ell}\}$$

is an orthonormal subset of \mathbb{R}^n . Since there does not exist any orthonormal subset of \mathbb{R}^n consisting of more than n vectors, $k + \ell \leq n$.

Suppose that $k + \ell < n$. Then $\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_k, \mathbf{v}_{k+1}, \dots, \mathbf{v}_{k+\ell}\})$ is a subspace W of \mathbb{R}^n of dimension $k + \ell < n$, and it cannot be all of \mathbb{R}^n which has dimension n . Then there exists a vector $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{x} \notin W$. Define

$$\mathbf{z} = \mathbf{x} - \sum_{j=1}^{k+\ell} (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j .$$

Since $\mathbf{x} \notin W$, $\mathbf{z} \neq \mathbf{0}$ and it is easy to check that \mathbf{z} is orthogonal to each vector in $\{\mathbf{v}_1, \dots, \dots, \mathbf{v}_{k+\ell}\}$, and hence to each vector in W . But W contains both V and V^\perp , so that \mathbf{z} is orthogonal to every vector in V and in V^\perp . This means \mathbf{z} is in V^\perp , and is also orthogonal to every vector in V^\perp . In particular, it is orthogonal to itself. Hence $\|\mathbf{z}\|^2 = \mathbf{z} \cdot \mathbf{z} = 0$, which is a contradiction. Hence $k + \ell = n$. This proves (1.73), and the final statement is now evident.

Next since every vector in V is orthogonal to every vector in V^\perp , $V \subset V^\perp$. Next by (1.73) applied with V^\perp in place of V ,

$$\dim(V^\perp) + \dim((V^\perp)^\perp) = n = \dim(V) + \dim(V^\perp) .$$

It follows that

$$\dim((V^\perp)^\perp) = \dim(V) ,$$

and we have already seen that $V \subset (V^\perp)^\perp$. By Theorem 18, this proves (1.74). \square

Theorem 20 (Span and orthogonal complements). *Let S be any subset of \mathbb{R}^n . Then $(S^\perp)^\perp = \text{Span}(S)$.*

Proof. Let W be any subspace of \mathbb{R}^n such that $S \subset W$. By (1.72), $W^\perp \subset S^\perp$, and then by (1.72) again and (1.74), $(S^\perp)^\perp \subset (W^\perp)^\perp = W$. In particular, since $S \subset \text{Span}(S)$, $(S^\perp)^\perp \subset \text{Span}(S)$.

On the other hand, since subspaces are closed under taking linear combinations, and since $(S^\perp)^\perp$ contains S , $(S^\perp)^\perp$ contains every finite linear combination of vectors in S ; i.e., $\text{Span}(S) \subset (S^\perp)^\perp$.

Having proved both $(S^\perp)^\perp \subset \text{Span}(S)$ and $\text{Span}(S) \subset (S^\perp)^\perp$, we have the equality. \square

Corollary 3. *Let V be a subspace of \mathbb{R}^n of dimension d . Then there is an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n such that*

$$V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_d\}) = \{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}^\perp . \quad (1.75)$$

Proof. Theorem 19 provides the existence of an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n such that $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_d\})$ and $V^\perp = \text{Span}(\{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\})$. A vector \mathbf{v} is orthogonal to every vector in $\{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}$ if and only if it is orthogonal to every linear combination of these vectors, and hence

$$(\text{Span}(\{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}))^\perp = \{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}^\perp .$$

By Theorem 20, $V = V^{\perp\perp} = (\text{Span}(\{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}))^\perp$, and altogether we have $V = \{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}^\perp$. \square

The answers to most questions concerning a d dimensional subspace V of \mathbb{R}^n can be answered by considering the orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_d, \mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n provided by Corollary 3, since the two identities in (1.75) give both a system of equations for V and a parameterization of V .

The identity $V = \{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}^\perp$ says that V is exactly the solution set of the system of equations $\mathbf{u}_j \cdot \mathbf{x} = 0$ for $j = d+1, \dots, n$. When $n = 3$ and $d = 2$, this reduces to our standard geometric form of the equation of a plane through the origin in \mathbb{R}^3 , namely, $\mathbf{u}_3 \cdot \mathbf{x} = 0$.

Likewise, the identity $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_d\})$ says that the function P_V define by

$$P_V((t_1, \dots, t_d)) = \sum_{j=1}^d t_j \mathbf{u}_j \quad (1.76)$$

is a parameterization of V . Its inverse is the coordinate function C_V given by

$$C_V(\mathbf{v}) = (\mathbf{u}_1 \cdot \mathbf{v}, \dots, \mathbf{u}_d \cdot \mathbf{v}). \quad (1.77)$$

As we have seen in the discussion leading up to Theorem 17, C_V and P_V take orthonormal sets to orthonormal sets.

1.3.5 Higher dimensional analogs of lines and planes

What about the generalization of lines and planes that do not pass through the origin? Given any subset A of \mathbb{R}^n , and any $\mathbf{x}_0 \in \mathbb{R}^n$, define the set $A - \mathbf{x}_0$ by

$$A - \mathbf{x}_0 = \{\mathbf{a} - \mathbf{x}_0 : \mathbf{a} \in A\}.$$

That is, $A - \mathbf{x}_0$ is the sets of all vectors that one obtains by subtracting \mathbf{x}_0 from a vector in A . This operation “translates” the set A by moving each vector in it the same distance in the same direction. It simply shifts the position of the set without deforming it in any way.

If A is a plane in \mathbb{R}^3 , and \mathbf{x}_0 is any point in A , the set $A - \mathbf{x}_0$ is therefore again a plane, but now it contains $\mathbf{0}$, so that it is a plane through the origin. Moreover, any point \mathbf{x}_0 in the plane can serve as the “base point” \mathbf{x}_0 . Since translation relates general planes in \mathbb{R}^3 so two-dimensional subspaces of \mathbb{R}^3 , it is natural to use translation to relate their higher-dimensional analogs to subspaces of \mathbb{R}^n .

Definition 21 (Affine subsets of \mathbb{R}^n). *A subset $A \subset \mathbb{R}^n$ is an affine subset if and only if for some $\mathbf{x}_0 \in A$, $A - \mathbf{x}_0$ is a subspace V of \mathbb{R}^n .*

Lemma 4. *Let $A \subset \mathbb{R}^n$, and suppose that for some $\mathbf{x}_0 \in A$, $V := A - \mathbf{x}_0$ is a subspace of \mathbb{R}^n . Then for every $\mathbf{x}_1 \in A$, $V = A - \mathbf{x}_1$.*

Proof. Let \mathbf{x}_0 and \mathbf{x}_1 be two elements of A . Suppose $V = A - \mathbf{x}_0$ is a subspace. Since $\mathbf{x}_1 \in A$, the vector \mathbf{v} defined by $\mathbf{x}_1 - \mathbf{x}_0$ belongs to V , and $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{v}$. Therefore, for any $\mathbf{a} \in A$,

$$\mathbf{a} - \mathbf{x}_1 = \mathbf{a} - (\mathbf{v} + \mathbf{x}_0) = (\mathbf{a} - \mathbf{x}_0) - \mathbf{v} \in V$$

since the difference of two vectors in V belongs to V . □

Since the subspace V does not depend on the choice of the base point, the following definition makes sense:

Definition 22 (The dimension of an affine subsets of \mathbb{R}^n). *Let A be an affine subset of \mathbb{R}^n . The dimension of A , $\dim(A)$, is defined by $\dim(A) = \dim(V)$ where V is the subspace $A - \mathbf{x}_0$ for any $x_0 \in A$. If A is an affine subset of \mathbb{R}^n and $\dim(A) = 1$, A is a line in \mathbb{R}^n . If A is an affine subset of \mathbb{R}^n and $\dim(A) = 2$, A is a plane in \mathbb{R}^n . If $n > 3$ and A is an affine subset of \mathbb{R}^n and $\dim(A) = n-1$, A is a hyperplane in \mathbb{R}^n .*

Let A be any d -dimensional affine subset of \mathbb{R}^n , $1 \leq d \leq n-1$. (The cases $d=0$ and $d=n$ are trivial.) Fix any $\mathbf{x}_0 \in A$, and let V be the subspace $V = A - \mathbf{x}_0$. Then by Corollary 3, there is an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n so that $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_d\}) = \{\mathbf{u}_{d+1}, \dots, \mathbf{u}_n\}^\perp$. We have then seen in the discussion following Corollary 3 that V is exactly the solution set of the system of equations $\mathbf{u}_j \cdot \mathbf{x} = 0$ for $j = d+1, \dots, n$. Therefore, A is exactly the solution set of the system of equations $\mathbf{u}_j \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ for $j = d+1, \dots, n$. When $n=3$ and $d=2$, this reduces to our standard geometric form of the equation of a plane in \mathbb{R}^3 , namely, $\mathbf{u}_3 \cdot (\mathbf{x} - \mathbf{x}_0) = 0$.

Likewise, the identity $V = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_d\})$ says that the function sending (t_1, \dots, t_d) to $\sum_{j=1}^d t_j \mathbf{u}_j$ is a parameterization of $V = A - \mathbf{x}_0$, and therefore the function P_A defined by

$$P_A((t_1, \dots, t_d)) = \mathbf{x}_0 + \sum_{j=1}^d t_j \mathbf{u}_j \quad (1.78)$$

is a parameterization of A .

It is now a simple matter to show that if A is any nonempty affine subset of \mathbb{R}^n , then for all $\mathbf{p} \in \mathbb{R}^n$, there is a unique $\mathbf{q} \in A$ such that $\|\mathbf{q} - \mathbf{p}\| < \|\mathbf{x} - \mathbf{p}\|$ for all $\mathbf{x} \in A$, $\mathbf{x} \neq \mathbf{q}$. This is trivial if $d=0$ or $d=n$, so suppose that $1 \leq d \leq n-1$. Pick any $\mathbf{x}_0 \in A$, and let V be the subspace given by $V = A - \mathbf{x}_0$. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be an orthonormal basis of \mathbb{R}^n such that (1.75) is satisfied. The general point in A us then given by $P_A((t_1, \dots, t_d))$ with P_A given by (1.78). By the Pythagorean Theorem, the squared distance from this point to \mathbf{p} is

$$\|P_A((t_1, \dots, t_d)) - \mathbf{p}\|^2 = \sum_{k=1}^n ((P_A((t_1, \dots, t_d)) - \mathbf{p}) \cdot \mathbf{u}_k)^2.$$

For $k > d$, $(P_A((t_1, \dots, t_d)) - \mathbf{p}) \cdot \mathbf{u}_k = (\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_k$ which is independent of t_1, \dots, t_d . For $k \leq d$

$$(P_A((t_1, \dots, t_d)) - \mathbf{p}) \cdot \mathbf{u}_k = (\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_k + t_k$$

Evidently, we minimize the sum of squares by taking $t_j = -(\mathbf{x}_0 - \mathbf{p}) \cdot \mathbf{u}_j$ for $j = 1, \dots, d$, and only by this choice. Therefore,

$$\mathbf{q} = \mathbf{x}_0 + \sum_{j=1}^d ((\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u}_j) \mathbf{u}_j ,$$

and $\|\mathbf{q} - \mathbf{p}\|$, the distance form \mathbf{p} to A is given by $\|\mathbf{q} - \mathbf{p}\|^2 = \sum_{j=1}^d ((\mathbf{p} - \mathbf{x}_0) \cdot \mathbf{u}_j)^2$. We could go on to determine the distance between two affine subsets of \mathbb{R}^n , generalizing our results on the distance between two lines in \mathbb{R}^3 , but this will be easier once we have learned some more about solving systems of linear equations in Chapter 4. The main point that we want to make here, with which we conclude this chapter, is that “well-adapted” orthonormal bases provide the key to a great many geometric problem that we shall consider, no matter how many variables are involved.

1.4 Exercises

1.1 Let $\mathbf{a} = (3, -1)$, $\mathbf{b} = (2, 1)$ and $\mathbf{c} = (1, 3)$. Express \mathbf{a} as a linear combination of \mathbf{b} and \mathbf{c} . That is, find numbers s and t so that $\mathbf{a} = s\mathbf{b} + t\mathbf{c}$.

1.2 Let $\mathbf{a} = (5, 2)$, $\mathbf{b} = (2, -1)$ and $\mathbf{c} = (1, 1)$. Express each of these three vectors as a linear combination of the other two.

1.3 Let $\mathbf{x} = (1, 4, 8)$ and $\mathbf{y} = (1, 2, -2)$. Compute the lengths of each of these vectors, and the angle between them.

1.4 Let $\mathbf{x} = (4, 7, -4, 1, 2, -2)$ and $\mathbf{y} = (2, 1, 2, 2, -1, -1)$. Compute the lengths of each of these vectors, and the angle between them.

1.5 Let $\mathbf{x} = (4, 7, 4)$ and $\mathbf{y} = (2, 1, 2)$. Compute the lengths of each of these vectors, and the angle between them.

1.6 Let $\mathbf{x} = (-5, 2, -5)$ and $\mathbf{y} = (1, 2, 1)$. Is the angle between \mathbf{x} and \mathbf{y} acute or obtuse? Justify your answer.

1.7 Let

$$\mathbf{u}_1 = \frac{1}{9}(1, -4, -8) \quad \mathbf{u}_2 = \frac{1}{9}(8, 4, -1) \quad \text{and} \quad \mathbf{u}_3 = \frac{1}{9}(4, -7, 4).$$

(a) Show that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an orthonormal basis of \mathbb{R}^3 . Is it a right-handed orthonormal basis? Justify your answer.

(b) Find numbers y_1 , y_2 and y_3 such that

$$y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + y_3\mathbf{u}_3 = (10, 11, -11).$$

What are the lengths of the vectors $(10, 11, -11)$ and (y_1, y_2, y_3) ? Give calculations or an explanation in each case.

1.8 Let \mathbf{a} , \mathbf{b} and \mathbf{c} be any three vectors in \mathbb{R}^3 with $\mathbf{a} \neq \mathbf{0}$. Show that $\mathbf{b} = \mathbf{c}$ if and only if

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} \quad \text{and} \quad \mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{c}.$$

1.9 Let $\mathbf{a} = (1, 1, 1)$.

(a) Find a vector \mathbf{x} such that

$$\mathbf{x} \times \mathbf{a} = (-7, 2, 5) \quad \text{and} \quad \mathbf{x} \cdot \mathbf{a} = 0.$$

(b) There is no vector \mathbf{x} such that

$$\mathbf{x} \times \mathbf{a} = (1, 0, 0) \quad \text{and} \quad \mathbf{x} \cdot \mathbf{a} = 0.$$

Show that no such vector exists.

1.10 (a) Let $\mathbf{a} = (-1, 1, 2)$ and $\mathbf{b} = (2, -1, 1)$. Find all vectors \mathbf{x} , if any exist, such that

$$\mathbf{a} \times \mathbf{x} = (-2, 4, -3) \quad \text{and} \quad \mathbf{b} \cdot \mathbf{x} = 2.$$

If none exist, explain why this is the case.

- (b) Let $\mathbf{a} = (-1, 1, 2)$ and $\mathbf{b} = (2, -1, 1)$. Find all vectors \mathbf{x} , if any exist, such that

$$\mathbf{a} \times \mathbf{x} = (2, 4, 3) \quad \text{and} \quad \mathbf{b} \cdot \mathbf{x} = 2 .$$

If none exist, explain why this is the case.

- (c) Among all vectors \mathbf{x} such that $(-1, 1, 2) \times \mathbf{x} = (-2, 4, -3)$, find the one that is closest to $(1, 1, 1)$.

- 1.11** Let \mathbf{a} and \mathbf{b} be orthogonal vectors. Define a sequence of vectors $\{\mathbf{b}_n\}$ by

$$\mathbf{b}_{n+1} = \mathbf{a} \times \mathbf{b}_n \quad \text{and} \quad \mathbf{b}_0 = \mathbf{b} .$$

Show that for all positive integers m

$$\mathbf{b}_{2m} = (-1)^m \|\mathbf{a}\|^{2m} \mathbf{b} .$$

How do you have to adjust the formula if the hypothesis that \mathbf{a} and \mathbf{b} are orthogonal is dropped?

- (b) Let $\mathbf{a} = \frac{1}{3}(2, -1, 2)$ and $\mathbf{b} = (1, 1, 1)$. With \mathbf{b}_n defined as in part (a), compute \mathbf{b}_{99} .

- 1.12** Let \mathbf{a} , \mathbf{b} and \mathbf{c} be three non-zero vectors in \mathbb{R}^3 . Define a transformation \mathbf{f} from \mathbb{R}^3 to \mathbb{R}^3 by

$$\mathbf{f}(\mathbf{x}) = \mathbf{a} \times (\mathbf{b} \times (\mathbf{c} \times \mathbf{x})) .$$

Show that $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in \mathbb{R}^3$ if and only if \mathbf{b} is orthogonal to \mathbf{c} , and \mathbf{a} is a multiple of \mathbf{c} .

- 1.13** Let \mathbf{a} , \mathbf{b} and \mathbf{c} be three non-zero vectors in \mathbb{R}^3 . Show that

$$|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| \leq \|\mathbf{a}\| \|\mathbf{b}\| \|\mathbf{c}\|$$

and there is equality if and only if $\left\{ \frac{1}{\|\mathbf{a}\|} \mathbf{a}, \frac{1}{\|\mathbf{b}\|} \mathbf{b}, \frac{1}{\|\mathbf{c}\|} \mathbf{c} \right\}$ is orthonormal.

- 1.14** Let P_1 the plane through the three points $\mathbf{a}_1 = (1, 2, 1)$ $\mathbf{a}_2 = (-1, 2, -3)$ and $\mathbf{a}_3 = (2, -3, -2)$. Let P_2 denote the plane through the three points $\mathbf{b}_1 = (1, 1, 0)$ $\mathbf{b}_2 = (1, 0, 1)$ and $\mathbf{b}_3 = (0, 1, 1)$.

- (a) Find equations for the planes P_1 and P_2 .

- (b) Parameterize the line given by $P_1 \cap P_2$, and find the distance between this line and the point \mathbf{a}_1 .

- (c) Consider the line through \mathbf{b}_1 and \mathbf{b}_2 . Determine the point of intersection of this line with the plane P_1 .

- 1.15** Let $\mathbf{v} = (1, 4, 3)$. Find an orthonormal basis of \mathbb{R}^3 whose third vector is a multiple of \mathbf{v} .

- 1.16** Let $\mathbf{a} = (1, 4, 3)$ and $\mathbf{b} = (3, 2, 1)$. Find an orthonormal basis of \mathbb{R}^3 whose third vector is orthogonal to both \mathbf{a} and \mathbf{b} .

- 1.17** Consider the plane passing through the three points

$$\mathbf{p}_1 = (-2, 0, 2) \quad \mathbf{p}_2 = (1, -2, 2) \quad \text{and} \quad \mathbf{p}_3 = (3, -1, -2)$$

and the line passing through

$$\mathbf{z}_0 = (1, 4, -2) \quad \text{and} \quad \mathbf{z}_1 = (0, -3, 1)$$

- (a) Find a parametric representation $\mathbf{x}(s, t) = \mathbf{x}_0 + s\mathbf{v}_1 + t\mathbf{v}_2$ for the plane.
- (b) Find a parametric representation $\mathbf{z}(u) = \mathbf{z}_0 + u\mathbf{w}$ for the line.
- (c) Find an equation for the plane.
- (d) Find a system of equations for the line.
- (e) Find the points, if any, where the line intersects the plane.
- (f) Find the distance from \mathbf{p}_1 to the line.
- (g) Find the distance from \mathbf{z}_0 to the plane.

1.18 Consider the plane passing through the three points

$$\mathbf{p}_1 = (-1, -3, 0) \quad \mathbf{p}_2 = (5, 1, 2) \quad \text{and} \quad \mathbf{p}_3 = (0, -3, 4)$$

and the line passing through

$$\mathbf{z}_0 = (1, 1, -1) \quad \text{and} \quad \mathbf{z}_1 = (1, -2, 2)$$

- (a) Find a parametric representation $\mathbf{x}(s, t) = \mathbf{x}_0 + s\mathbf{v}_1 + t\mathbf{v}_2$ for the plane.
- (b) Find a parametric representation $\mathbf{z}(u) = \mathbf{z}_0 + u\mathbf{w}$ for the line.
- (c) Find an equation for the plane.
- (d) Find a system of equations for the line.
- (e) Find the points, if any, where the line intersects the plane.
- (f) Find the distance from \mathbf{p}_1 to the line.
- (g) Find the distance from \mathbf{z}_0 to the plane.

1.19 Consider the two lines parameterized by

$$(1, 1, 0) + t(1, -1, 2) \quad \text{and} \quad (2, 0, 2) + s(-1, 1, 0).$$

- (a) These lines intersect. Find the point of intersection.
- (b) Find an equation for the plane P containing these two lines.

1.20 Consider the plane given by

$$2x - y + 3z = 4.$$

Let $\mathbf{p} = (-1, -3, 0)$. What is the distance from \mathbf{p} to the plane?

1.21 Consider the plane given by

$$x - 3y + z = 2.$$

Let $\mathbf{p} = (-2, -5, 1)$. What is the distance from \mathbf{p} to the plane?

1.22 Consider the line ℓ given by

$$2x - y + 3z = 4$$

$$x + y + z = 2.$$

Find a parametric representation of the line obtained by reflecting this line through the plane $x + 3y - z = 1$. That is; the outgoing line should have as its base point the intersection of the incoming line and the plane, and its direction vector should be $\mathbf{h}_\mathbf{u}(\mathbf{v})$ where \mathbf{v} is the incoming direction vector, and \mathbf{u} is a unit normal vector to the plane.

1.23 Consider the line ℓ given by

$$\begin{aligned}x - 3y + z &= 2 \\2y + z &= 3.\end{aligned}$$

Find a parametric representation of the line obtained by reflecting this line through the plane $x + 2y - z = 1$. Find a parametric representation of the line obtained by reflecting this line through the plane $x + 3y - z = 1$. (See the previous exercise.)

1.24 Let $\mathbf{x} = (5, 2, 4, 2)$. Let \mathbf{u} be a unit vector such that $\mathbf{h}_\mathbf{u}(\mathbf{x})$ is a multiple of \mathbf{e}_1 . What are the possible values of this multiple? Find four such unit vectors \mathbf{u} .

1.25 Consider two lines in \mathbb{R}^3 given parametrically by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$ and $\mathbf{x}_2(t) = \mathbf{x}_2 + t\mathbf{v}_2$ where

$$\mathbf{x}_1 = (1, 2, 1) \quad \mathbf{x}_2 = (1, -1, 0) \quad \mathbf{v}_1 = (1, 0, -1) \quad \text{and} \quad \mathbf{v}_2 = (2, 1, 1).$$

Compute the distance between these two lines.

1.26 Consider two lines in \mathbb{R}^3 given parametrically by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$ and $\mathbf{x}_2(t) = \mathbf{x}_2 + t\mathbf{v}_2$ where

$$\mathbf{x}_1 = (1, 2, 3) \quad \mathbf{x}_2 = (2, 0, 2) \quad \mathbf{v}_1 = (1, 2, 2) \quad \text{and} \quad \mathbf{v}_2 = (-2, 1, 1).$$

Compute the distance between these two lines.

1.27 Consider two lines in \mathbb{R}^3 given parametrically by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$ and $\mathbf{x}_2(t) = \mathbf{x}_2 + t\mathbf{v}_2$ where

$$\mathbf{x}_1 = (3, 2, 1) \quad \mathbf{x}_2 = (1, 1, -1) \quad \mathbf{v}_1 = (3, -5, -1) \quad \text{and} \quad \mathbf{v}_2 = (-1, 3, 3).$$

Find the point on the first line that is closest to the second line, the point on the second line that is closest to the first line, and the distance between these two lines.

1.28 Consider two lines in \mathbb{R}^3 given parametrically by $\mathbf{x}_1(s) = \mathbf{x}_1 + s\mathbf{v}_1$ and $\mathbf{x}_2(t) = \mathbf{x}_2 + t\mathbf{v}_2$ where

$$\mathbf{x}_1 = (1, 2, -1) \quad \mathbf{x}_2 = (2, 1, -5) \quad \mathbf{v}_1 = (1, -4, -2) \quad \text{and} \quad \mathbf{v}_2 = (1, 1, -2).$$

Find the point on the first line that is closest to the second line, the point on the second line that is closest to the first line, and the distance between these two lines.

1.29 Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be given by

$$\mathbf{u}_1 = \frac{1}{3}(1, 2, -2) \quad \mathbf{u}_2 = \frac{1}{3}(2, 1, 2) \quad \text{and} \quad \mathbf{u}_3 = \frac{1}{3}(2, -2, -1).$$

(a) Verify whether $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is, or is not, an orthonormal basis of \mathbb{R}^3 .

(b) Find a unit vector \mathbf{u} so that $\mathbf{h}_\mathbf{u}(\mathbf{u}_1) = \mathbf{e}_1$.

(c) With this same choice of \mathbf{u} , compute $\mathbf{h}_\mathbf{u}(\mathbf{u}_2)$ and $\mathbf{h}_\mathbf{u}(\mathbf{u}_3)$.

1.30 Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be given by

$$\mathbf{u}_1 = \frac{1}{9}(1, 4, 8) \quad \mathbf{u}_2 = \frac{1}{9}(8, -4, 1) \quad \text{and} \quad \mathbf{u}_3 = \frac{1}{9}(4, 7, -4) .$$

- (a) Verify whether $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is, or is not, an orthonormal basis of \mathbb{R}^3 .
- (b) Find a unit vector \mathbf{u} so that $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_1) = \mathbf{e}_1$.
- (c) With this same choice of \mathbf{u} , compute $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_2)$ and $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_3)$.

1.31 Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be given by

$$\mathbf{u}_1 = \frac{1}{3}(1, 2, -2) \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}(0, 1, 1) \quad \text{and} \quad \mathbf{u}_3 = \frac{1}{3\sqrt{2}}(4, 1, -1) .$$

- (a) Verify whether $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is, or is not, an orthonormal basis of \mathbb{R}^3 .
- (b) Find a unit vector \mathbf{u} so that $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_1) = \mathbf{e}_1$.
- (c) With this same choice of \mathbf{u} , compute $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_2)$ and $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_3)$.

1.32 Let V_1 and V_2 be two subspaces of \mathbb{R}^n .

- (a) Show that $V_1 \cap V_2$ is a subspace of \mathbb{R}^n .
- (b) Define $V_1 + V_2$ to be the set of all vectors $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{z} = \mathbf{x} + \mathbf{y}$ for some $\mathbf{x} \in V_1$ and some $\mathbf{y} \in V_2$. Show that $V_1 + V_2$ is a subspace of \mathbb{R}^n .

1.33 Let V_1 and V_2 be two subspaces of \mathbb{R}^n . Using the results and notation from the previous exercise, show that

$$\dim(V_1 \cap V_2) + \dim(V_1 + V_2) = \dim(V_1) + \dim(V_2) .$$

1.34 For $n > 3$, an $n - 1$ dimensional V subspace of \mathbb{R}^n is called a *hyperplane through the origin*. The orthogonal complement V^\perp is a one dimensional subspace. In this case, starting from the equation of the hyperplane, it is easy to write down an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_{n-1}, \mathbf{u}_n\}$ such that $\{\mathbf{u}_1, \dots, \mathbf{u}_{n-1}\}$ is an orthonormal basis of V , and such that $\{\mathbf{u}_n\}$ is an orthonormal basis of V^\perp :

Let \mathbf{a} be a non-zero vector in \mathbb{R}^n , and let V be the solution set of $\mathbf{a} \cdot \mathbf{x} = 0$. Define the unit vector $\mathbf{w} = (1/\|\mathbf{a}\|)\mathbf{a}$. Let \mathbf{u} be a unit vector such that the Householder reflection $\mathbf{h}_{\mathbf{u}}$ satisfies

$$\mathbf{h}_{\mathbf{u}}(\mathbf{w}) = \mathbf{e}_n .$$

Define

$$\mathbf{u}_j = \mathbf{h}_{\mathbf{u}}(\mathbf{e}_j) \quad \text{for } j = 1, \dots, n .$$

Show that with these definitions, $\{\mathbf{u}_1, \dots, \mathbf{u}_{n-1}\}$ is an orthonormal basis of V , and such that $\{\mathbf{u}_n\}$ is an orthonormal basis of V^\perp .

1.35 We use the notation and results of the previous exercise. Consider the hyperplane V through the origin in \mathbb{R}^4 given by

$$2x + 2y - 7z + 4w = 0 .$$

Let $\mathbf{b} = (1, 2, 0, 2)$. Find the point $\mathbf{x} \in V$ that is closest to \mathbf{b} and find the distance between \mathbf{b} and V .

1.36 Show that for all vectors \mathbf{a} , \mathbf{b} , \mathbf{c} and \mathbf{d} in \mathbb{R}^3 ,

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}) .$$

1.37 Show that for all \mathbf{a} , \mathbf{b} and \mathbf{c} in \mathbb{R}^3 ,

$$(\mathbf{b} \times \mathbf{c}) \cdot [(\mathbf{c} \times \mathbf{a}) \times (\mathbf{a} \times \mathbf{b})] = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|^2 .$$

Chapter 2

DESCRIPTION OF MOTION

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

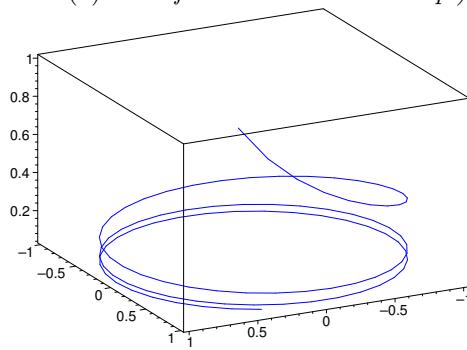
In many ways, the simplest multivariable functions are functions from \mathbb{R} to \mathbb{R}^n for some $n \geq 1$. These are functions that have one input variable (one independent variable), and n output variables (n dependent variables).

If $n = 2$ or if $n = 3$, we can think of the output variable \mathbf{x} as the coordinate vector of a point moving in \mathbb{R}^2 or \mathbb{R}^3 , and we can think of the input variable t as the time so that the function gives us the location of a moving point at time t . We then write the function as $\mathbf{x}(t)$. Such functions are also called *vector valued functions of a real variable*.

Example 25. Consider the function $\mathbf{x}(t)$ of the real variable t with values in \mathbb{R}^3 given by

$$\mathbf{x}(t) = (\cos(t), \sin(t), 1/t). \quad (2.1)$$

Here is a three dimensional plot of the curve traced out by $\mathbf{x}(t)$ as for $1 \leq t \leq 20$. (Since $x_3(t) = 1/t$ is a decreasing function of t , the $\mathbf{x}(0)$ end of the curve is at the top.)



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

Vector valued functions of one real variable that describe particle motion usually have certain *regularity properties*: For example, particle motions are usually at least *continuous*:

Definition 23 (Convergence and continuity in \mathbb{R}^n). A sequence of vectors $\{\mathbf{x}_j\}$ in \mathbb{R}^n converges to $\mathbf{z} \in \mathbb{R}^n$ in case for each $\epsilon > 0$, there is a natural number N_ϵ so that

$$j \geq N_\epsilon \quad \Rightarrow \quad \|\mathbf{x}_j - \mathbf{z}\| \leq \epsilon ,$$

in which case we say that \mathbf{z} is the limit of the sequence and write

$$\lim_{n \rightarrow \infty} \mathbf{x}_j = \mathbf{z} .$$

A function $\mathbf{x}(t)$ defined on an open interval $(a, b) \subset \mathbb{R}$ with values in \mathbb{R}^n is continuous at $t_0 \in (a, b)$ in case for each $\epsilon > 0$, there is a real number $\delta_\epsilon > 0$ so that

$$|t - t_0| \leq \delta_\epsilon \quad \Rightarrow \quad \|\mathbf{x}(t) - \mathbf{x}(t_0)\| \leq \epsilon , \quad (2.2)$$

in which case we write $\lim_{t \rightarrow t_0} \mathbf{x}(t) = \mathbf{x}(t_0)$. The function $\mathbf{x}(t)$ is said to be continuous if it is continuous at each point in its domain. Such a function is often called a curve in \mathbb{R}^n .

Checking continuity for functions from (a, b) to \mathbb{R}^n can be done one coordinate function at a time:

Theorem 21. A vector valued function $\mathbf{x}(t)$ of a real variable t is continuous at t_0 if and only if each of its entry functions $x_j(t)$ is a continuous at t_0 .

This theorem means we can use everything we know about continuity for real valued functions of one real variable to answer questions about vector valued functions of one real variable. For example, we know for single variable functions that if limits exist, they are unique: As $t \rightarrow t_0$, $x(t)$ cannot converge to two different numbers x and y . It follows that if $\lim_{t \rightarrow t_0} \mathbf{x}(t) = \mathbf{x}$ and $\lim_{t \rightarrow t_0} \mathbf{x}(t) = \mathbf{y}$, then $\mathbf{y} = \mathbf{x}$. It makes sense to talk about “the limit” whenever limits exist.

The proof of Theorem 21 turns on some very simple general observations that we review before the giving the proof. Observe that if $\{a_1, \dots, a_n\}$ is any set of n non-negative numbers, then

$$\max\{a_1, \dots, a_n\} \leq \sum_{k=1}^n a_k \leq n (\max\{a_1, \dots, a_n\}) . \quad (2.3)$$

Now let \mathbf{x} and \mathbf{y} be any two vectors in \mathbb{R}^n . Apply (2.3) with $a_j = |x_j - y_j|^2$, and recall that since the function $f(a) = a^2$ is monotone for $a > 0$, $\max\{a^2, b^2\} = (\max\{a, b\})^2$ for all $a, b \geq 0$. We obtain

$$\max\{|x_1 - y_1|, \dots, |x_n - y_n|\} \leq \|\mathbf{x} - \mathbf{y}\| \leq \sqrt{n} (\max\{|x_1 - y_1|, \dots, |x_n - y_n|\}) . \quad (2.4)$$

Proof of Theorem 21. Suppose that $\mathbf{x}(t)$ is continuous at t_0 , meaning that $\lim_{t \rightarrow t_0} \|\mathbf{x}(t) - \mathbf{x}(t_0)\| = 0$. Then by the inequality on the left in (2.4), for each j , $\lim_{t \rightarrow t_0} |x_j(t) - x_j(t_0)| = 0$, which means that $x_j(t)$ is continuous at t_0 . On the other hand, suppose that for each j , $x_j(t)$ is continuous at t_0 . Then $\lim_{t \rightarrow t_0} |x_j(t) - x_j(t_0)| = 0$ for each j , and consequently

$$\lim_{t \rightarrow t_0} (\sqrt{n} \max\{|x_1(t) - x_1(t_0)|, \dots, |x_n(t) - x_n(t_0)|\}) = 0 .$$

Then by the inequality on the right in (2.4), (2.2) is valid. \square

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

The motion of physical particles is only continuous, but differentiable. In fact, as we shall explain later, as a consequence of Newton's second law, as long as no infinite forces act on a particle, its motion will be described by a curve that is at least twice differentiable.

To say that $\mathbf{x}(t)$ is differentiable means, roughly speaking, that if you observe the motion described by $\mathbf{x}(t)$ over a sufficiently short time interval, it looks like constant speed motion along a parameterized line.

A parameterized line in \mathbb{R}^n is a function from \mathbb{R} to \mathbb{R}^n of the form

$$\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{v} \quad (2.5)$$

for fixed vectors \mathbf{x}_0 and \mathbf{v} in \mathbb{R}^n , with $\mathbf{v} \neq \mathbf{0}$. For any $t_0 \in \mathbb{R}$, $\mathbf{x}(t) = \mathbf{x}(t_0) + (t - t_0)\mathbf{v}$ for all $t \in \mathbb{R}$.

Remark 3. Let \mathbf{x}_0 and \mathbf{v} be vectors in \mathbb{R}^n with $\mathbf{v} \neq \mathbf{0}$. The sets traced out by

$$\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{v} \quad \text{and} \quad \tilde{\mathbf{x}}(t) = \mathbf{x}_0 + t^3\mathbf{v}$$

as t varies over \mathbb{R} are the same line in \mathbb{R}^n , as a subset of \mathbb{R}^n , but they are different parameterizations of the same line. Note that $\mathbf{x}'(t) = \mathbf{v}$ is constant, while $\tilde{\mathbf{x}}'(t) = 3t^2\mathbf{v}$ varies with t . When we refer to a function $\mathbf{x}(t)$ as a parameterized line, we always mean a parameterization of the form (2.5) for which $\mathbf{x}'(t)$ is constant unless some other more complicated parameterization is explicitly specified.

Lemma 5. Let $\mathbf{x}(t)$ be a parameterized curve in \mathbb{R}^n defined for $t \in (a, b)$. Fix $t_0 \in (a, b)$, and define $\mathbf{x}_0 = \mathbf{x}(t_0)$. There is at most one vector $\mathbf{v} \in \mathbb{R}^n$ such that

$$\lim_{t \rightarrow 0} \frac{\|\mathbf{x}(t) - \mathbf{y}(t)\|}{|t - t_0|} = 0 . \quad (2.6)$$

is satisfied with $\mathbf{y}(t) = \mathbf{x}_0 + (t - t_0)\mathbf{v}$.

Proof. Suppose that $\mathbf{y}(t) = \mathbf{x}_0 + (t - t_0)\mathbf{v}$ and $\mathbf{z}(t) = \mathbf{x}_0 + (t - t_0)\mathbf{w}$ are two parameterized lines through $\mathbf{x}_0 = \mathbf{x}(t_0)$. Suppose that

$$\lim_{t \rightarrow t_0} \frac{\|\mathbf{x}(t) - \mathbf{y}(t)\|}{|t - t_0|} = 0 \quad \text{and} \quad \lim_{t \rightarrow t_0} \frac{\|\mathbf{x}(t) - \mathbf{z}(t)\|}{|t - t_0|} = 0 . \quad (2.7)$$

Then since

$$\|\mathbf{y}(t) - \mathbf{z}(t)\| = \|(\mathbf{y}(t) - \mathbf{x}(t)) + (\mathbf{x}(t) - \mathbf{z}(t))\| \leq \|\mathbf{x}(t) - \mathbf{y}(t)\| + \|\mathbf{x}(t) - \mathbf{z}(t)\| ,$$

it follows that

$$\lim_{t \rightarrow t_0} \frac{\|\mathbf{y}(t) - \mathbf{z}(t)\|}{|t - t_0|} = 0 . \quad (2.8)$$

But $\frac{\|\mathbf{y}(t) - \mathbf{z}(t)\|}{|t - t_0|} = \|\mathbf{v} - \mathbf{w}\|$, and so whenever (2.7) is true, $\mathbf{v} = \mathbf{w}$. \square

We now define the important class of functions for which such a line exists.

Definition 24 (Differentiable curves). Let $\mathbf{x}(t)$ be an \mathbb{R}^n valued function of the variable t . We say that $\mathbf{x}(t)$ is differentiable at $t = t_0$ in case there is a parameterized line $\mathbf{x}_0 + (t - t_0)\mathbf{v}$ such that $\mathbf{x}_0 = \mathbf{x}(t_0)$ and

$$\lim_{t \rightarrow t_0} \frac{\|\mathbf{x}(t) - (\mathbf{x}_0 + (t - t_0)\mathbf{v})\|}{|t - t_0|} = 0.$$

The unique vector \mathbf{v} for which this is true is the derivative of the function $\mathbf{x}(t)$ at $t = t_0$, and it is denoted by $\mathbf{x}'(t_0)$. The parameterized line $\mathbf{y}(t) = \mathbf{x}(t_0) + (t - t_0)\mathbf{x}'(t_0)$ is called the tangent line to the curve $\mathbf{x}(t)$ at $t = t_0$.

How do we check for differentiability, and supposing that $\mathbf{x}(t)$ is differentiable at $t = t_0$, how do we compute the derivative $\mathbf{v} = \mathbf{x}'(t_0)$? For any choice of \mathbf{v} , let $\mathbf{y}(t) = \mathbf{x}(t_0) + (t - t_0)\mathbf{v}$, and note that

$$\frac{\|\mathbf{x}(t) - \mathbf{y}(t)\|}{|t - t_0|} = \frac{\|\mathbf{x}(t) - \mathbf{x}(t_0) - (t - t_0)\mathbf{v}\|}{|t - t_0|} = \left\| \frac{\mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0} - \mathbf{v} \right\|.$$

Therefore, (2.6) is true for this choice of \mathbf{v} if and only if

$$\lim_{t \rightarrow t_0} \frac{\mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0} = \mathbf{v},$$

and then by Theorem 21, this is the case if and only if for each $j = 1, \dots, n$,

$$\lim_{t \rightarrow t_0} \frac{x_j(t) - x_j(t_0)}{t - t_0} = v_j. \quad (2.9)$$

Summarizing $\mathbf{x}(t)$ is differentiable at $t = t_0$ if and only if each of its coordinate functions $x_j(t)$ is differentiable at $t = t_0$ in the usual single variable sense, and in that case, if we define $v_j = x'_j(t_0)$ for each j , then $\mathbf{v} = (v_1, \dots, v_n) = \mathbf{x}'(t_0)$. Thus as far as computation of derivatives *per se* is concerned, there is nothing really new going on here: We just differentiate the entries of a vector valued function of t separately, one at a time.

Example 26 (Computing the derivative of a vector valued function of t). Let $\mathbf{x}(t)$ be given by (2.1). Then for any $t \neq 0$,

$$\mathbf{x}'(t) = (-\sin(t), \cos(t), -1/t^2).$$

Example 27 (A tangent line). Consider the vector valued function

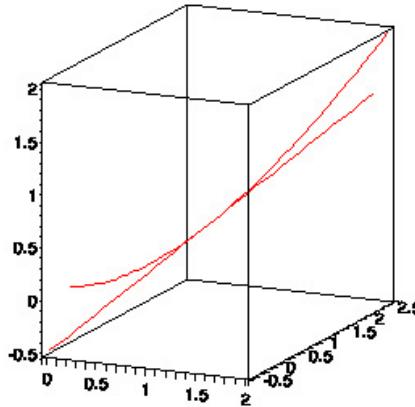
$$\mathbf{x}(t) = (x(t), y(t), z(t)) = (t, 2^{3/2}t^{3/2}/3, t^2/2).$$

To compute $\mathbf{x}'(t)$, compute $x'(t) = 1$, $y'(t) = 2^{1/2}t^{1/2}$ and $z'(t) = t$. Hence $\mathbf{x}'(t) = (1, 2^{1/2}t^{1/2}, t)$.

We now compute the tangent line at $t_0 = 1$. This is parameterized by

$$\mathbf{x}(1) + (t - 1)\mathbf{x}'(1) = (1, 2^{3/2}/3, 1/2) + (t - 1)(1, 2^{1/2}, 1). \quad (2.10)$$

Here is a graph showing both the curve $\mathbf{x}(t)$ and the tangent line $\mathbf{x}(1) + (t - 1)\mathbf{x}'(1)$ for $-1 \leq t \leq 1$:



As you can see, the straight line is a very close match to the curve for $t \approx t_0 = 1$: Both curves pass through $\mathbf{x}(1)$ at $t = 1$, and they do so moving in the same direction. What you cannot see in this static picture is that they also move through this point at the same speed. That is, the linear motion and the curved motion “track each other” very well.

Had we “zoomed in more”, and shown the two graphs only for $-0.1 \leq t \leq 0.1$, the two graphs would have been pretty much indistinguishable. If we keep “zooming in” the two curves, and the motions along them, will look more and more “equivalent”.

Because we differentiate vectors entry by entry without mixing the entries up in any way, familiar rules for differentiating scalar valued functions hold for vector valued functions as well. In particular, the derivative of a sum is still the sum of the derivatives, etc.:

$$(\mathbf{x}(t) + \mathbf{y}(t))' = \mathbf{x}'(t) + \mathbf{y}'(t) . \quad (2.11)$$

We now turn to product rules. There are now several types of products to consider: product rules for scalar-vector multiplication and product rules for both the dot and cross products.

Theorem 22 (Differentiating dot and cross products). *Suppose that $\mathbf{v}(t)$ and $\mathbf{w}(t)$ are differentiable vector valued functions for t in (a, b) with values in \mathbb{R}^n , and that both of these functions are differentiable at $t_0 \in (a, b)$. Then $\mathbf{v}(t) \cdot \mathbf{w}(t)$ is differentiable at t_0 and*

$$\frac{d}{dt} (\mathbf{v}(t) \cdot \mathbf{w}(t)) \Big|_{t=t_0} = \mathbf{v}'(t_0) \cdot \mathbf{w}(t_0) + \mathbf{v}(t_0) \cdot \mathbf{w}'(t_0) . \quad (2.12)$$

Also, if $n = 3$ so that the cross product is defined, $\mathbf{v}(t) \times \mathbf{w}(t)$ is differentiable at t_0 and

$$\frac{d}{dt} (\mathbf{v}(t) \times \mathbf{w}(t)) \Big|_{t=t_0} = \mathbf{v}'(t_0) \times \mathbf{w}(t_0) + \mathbf{v}(t_0) \times \mathbf{w}'(t_0) . \quad (2.13)$$

Proof. By definition we have

$$\frac{d}{dt} (\mathbf{v}(t) \cdot \mathbf{w}(t)) \Big|_{t=t_0} = \lim_{h \rightarrow 0} \frac{1}{h} (\mathbf{v}(t_0 + h) \cdot \mathbf{w}(t_0 + h) - \mathbf{v}(t_0) \cdot \mathbf{w}(t_0)) . \quad (2.14)$$

Now we use the device of “adding and subtracting” that is used to prove the single variable product rule to write

$$\begin{aligned}\mathbf{v}(t_0 + h) \cdot \mathbf{w}(t_0 + h) - \mathbf{v}(t_0) \cdot \mathbf{w}(t_0) &= [\mathbf{v}(t_0 + h) - \mathbf{v}(t_0)] \cdot \mathbf{w}(t_0 + h) \\ &= \mathbf{v}(t_0) \cdot [\mathbf{w}(t_0 + h) - \mathbf{w}(t_0)]\end{aligned}\quad (2.15)$$

Note that this identity is true because we have simply added $\mathbf{v}(t_0) \cdot \mathbf{w}(t_0 + h)$ in the first line on the right, and subtracted it back out in the second. The advantage is that now in each term, only one of \mathbf{v} and \mathbf{w} is changing.

Combining (2.14) and (2.15), we have

$$\begin{aligned}\frac{d}{dt} (\mathbf{v}(t) \cdot \mathbf{w}(t)) \Big|_{t=t_0} &= \lim_{h \rightarrow 0} \left(\frac{\mathbf{v}(t_0 + h) - \mathbf{v}(t_0)}{h} \cdot \mathbf{w}(t_0 + h) \right) \\ &\quad + \lim_{h \rightarrow 0} \left(\mathbf{v}(t_0) \cdot \frac{\mathbf{w}(t_0 + h) - \mathbf{w}(t_0)}{h} \right) \\ &= \mathbf{v}'(t_0) \cdot \mathbf{w}(t_0) + \mathbf{v}(t_0) \cdot \mathbf{w}'(t_0)\end{aligned}$$

The proof for cross products is exactly the same; simply replace each dot product with a cross product in the lines above. \square

Finally there is the case of the product rule for scalar vector multiplication. If $g(t)$ is a real valued function defined on (a, b) , and $\mathbf{x}(t)$ is an \mathbb{R}^n valued function defined on (a, b) , and if both are differentiable at $t_0 \in (a, b)$, then

$$\frac{d}{dt} (g(t)\mathbf{x}(t)) \Big|_{t=t_0} = g'(t_0)\mathbf{x}(t_0) + g(t_0)\mathbf{x}'(t_0). \quad (2.16)$$

We leave the proof of this to the reader - treat the components one at a time.

We next present a simple consequence of Theorem 22 that we shall frequently use.

Theorem 23 (Orthogonality for constant magnitude curves). *Let $\mathbf{w}(t)$ be a differentiable curve in \mathbb{R}^n defined on (a, b) such that for some $\varrho > 0$, $\|\mathbf{w}(t)\| = \varrho$ for all $t \in (a, b)$. That is, suppose the vector $\mathbf{w}(t)$ has constant magnitude. Then for all $t \in (a, b)$,*

$$\mathbf{w}(t) \cdot \mathbf{w}'(t) = 0.$$

Proof.

$$0 = \frac{d}{dt} \varrho^2 = \frac{d}{dt} \mathbf{w}(t) \cdot \mathbf{w}(t) = 2\mathbf{w}(t) \cdot \mathbf{w}'(t).$$

\square

2.1.3 Velocity and acceleration

Let $\mathbf{x}(t)$ be a function defined on (a, b) with values in \mathbb{R}^n . If $n = 3$, we can think of $\mathbf{x}(t)$ as representing the *position* of a point particle in physical space at time t . In this case it is natural to call $\mathbf{x}'(t)$ *velocity*, and we shall do for all values of n . The velocity gives the rate of change of the

position, or more generally the *configuration* of some physical system more complicated than a point particle.

If the function $\mathbf{v}(t) = \mathbf{x}'(t)$ is differentiable, then $\mathbf{v}'(t)$ is called the *acceleration*, and is often denoted by $\mathbf{a}(t)$, so that $\mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{x}''(t)$. In this case we say that $\mathbf{x}(t)$ is *twice differentiable*, and *twice continuously differentiable* in case $\mathbf{a}(t) = \mathbf{x}''(t)$ is continuous. Thus, the acceleration is the second time derivative of the position (if it is twice differentiable) and gives the rate of change of the velocity vector.

For a parameterized line $\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{v}$, we have $\mathbf{v}(t) = \mathbf{x}'(t) = \mathbf{v}$, and so the velocity is constant. Therefore, $\mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{0}$ for all t . That is, parameterized lines have zero acceleration. For parameterized circles, matters are different.

Example 28 (Parameterized circle in \mathbb{R}^3). *Let \mathbf{c} and \mathbf{a} be vectors on \mathbb{R}^3 with $\mathbf{a} \neq \mathbf{0}$. Let $\varrho > 0$. Consider the system of equations*

$$\begin{aligned}\|\mathbf{x} - \mathbf{c}\|^2 &= \varrho^2 \\ \mathbf{a} \cdot (\mathbf{x} - \mathbf{c}) &= 0.\end{aligned}\tag{2.17}$$

The first equation in this system is the equation for the sphere of radius ϱ centered at \mathbf{c} . The second is the equation of the plane passing through \mathbf{c} with normal direction along \mathbf{a} . The intersection of the plane and the sphere is a circle of radius ϱ in \mathbb{R}^3 . In fact, if you “slice” a sphere by a plane through the center of the sphere, you get a so-called great circle on the sphere. Segments of great circles have a “geodesic property” that we shall study later in this chapter.

In the mean time, let us parameterize the solutions set to (2.17). Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be an orthonormal basis of \mathbb{R}^3 such that $\mathbf{u}_3 = \|\mathbf{a}\|^{-1}\mathbf{a}$. We have seen how to construct such an orthonormal basis.

Note that $\mathbf{a} \cdot (\mathbf{x} - \mathbf{c}) = 0$ if and only if $\mathbf{u}_3 \cdot (\mathbf{x} - \mathbf{c}) = 0$, and so \mathbf{x} satisfies the second equation in (2.17) if and only if $\mathbf{x} - \mathbf{c} = ((\mathbf{x} - \mathbf{c}) \cdot \mathbf{u}_1)\mathbf{u}_1 + ((\mathbf{x} - \mathbf{c}) \cdot \mathbf{u}_2)\mathbf{u}_2$. Then \mathbf{x} also satisfies the first equation in (2.17) if and only if

$$((\mathbf{x} - \mathbf{c}) \cdot \mathbf{u}_1)^2 + ((\mathbf{x} - \mathbf{c}) \cdot \mathbf{u}_2)^2 = \varrho^2.$$

Since all of the solutions of $a^2 + b^2 = \varrho^2$ are given by $(a, b) = \varrho(\cos \theta, \sin \theta)$ for some $0 \leq \theta < 2\pi$, we must have that $(\mathbf{x} - \mathbf{c}) \cdot \mathbf{u}_1 = \cos \theta$ and $(\mathbf{x} - \mathbf{c}) \cdot \mathbf{u}_2 = \sin \theta$ for some $0 \leq \theta < 2\pi$,

Thus,

$$\mathbf{x}(\theta) := \mathbf{c} + \varrho \cos \theta \mathbf{u}_1 + \varrho \sin \theta \mathbf{u}_2$$

for $0 \leq \theta < 2\pi$ is a parameterization of the solution set of (2.17).

Now suppose that the angle θ is increasing at a constant rate; i.e., that for some $\omega > 0$, the angle $\theta(t)$ at time t is given by $\theta(t) = \omega(t - t_0)$ for some $t_0 \in \mathbb{R}$. Then writing $\mathbf{x}(t)$ to denote $\mathbf{x}(\theta(t))$, we have

$$\mathbf{x}(t) = \mathbf{c} + \varrho[\cos(\omega(t - t_0))\mathbf{u}_1 + \sin(\omega(t - t_0))\mathbf{u}_2].$$

With this parameterization $\mathbf{x}(t_0) = \mathbf{c} + \varrho\mathbf{u}_1$, $\mathbf{x}(t_0 + \pi/2\omega) = \mathbf{c} + \varrho\mathbf{u}_2$, and the period of the motion is $2\pi/\omega$.

Now, let us compute the velocity and acceleration of $\mathbf{x}(t)$. We compute:

$$\mathbf{v}(t) = \mathbf{x}'(t) = \varrho\omega[-\sin(\omega(t - t_0))\mathbf{u}_1 + \cos(\omega(t - t_0))\mathbf{u}_2] ,$$

and

$$\mathbf{a}(t) = \mathbf{v}'(t) = -\varrho\omega^2[\cos(\omega(t - t_0))\mathbf{u}_1 + \sin(\omega(t - t_0))\mathbf{u}_2] .$$

Note that

$$\|\mathbf{v}(t)\| = \varrho\omega \quad \text{and} \quad \|\mathbf{a}(t)\| = \varrho\omega^2 . \quad (2.18)$$

Since $\|\mathbf{v}(t)\|$ is constant, it follows from Theorem 23 that $\mathbf{v}(t)$ and $\mathbf{a}(t)$ are orthogonal for each t , as you can readily check.

In the previous example, speed was constant, and so the acceleration was non-zero only because the direction of the velocity vector was changing, not its magnitude. In general it is useful to separate the acceleration vector into two components, one having to do with the rate of change of the speed, and the other having to do with the rate of change of the direction of motion.

Definition 25 (Speed and the unit tangent vector). *The magnitude of the velocity vector is called the speed. We denote it by $v(t)$. That is,*

$$v(t) = |\mathbf{v}(t)| .$$

Provided that $v(t) \neq 0$, we can define a unit vector valued function $\mathbf{T}(t)$ by

$$\mathbf{T}(t) = \frac{1}{v(t)}\mathbf{v}(t) \quad \text{so that} \quad \mathbf{v}(t) = v(t)\mathbf{T}(t) .$$

The vector $\mathbf{T}(t)$ is called the unit tangent vector at time t . It specifies the instantaneous direction of motion.

Example 29 (Speed and the unit tangent vector). Let $\mathbf{x}(t) = (t, 2^{3/2}t^{3/2}/3, t^2/2)$ for $t > 0$ as in Example 26. There we found that $\mathbf{v}(t) = (1, 2^{1/2}t^{1/2}, t)$, and so the speed $v(t)$ is given by

$$v(t) = \sqrt{1 + 2t + t^2} = 1 + t .$$

which is strictly positive for all $t > 0$, and then we have

$$\mathbf{T}(t) = \frac{1}{1+t}(1, 2^{1/2}t^{1/2}, t) .$$

Theorem 24. Let $\mathbf{x}(t)$ be a twice differentiable curve, and suppose that the speed $v(t)$ is nonzero on some open interval (b, c) so that $\mathbf{T}(t)$ is defined for all t in this interval. Let $\mathbf{a}(t) = \mathbf{a}_{\parallel}(t) + \mathbf{a}_{\perp}(t)$ where we decompose $\mathbf{a}(t)$ using the direction $\mathbf{T}(t)$. Then

$$\mathbf{a}_{\parallel}(t) = v'(t)\mathbf{T}(t) \quad \text{and} \quad \mathbf{a}_{\perp}(t) = v(t)\mathbf{T}'(t) . \quad (2.19)$$

Proof. Since $\mathbf{v}(t) = v(t)\mathbf{T}(t)$, we have from (2.16) that

$$\mathbf{a}(t) = (v(t)\mathbf{T}(t))' = v'(t)\mathbf{T}(t) + v(t)\mathbf{T}'(t) .$$

By Theorem 23, $\mathbf{T}(t)$ and $\mathbf{T}'(t)$ are orthogonal, and so $v'(t)\mathbf{T}(t) + v(t)\mathbf{T}'(t)$ are orthogonal, and clearly the first of these vectors is a multiple of $\mathbf{T}(t)$. This proves (2.19). \square

We refer to \mathbf{a}_{\parallel} as the *tangential component* of the acceleration, and to \mathbf{a}_{\perp} as the *normal component* of the acceleration. We see from (2.19) that the tangential component of the acceleration has to do with the rate of change of the speed, while the normal component has to do with the rate of change of the direction of the velocity vector, $\mathbf{T}(t)$. In particular, when the speed is constant, $\mathbf{a}_{\parallel}(t) = 0$ as in Example 28.

Example 30 (Constant speed circular motion). *Let $\mathbf{x}(t)$ be the curve in \mathbb{R}^3*

$$\mathbf{x}(t) = \mathbf{c} + \varrho[\cos(\omega(t - t_0))\mathbf{u}_1 + \sin(\omega(t - t_0))\mathbf{u}_2] . \quad (2.20)$$

that we considered in Example 28. Recall that $\varrho, \omega > 0$. As we saw there, the speed $v(t)$ has the constant value $v = \varrho\omega$, and so there is no tangential component of the acceleration. By our computations there,

$$\|\mathbf{a}_{\perp}\| = \|\mathbf{a}\| = \varrho\omega^2 = \frac{v^2}{\varrho} .$$

Note that the smaller the radius of the circle, the more “tightly curved” the circle is, and the greater the magnitude of the acceleration at any given speed $v > 0$.

The previous example motivates the following definition.

Definition 26 (Curvature and the unit normal vector). *Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^n , and suppose that the speed $v(t)$ is nonzero on some open interval (a, b) so that $\mathbf{T}(t)$ is defined for all t in this interval. The curvature $\kappa(t)$ at time t is defined by*

$$\kappa(t) = \frac{\|\mathbf{a}_{\perp}\|}{v^2(t)} , \quad (2.21)$$

and the radius of curvature $\varrho(t)$ at time t is defined by $\varrho(t) = \frac{1}{\kappa(t)}$. Furthermore, if $\|\mathbf{a}_{\perp}\| \neq 0$, we define the unit normal vector $\mathbf{N}(t)$ by

$$\mathbf{N}(t) = \frac{1}{\|\mathbf{a}_{\perp}\|}\mathbf{a}_{\perp} , \quad (2.22)$$

Comparing (2.19) and (2.22), we see that $\mathbf{N}(t)$ points in the same direction as $\mathbf{T}'(t)$. Thus, it points in the direction in which the curve is turning. Moreover, since whenever $\|\mathbf{a}_{\perp}\| \neq 0$,

$$\mathbf{a}_{\perp} = \|\mathbf{a}_{\perp}\| \frac{1}{\|\mathbf{a}_{\perp}\|} \mathbf{a}_{\perp} = \|\mathbf{a}_{\perp}\| \mathbf{N} ,$$

it follows from the definition that $\mathbf{a}_{\perp} = v^2 \kappa \mathbf{N}$. Combining this with Theorem 24 yields:

Theorem 25. *Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^n . Then*

$$\mathbf{a}(t) = v'(t)\mathbf{T}(t) + v^2(t)\kappa(t)\mathbf{N}(t) , \quad (2.23)$$

and

$$\mathbf{T}'(t) = v(t)\kappa(t)\mathbf{N}(t) . \quad (2.24)$$

Example 31 (Normal and tangential acceleration). Let $\mathbf{x}(t) = (t, (2t)^{3/2}/3, t^2/2)$ for $t > 0$. We have computed in Example 29 that

$$v(t) = 1 + t \quad \text{and} \quad \mathbf{T}(t) = \frac{1}{1+t}(1, (2t)^{1/2}, t) .$$

Therefore, $v'(t) = 1$, and so $\mathbf{a}_{\parallel}(t) = \mathbf{T}(t)$. Thus, $\mathbf{a}_{\parallel}(t) = \frac{1}{1+t}(1, (2t)^{1/2}, t)$. This is the tangential component of the acceleration.

We next compute $\mathbf{a}(t) = \mathbf{x}''(t) = (0, (2t)^{-1/2}, 1)$. The normal component is

$$(0, (2t)^{-1/2}, 1) - \frac{1}{1+t}(1, (2t)^{1/2}, t) = \frac{1}{1+t}(1, (1-t)(2t)^{-1/2}, 1) .$$

From here we compute $\|\mathbf{a}_{\perp}(t)\| = \frac{1}{\sqrt{2t}}$. Hence

$$\mathbf{N}(t) = \frac{\sqrt{2t}}{1+t}(-1, (1-t)(2t)^{-1/2}, 1)$$

and

$$\kappa(t) = \frac{\sqrt{2t}}{(1+t)^2} \quad \text{and} \quad \varrho(t) = \frac{(1+t)^2}{\sqrt{2t}} .$$

What we have done so far for a twice continuously differentiable curve $\mathbf{x}(t)$ can be summarized as follows: We applied the Gram-Schmidt Algorithm to $\{\mathbf{x}'(t), \mathbf{x}''(t)\}$ to produce the orthonormal set $\{\mathbf{T}(t), \mathbf{N}(t)\}$. Then since

$$\mathbf{T}'(t) = (v(t)^{-1}\mathbf{x}'(t))' = \frac{v'(t)}{v^2(t)}\mathbf{x}'(t) + v(t)\mathbf{x}''(t) ,$$

$\mathbf{T}'(t)$ lies in $\text{Span}(\{\mathbf{x}'(t), \mathbf{x}''(t)\}) = \text{Span}(\{\mathbf{T}(t), \mathbf{N}(t)\})$. Hence

$$\mathbf{T}'(t) = (\mathbf{T}'(t) \cdot \mathbf{T}(t))\mathbf{T}(t) + (\mathbf{T}'(t) \cdot \mathbf{N}(t))\mathbf{N}(t) . \quad (2.25)$$

By Theorem 23, $\mathbf{T}'(t) \cdot \mathbf{T}(t) = 0$. We then defined the curvature $\kappa(t)$ by $v(t)\kappa(t) = \mathbf{T}'(t) \cdot \mathbf{N}(t)$, so that (2.25) reduces to

$$\mathbf{T}'(t) = v(t)\kappa(t)\mathbf{N}(t) . \quad (2.26)$$

In the next subsection, for curves in \mathbb{R}^3 , we complete $\{\mathbf{T}(t), \mathbf{N}(t)\}$ to produce a time-dependent orthonormal basis $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ of \mathbb{R}^3 . One way to do this – which would extend to higher dimensions – would be to apply the Gram-Schmidt algorithm to $\{\mathbf{x}'(t), \mathbf{x}''(t), \mathbf{x}'''(t)\}$. However, in three dimensions, it is traditional to work with right-handed orthonormal bases, so we shall define $\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t)$.

2.1.4 Torsion and the Frenet–Serret formulae for a curve in \mathbb{R}^3

Definition 27 (Binormal vector and osculating plane). Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^3 . Then at each t_0 for which $v(t_0) \neq 0$ and $\kappa(t_0) \neq 0$, so that $\mathbf{T}(t_0)$ and $\mathbf{N}(t_0)$ are well defined, the binormal vector $\mathbf{B}(t_0)$ is defined by

$$\mathbf{B}(t_0) = \mathbf{T}(t_0) \times \mathbf{N}(t_0) , \quad (2.27)$$

and the osculating plane at t_0 is the plane specified by the equation

$$\mathbf{B}(t_0) \cdot (\mathbf{x} - \mathbf{x}(t_0)) = 0 . \quad (2.28)$$

Since $\mathbf{B}(t_0)$ is orthogonal to $\mathbf{T}(t_0)$ and $\mathbf{N}(t_0)$, (2.28) is the equation of the plane through $\mathbf{x}(t_0)$ that contains both $\mathbf{T}(t_0)$ and $\mathbf{N}(t_0)$. Since $\mathbf{v} = v\mathbf{T}$ and $\mathbf{a} = v'\mathbf{T} + v^2\kappa\mathbf{N}$, $\mathbf{v} \times \mathbf{a} = v^3\kappa\mathbf{B}$, which yields the useful formulas

$$\mathbf{B}(t_0) = \frac{1}{v^3(t_0)\kappa(t_0)}\mathbf{v}(t_0) \times \mathbf{a}(t_0) = \frac{1}{\|\mathbf{v}(t_0) \times \mathbf{a}(t_0)\|}\mathbf{v}(t_0) \times \mathbf{a}(t_0) . \quad (2.29)$$

It follows that the direction of \mathbf{B} is the same as that of $\mathbf{v} \times \mathbf{a}$. Therefore, the osculating plane at time t_0 is the plane through $\mathbf{x}(t_0)$ that contains $\mathbf{v}(t_0)$ and $\mathbf{a}(t_0)$. For this reason, the osculating plane is sometimes called the *instantaneous plane of motion*, and another equation for the osculating plane at $t = t_0$ is

$$(\mathbf{v}(t_0) \times \mathbf{a}(t_0)) \cdot (\mathbf{x} - \mathbf{x}(t_0)) = 0 .$$

In particular, it is not necessary to go through all the work of computing \mathbf{T} , \mathbf{N} and then \mathbf{B} if all you wanted to find was an equation for the osculating plane. You can find the equation directly from a computation of \mathbf{v} , \mathbf{a} and $\mathbf{v} \times \mathbf{a}$.

We emphasize that we are assuming throughout these paragraphs, as in Definition 27, that $v(t_0) \neq 0$ and $\kappa(t_0) \neq 0$, so that $\mathbf{T}(t_0)$ and $\mathbf{N}(t_0)$ are well defined. Otherwise, it does not make sense to refer to “the” plane through $\mathbf{x}(t_0)$ containing $\mathbf{v}(t_0)$ and $\mathbf{a}(t_0)$.

Example 32 (An osculating plane). Let $\mathbf{x}(t) = (t, 2^{3/2}t^{3/2}/3, t^2/2)$ for $t > 0$. We have computed that

$$\mathbf{x}(1) = (1, 2^{3/2}/3, 1/2) \quad \mathbf{v}(1) = (1, 2^{1/2}, 1) \quad \text{and} \quad \mathbf{a}(1) = (0, 2^{-1/2}, 1) .$$

We now compute $\mathbf{v}(1) \times \mathbf{a}(1) = (2^{-1/2}, -1, 2^{-1/2})$. The equation for the osculating plane then is

$$(2^{-1/2}, -1, 2^{-1/2}) \cdot (x - 1, y - 2^{3/2}/3, z - 1/2) = 0$$

which reduces to $x - 2^{1/2}y + z = 6$.

Definition 28 (Planar curve in \mathbb{R}^3). A curve $\mathbf{x}(t)$ in \mathbb{R}^3 , $a < t < b$, is planar in case there is some plane in \mathbb{R}^3 that contains $\mathbf{x}(t)$ for all $a < t < b$. In other words, $\mathbf{x}(t)$ is planar in case there exists a non-zero vector \mathbf{n} and a constant d such that $\mathbf{n} \cdot \mathbf{x}(t) = d$ for all $a < t < b$.

Planar curves are easy to recognize when the plane is one of the coordinate planes: For example $\mathbf{x}(t) := (t, t^2, 0)$ is clearly a planar curve – a parabola in the x, y plane. But planar curves are not always so easy to recognize. Consider the curve

$$\mathbf{x}(t) := (-1 + 2t - t^3, t + 3t^2 - 2t^3, 1 + 2t - 6t^2 + 2t^3) . \quad (2.30)$$

As we shall see, this is in fact a planar curve. How can we recognize that, and what is the plane that contains the curve?

Theorem 26 (The binormal vector and planar curves). *Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^3 defined on (a, b) such that $v(t)$ and $\kappa(t)$ are non-zero for all $a < t < b$. Then $\mathbf{x}(t)$ is planar if and only if $\mathbf{B}(t)$ is a constant vector on (a, b) . In this case, there is exactly one plane containing the curve, and for any $t_0 \in (a, b)$, if we define $\mathbf{n} = \mathbf{B}(t_0)$ and $d = \mathbf{x}(t_0) \cdot \mathbf{B}(t_0)$, then $\mathbf{n} \cdot \mathbf{x} = d$ is an equation for the unique plane containing the curve.*

Proof. Suppose that $\mathbf{B}(t)$ is constant on (a, b) . Then for all $a < b < t$, and any $t_0 \in (a, b)$

$$(\mathbf{x}(t) \cdot \mathbf{B}(t_0))' = \mathbf{x}'(t) \cdot \mathbf{B}(t_0) = \mathbf{v}(t) \cdot \mathbf{B}(t) = \frac{1}{\|\mathbf{v}(t) \times \mathbf{a}(t)\|} \mathbf{v}(t) \cdot (\mathbf{v}(t) \times \mathbf{a}(t)) = 0$$

by the triple product identity. Hence for all $t \in (a, b)$

$$\mathbf{x}(t) \cdot \mathbf{B}(t_0) = \mathbf{x}(t_0) \cdot \mathbf{B}(t_0) .$$

This shows that with $\mathbf{n} := \mathbf{B}(t_0)$ and $d := \mathbf{B}(t_0) \cdot \mathbf{x}(t_0)$, the plane specified by $\mathbf{n} \cdot \mathbf{x} = d$ contains $\mathbf{x}(t)$ for all $t \in (a, b)$. There can be no other plane containing the curve since the intersection of two distinct planes in \mathbb{R}^3 is either empty or is a line. Since by hypothesis $\mathbf{x}(t)$ has non-zero curvature, it is not contained in any line.

On the other hand, suppose that $\mathbf{x}(t)$ is planar, and therefore satisfies $\mathbf{n} \cdot \mathbf{x}(t) = d$ for some non-zero \mathbf{n} and some d . Differentiating twice we obtain

$$0 = (\mathbf{n} \cdot \mathbf{x}(t))' = \mathbf{n} \cdot \mathbf{v}(t) \quad \text{and then} \quad 0 = (\mathbf{n} \cdot \mathbf{v}(t))' = \mathbf{n} \cdot \mathbf{a}(t) .$$

Hence for all $t \in (a, b)$, \mathbf{n} is orthogonal to both $\mathbf{v}(t)$ and $\mathbf{a}(t)$. Since by hypothesis the curvature is non-zero, $\mathbf{v}(t)$ and $\mathbf{a}(t)$ are not multiples of one another, and so

$$\mathbf{B}(t) = \frac{1}{\|\mathbf{v}(t) \times \mathbf{a}(t)\|} \mathbf{v}(t) \times \mathbf{a}(t) = \pm \frac{1}{\|\mathbf{n}\|} \mathbf{n} .$$

Since $\mathbf{B}(t)$ is continuous, the sign cannot change anywhere in (a, b) , and so $\mathbf{B}(t)$ is constant whenever the curve is planar. \square

Example 33 (Identifying a planar curve). *Let $\mathbf{x}(t)$ be given by (2.30). Let us compute $\mathbf{B}(t)$ for this curve, and see whether it is constant or not. Before beginning the computation, it will pay to regroup the entries in $\mathbf{x}(t)$. Note that $\mathbf{x}(t) = \mathbf{w}_0 + t\mathbf{w}_1 + t^2\mathbf{w}_2 + t^3\mathbf{w}_3$ where*

$$\mathbf{w}_0 := (-1, 0, 1) , \quad \mathbf{w}_1 := (2, 1, 2) , \quad \mathbf{w}_2 := (0, 3, 6) , \quad \text{and} \quad \mathbf{w}_3 := (-1, -1, 2) .$$

Then we have $\mathbf{v}(t) = \mathbf{w}_1 + 2t\mathbf{w}_2 + 3t^2\mathbf{w}_3$ and $\mathbf{a}(t) = 2\mathbf{w}_2 + 6t\mathbf{w}_3$. Therefore

$$\begin{aligned} \mathbf{v}(t) \times \mathbf{a}(t) &= (\mathbf{w}_1 + 2t\mathbf{w}_2 + 3t^2\mathbf{w}_3) \times (2\mathbf{w}_2 + 6t\mathbf{w}_3) \\ &= 2(\mathbf{w}_1 + 3t^2\mathbf{w}_3) \times \mathbf{w}_2 + 6t(\mathbf{w}_1 + 2t\mathbf{w}_2) \times \mathbf{w}_3 \\ &= 2\mathbf{w}_1 \times \mathbf{w}_2 + 6t\mathbf{w}_1 \times \mathbf{w}_3 + 6t^2\mathbf{w}_2 \times \mathbf{w}_3 . \end{aligned}$$

We then compute

$$\mathbf{w}_1 \times \mathbf{w}_2 = 6(-2, 2, 1) , \quad \mathbf{w}_1 \times \mathbf{w}_3 = -3(-2, 2, 1) , \quad \text{and} \quad \mathbf{w}_2 \times \mathbf{w}_3 = 3(-2, 2, 1) .$$

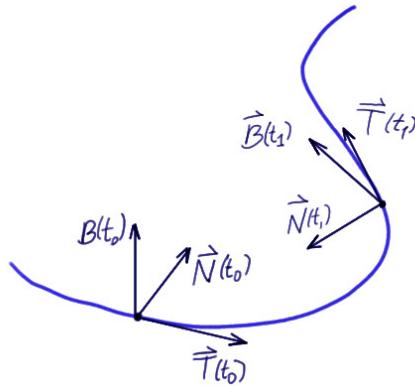
Altogether then $\mathbf{v}(t) \times \mathbf{a}(t) = (12 - 18t + 18t^2)(-2, 2, 1)$ and hence $\mathbf{B}(t) = \frac{1}{3}(-2, 2, 1)$. Since $\mathbf{x}(0) = (1, 0, 1)$, $\mathbf{B}(0) \cdot \mathbf{x}(0) = 1$. Thus, the plane containing the curve satisfies the equation

$$-2x + 2y + z = 1.$$

As we have seen in our examples so far, the rate of change of the basis $\mathbf{T}(t)$ and $\mathbf{B}(t)$ tell us important information about the shape of a curve: The curvature $\kappa(t)$ is related to $\mathbf{T}'(t)$ through $\mathbf{T}'(t) = v(t)\kappa(t)\mathbf{N}(t)$, and the curve is planar if and only if $\mathbf{B}'(t) = 0$ for all t . But this only scratches the surface. There is much more to be learned by considering the rates of change of the vectors in the right handed orthonormal basis

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

that is carried along by any twice differentiable curve in \mathbb{R}^3 with non-zero speed and curvature.



First, let us consider $\mathbf{B}'(t)$.

Lemma 6. Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^3 with non-zero speed and curvature. Then for each t , $\mathbf{B}'(t)$ is a multiple of $\mathbf{N}(t)$.

Proof. $\mathbf{B} = \mathbf{T} \times \mathbf{N}$ and so by Theorem 22

$$\mathbf{B}' = \mathbf{T}' \times \mathbf{N} + \mathbf{T} \times \mathbf{N}' = \mathbf{T} \times \mathbf{N}'$$

since \mathbf{T}' is a multiple of \mathbf{N} . But $\mathbf{T} \times \mathbf{N}'$ is orthogonal to \mathbf{T} , and so \mathbf{B}' is orthogonal to \mathbf{T} . Since \mathbf{B} has constant magnitude, \mathbf{B}' is orthogonal to \mathbf{B} by Theorem 23. Since \mathbf{B}' is orthogonal to both \mathbf{T} and \mathbf{B} , and since $\{\mathbf{T}, \mathbf{N}, \mathbf{B}\}$ is an orthonormal basis, \mathbf{B}' must be a multiple of \mathbf{N} . \square

We now define *torsion*, $\tau(t)$, which quantifies the rate of change of the binormal vector $\mathbf{B}(t)$, and therefore quantifies the extent to which the curve is “twisting out of its osculating plane”:

Definition 29 (Torsion). Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^3 with non-zero speed and curvature for all $t \in (a, b)$. Then the torsion at $t \in (a, b)$ is the quantity $\tau(t)$ defined by

$$\mathbf{B}'(t) = -v(t)\tau(t)\mathbf{N}(t). \quad (2.31)$$

We have already seen that

$$\mathbf{T}'(t) = v(t)\kappa(t)\mathbf{N}(t). \quad (2.32)$$

Notice the similarity between (2.31) and (2.32). The reason for including the minus sign in (2.31) will become evident soon.

Lemma 7. Let $\mathbf{x}(t)$ be a thrice differentiable curve in \mathbb{R}^3 with non-zero speed and curvature for all $t \in (a, b)$. Then for all $t \in (a, b)$,

$$\mathbf{N}'(t) = -v(t)\kappa(t)\mathbf{T}(t) + v(t)\tau(t)\mathbf{B}(t) . \quad (2.33)$$

Proof. Since $\{\mathbf{T}, \mathbf{N}, \mathbf{B}\}$ is a right handed orthonormal basis, $\mathbf{N} = \mathbf{B} \times \mathbf{T}$. Therefore, by Theorem 22

$$\begin{aligned} \mathbf{N}' &= (\mathbf{B} \times \mathbf{T})' = \mathbf{B}' \times \mathbf{T} + \mathbf{B} \times \mathbf{T}' \\ &= -v\tau\mathbf{N} \times \mathbf{T} + \mathbf{B} \times (v\kappa\mathbf{N}) \\ &= v\tau\mathbf{B} - v\kappa\mathbf{T} , \end{aligned}$$

where the last equality again uses the fact that $\{\mathbf{T}, \mathbf{N}, \mathbf{B}\}$ is a right handed orthonormal basis of \mathbb{R}^3 , and Theorem 10. \square

Summarizing the results, we have proved the following:

Theorem 27 (Frenet–Serret formulae). Let $\mathbf{x}(t)$ be a thrice differentiable curve in \mathbb{R}^3 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 and differentiable on this interval. Then for all t in this interval,

$$\begin{aligned} \mathbf{T}'(t) &= v(t)\kappa(t)\mathbf{N}(t) \\ \mathbf{N}'(t) &= -v(t)\kappa(t)\mathbf{T}(t) + v(t)\tau(t)\mathbf{B}(t) \\ \mathbf{B}'(t) &= -v(t)\tau(t)\mathbf{N}(t) . \end{aligned}$$

The Frenet–Serret formulae can be used to show that, up to a sign on \mathbf{B} , $\{\mathbf{T}, \mathbf{N}, \mathbf{B}\}$ is exactly the orthonormal basis one gets by applying the Gram-Schmidt algorithm to $\{\mathbf{x}', \mathbf{x}'', \mathbf{x}'''\}$. Indeed, $\mathbf{x}' = v\mathbf{T}$ by definition, so \mathbf{T} is what one gets in the first step. Then differentiating this,

$$\mathbf{x}'' = v'\mathbf{T} + v\mathbf{T}' = v'\mathbf{T} + v^2\kappa\mathbf{N} . \quad (2.34)$$

Since the coefficient $v^2\kappa$ is always positive, \mathbf{N} is exactly what comes from subtracting off the component of \mathbf{x}'' that is parallel to \mathbf{T} and normalizing. That is, \mathbf{N} is exactly what the second step in the Gram-Schmidt algorithm provides. Next, by the Frenet–Serret formulae and (2.34),

$$\mathbf{x}''' = (v'\mathbf{T} + v^2\kappa\mathbf{N})' = (v'' + v^3\kappa^2)\mathbf{T} + (3vv'\kappa + v^2\kappa')\mathbf{N} + v^3\kappa\tau\mathbf{B} . \quad (2.35)$$

Thus, $v^3\kappa\tau\mathbf{B}$ is what you get by subtracting off the components of \mathbf{x}''' that are parallel to \mathbf{T} and \mathbf{N} . However, τ can be either positive or negative, so if one normalizes this, one gets $\pm\mathbf{B}$. Thus, $\{\mathbf{T}, \mathbf{N}, \mathbf{B}\}$ is the orthonormal frame one gets by applying the Gram-Schmidt algorithm to $\{\mathbf{x}', \mathbf{x}'', \mathbf{x}'''\}$, and then adjusting the sign of the last vector, if need be, to get a right-handed orthonormal basis.

We can extract some useful formulas for curvature and torsion from (2.34) and (2.35): From (2.34), we have $\mathbf{x}' \times \mathbf{x}'' = v^3\kappa\mathbf{B}$, and hence

$$\kappa = \frac{1}{v^3} \|\mathbf{x}' \times \mathbf{x}''\| . \quad (2.36)$$

Next $(\mathbf{x}' \times \mathbf{x}'') \cdot \mathbf{x}''' = v^3 \kappa \mathbf{B} \times \mathbf{x}''$, and then from (2.35)

$$\tau = \frac{1}{v^6 \kappa^2} (\mathbf{x}' \times \mathbf{x}'') \cdot \mathbf{x}''' . \quad (2.37)$$

There is a more useful way to express the three Frenet–Serret formulae.

Definition 30 (Darboux vector). *Let $\mathbf{x}(t)$ 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 $\boldsymbol{\omega}$ is defined on this interval by*

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

The point of the definition is that since $\{\mathbf{T}, \mathbf{N}, \mathbf{B}\}$ is constructed to be a right-handed orthonormal basis of \mathbb{R}^3 , Theorem 10 says that

$$\mathbf{T} \times \mathbf{N} = \mathbf{B} \quad \mathbf{N} \times \mathbf{B} = \mathbf{T} \quad \text{and} \quad \mathbf{B} \times \mathbf{T} = \mathbf{N} ,$$

and thus,

$$\begin{aligned} \boldsymbol{\omega} \times \mathbf{T} &= (\tau\mathbf{T} + \kappa\mathbf{B}) \times \mathbf{T} = \kappa\mathbf{N} \\ \boldsymbol{\omega} \times \mathbf{N} &= (\tau\mathbf{T} + \kappa\mathbf{B}) \times \mathbf{N} = -\kappa\mathbf{T} + \tau\mathbf{B} \\ \boldsymbol{\omega} \times \mathbf{B} &= (\tau\mathbf{T} + \kappa\mathbf{B}) \times \mathbf{B} = -\tau\mathbf{N} . \end{aligned}$$

Comparing with Theorem 2.34, we see that

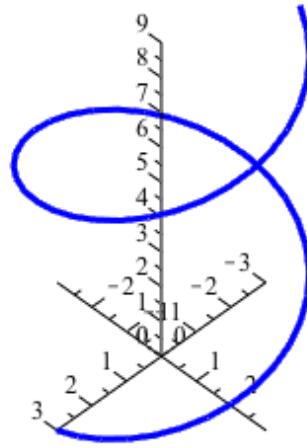
$$\begin{aligned} \mathbf{T}'(t) &= v(t)\boldsymbol{\omega}(t) \times \mathbf{T}(t) \\ \mathbf{N}'(t) &= v(t)\boldsymbol{\omega}(t) \times \mathbf{N}(t) \\ \mathbf{B}'(t) &= v(t)\boldsymbol{\omega}(t) \times \mathbf{B}(t) . \end{aligned} \quad (2.38)$$

As we shall see later in this chapter, this means that for small $h > 0$, the orthonormal basis $\{\mathbf{T}(t+h), \mathbf{B}(t+h), \mathbf{B}(t+h)\}$ 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 $\boldsymbol{\omega}(t)$. That is, the Darboux vector describes the instantaneous rate and direction of rotation of the orthonormal basis $\{\mathbf{T}(t), \mathbf{B}(t), \mathbf{B}(t)\}$.

Example 34 (Curvature and torsion for helices). *Consider the curve $\mathbf{x}(t)$ given by*

$$\mathbf{x}(t) := (r \cos(ct), r \sin(ct), bct)$$

for some $r > 0$ and $b, c \neq 0$. This curve is a helix: There is circular motion in the x, y variables, and linear motion in the z variable. A plot of the curve will look something like:



The plot was made using the values $r = 3$ and $b = 1$ for $0 \leq t \leq 9$.

Let us compute the curvature, torsion, the orthonormal basis $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ and the Darboux vector $\boldsymbol{\omega}(t)$. To begin, we compute

$$\mathbf{v}(t) = c(-r \sin(ct), r \cos(ct), b)$$

from which it follows that

$$v(t) = |c| \sqrt{r^2 + b^2} \quad \text{and} \quad \mathbf{T}(t) = \operatorname{sgn}(c) \frac{r}{\sqrt{r^2 + b^2}} (-\sin(ct), \cos(ct), b/r).$$

We next compute

$$\mathbf{a}(t) = c^2 (-r \cos(ct) - r \sin(ct), 0).$$

Then since $\mathbf{v}(t) \cdot \mathbf{a}(t) = 0$, the parallel component of the acceleration is zero, and so $\mathbf{a}_\perp(t) = \mathbf{a}(t)$. Since $\mathbf{N}(t)$ is the normalization of $\mathbf{a}_\perp(t)$, and hence in this case of $\mathbf{a}(t)$, it follows that

$$\|\mathbf{a}_\perp(t)\| = \|\mathbf{a}(t)\| = c^2 r \quad \text{and} \quad \mathbf{N}(t) = (-\cos(ct), -\sin(ct), 0).$$

The curvature is $\kappa(t) := \frac{\|\mathbf{a}_\perp(t)\|}{v^2(t)} = \frac{r}{r^2 + b^2}$. Let us pause to note that this is reasonable: If $b = 0$, the helix is simply a circle of radius r in the x, y plane, and so as b approaches zero, we must have that the curvature approaches $1/r$. On the other hand, if b is very large, the motion is essentially vertical, and the curvature is very small. This is in agreement with the formula we have found.

We next compute

$$\mathbf{v}(t) \times \mathbf{a}(t) = c^3 r b (\sin(ct), -\cos(ct), r/b).$$

Hence

$$\mathbf{B}(t) = \frac{1}{\|\mathbf{v}(t) \times \mathbf{a}(t)\|} \mathbf{v}(t) \times \mathbf{a}(t) = \operatorname{sgn}(c) \frac{b}{\sqrt{r^2 + b^2}} (\sin(ct), -\cos(ct), r/b).$$

Since $\mathbf{B}(t)$ and $\mathbf{N}(t)$ are so simple in this case, the easiest way to compute the torsion $\tau(t)$ is directly from the defining relation $\mathbf{B}'(t) = -v(t)\tau(t)\mathbf{N}(t)$. We compute

$$\mathbf{B}'(t) = \frac{|c|b}{\sqrt{r^2 + b^2}} (\cos(ct), \sin(ct), 0) = -\frac{|c|b}{\sqrt{r^2 + b^2}} \mathbf{N}(t),$$

Thus

$$-\frac{|c|b}{\sqrt{r^2 + b^2}} = -v(t)\tau(t) \quad \text{so that} \quad \tau(t) = \frac{b}{r^2 + b^2} .$$

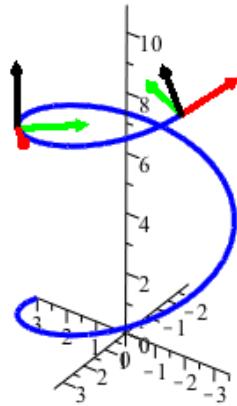
Of course, you could also have computed the curvature and torsion using the formulas (2.36) and (2.37) respectively. However, those formulae do not always provide the simplest approach.

Notice that both the curvature and the torsion turn out to be constant. Let us now compute the Darboux vector $\omega(t)$:

$$\omega(t) = \tau(t)\mathbf{T}(t) + \kappa(t)\mathbf{B}(t) = \frac{rb}{r^2 + b^2}(b/r + r/b)(0, 0, 1) .$$

Notice that this, too, is constant, despite the fact that neither $\mathbf{T}(t)$ nor $\mathbf{B}(t)$ are constant.

Here is a plot, once more for $r = 3$ and $b = 1$ for $0 \leq t \leq 9$, but this time showing $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ for $t = 7$ and $t = 9$:



The final thing to note before leaving this example is that the curvature and torsion are independent of the parameter c . The parameter c determines how fast our parameterization traces out the helix, and the direction of motion – up for c positive, down for c negative. If we change the value of c , we do not change either the curvature or the torsion: They are intrinsic geometric properties of the helix itself, independent of how fast or slow our parameterization may run along it. In the next subsection we shall see this from a more general point of view.

2.1.5 Curvature and torsion are independent of parameterization.

The same path can be parameterized many ways. For example, consider

$$\mathbf{x}(t) = (\cos(t), \sin(t)) \quad \text{and} \quad \mathbf{y}(u) = (\cos(-u^3), \sin(-u^3)) .$$

As t and u vary over \mathbb{R} , both of these curves trace out the unit circle in \mathbb{R}^2 , but they trace it out in different speeds, and one traces it out counterclockwise, and the other clockwise.

Definition 31 (Reparameterization). Let $\mathbf{x}(t)$ be a curve in \mathbb{R}^n defined on an open interval $(a, b) \subset \mathbb{R}$, and let $\mathbf{y}(u)$ be another curve in \mathbb{R}^n defined on an open interval $(c, d) \subset \mathbb{R}$. Either a or c may be $-\infty$, and either b or d may be $+\infty$. Then $\mathbf{y}(u)$ is a reparameterization of $\mathbf{x}(t)$ in case there is a a continuous, strictly monotone increasing or decreasing function $t(u)$ from (c, d) onto (a, b) such that

$$\mathbf{y}(u(t)) = \mathbf{x}(t) \quad \text{for all } t \in (a, b) .$$

Example 35. Define $t(u) = -u^3$ and $u(t) = -t^{1/3}$. Then with $\mathbf{x}(t) = (\cos(t), \sin(t))$ and $\mathbf{y}(u) = (\cos(-u^3), \sin(-u^3))$, we have both

$$\mathbf{y}(u) = \mathbf{x}(t(u)) \quad \text{for all } u \in \mathbb{R}$$

and

$$\mathbf{x}(t) = \mathbf{y}(u(t)) \quad \text{for all } u \in \mathbb{R}.$$

Thus the $\mathbf{x}(t)$ and $\mathbf{y}(u)$ are reparameterizations of each other, and they both parameterize the unit circle.

As in the example, whenever $\mathbf{y}(u)$ is a reparameterization of $\mathbf{x}(t)$, then $\mathbf{x}(t)$ is a reparameterization of $\mathbf{y}(u)$. Indeed, if $t(u)$ is any continuous, strictly monotone increasing function $t(u)$ from (c, d) onto (a, b) , then it is both one-to-one and onto, and so it has an inverse function $u(t)$ from (c, d) to (a, b) which is also continuous and strictly monotone increasing.

- It turns out that while any curve can be parameterized in infinitely many ways, the curvature at a point on the path is a purely geometric property of the path traced out by the curve – it is independent of the parameterization. Not only that, so is the unit normal vector, and, up to a sign, so is the unit tangent vector.

To see this, suppose that $\mathbf{x}(t)$ and $\mathbf{y}(u)$ are two parameterizations of the same path in \mathbb{R}^n . Suppose that

$$\mathbf{x}(t_0) = \mathbf{y}(u_0)$$

so that when $t = t_0$ and $u = u_0$, both curves pass through the same point. Let us suppose also that the two parameterizations are related in a smooth way, so that $t(u)$ is twice continuously differentiable in u .

Then, by the chain rule,

$$\mathbf{y}'(u) = \frac{d}{du} \mathbf{y}(u) = \frac{d}{du} \mathbf{x}(t(u)) = \left(\frac{dt}{du} \right) \mathbf{x}'(t(u)).$$

Evaluating at $u = u_0$, and recalling that $t_0 = t(u_0)$, we get the following relation between the speeds at which the two curve pass through the point in question:

$$\|\mathbf{y}'(u_0)\| = \left| \frac{dt}{du} \right| \|\mathbf{x}'(t_0)\|.$$

Therefore,

$$\begin{aligned} \frac{1}{\|\mathbf{y}'(u_0)\|} \mathbf{y}'(u_0) &= \left(\left| \frac{dt}{du} \right|^{-1} \frac{dt}{du} \right) \frac{1}{\|\mathbf{x}'(t_0)\|} \mathbf{x}'(t_0) \\ &= \pm \frac{1}{\|\mathbf{x}'(t_0)\|} \mathbf{x}'(t_0). \end{aligned}$$

The plus sign is correct if t is an increasing function of u , in which case the two parameterizations trace the path out in the same direction, and otherwise the minus sign is correct.

This shows that up to a sign, the unit tangent vector \mathbf{T} at the point in question comes out the same for the two parameterizations.

Next, let us differentiate once more. We find

$$\begin{aligned}\mathbf{y}''(u) &= \frac{d}{du} \mathbf{y}'(u) = \frac{d}{du} \left(\left(\frac{dt}{du} \right) \mathbf{x}'(t(u)) \right) \\ &= \left(\frac{d^2t}{du^2} \right) \mathbf{x}'(t(u)) + \left(\frac{dt}{du} \right)^2 \mathbf{x}''(t(u)).\end{aligned}$$

Evaluating at $u = u_0$, and recalling that $t_0 = t(u_0)$, we find the following formula relating the acceleration along the two curves as they pass through the point in question:

$$\mathbf{y}''(u_0) = \left(\frac{d^2t}{du^2} \right) \mathbf{x}'(t_0) + \left(\frac{dt}{du} \right)^2 \mathbf{x}''(t_0).$$

Notice that the first term on the right is a multiple of \mathbf{T} , and hence when we decompose $\mathbf{y}''(u_0)$ into its tangential and orthogonal components, this piece contributes only to the tangential component. Hence

$$\mathbf{y}_\perp''(u_0) = \left(\frac{dt}{du} \right)^2 \mathbf{x}_\perp''(t_0).$$

Because of the square, $\mathbf{y}_\perp''(u_0)$ is a positive multiple of $\mathbf{x}_\perp''(t_0)$, and so these two vectors point in the exact same direction. That is,

$$\mathbf{N} = \frac{1}{\|\mathbf{y}_\perp''(u_0)\|} \mathbf{y}_\perp''(u_0) = \frac{1}{\|\mathbf{x}_\perp''(t_0)\|} \mathbf{x}_\perp''(t_0),$$

showing that the normal vector \mathbf{N} is independent of the parameterization.

Next, we consider the curvature. Since

$$\begin{aligned}\frac{1}{\|\mathbf{y}'(u_0)\|^2} \|\mathbf{y}_\perp''(u_0)\| &= \left(\frac{dt}{du} \right)^{-2} \frac{1}{|\mathbf{x}'(t_0)|^2} \left(\frac{dt}{du} \right)^2 \|\mathbf{x}_\perp''(t_0)\| \\ &= \frac{1}{\|\mathbf{x}'(t_0)\|^2} \|\mathbf{x}_\perp''(t_0)\|,\end{aligned}$$

we get the exact same value for the curvature at the same point, using either parameterization. This shows that although in practice we use a particular parameterization to compute the curvature κ and the unit normal \mathbf{N} , the results do not depend on the choice of the parameterization, and are in fact an intrinsically geometric property of the path that the curve traces out.

So far what we have said about reparameterization is valid in \mathbb{R}^n for all $n \geq 2$. In \mathbb{R}^3 , there is more to say. In \mathbb{R}^3 , we also have the binormal vector $\mathbf{B} = \mathbf{T} \times \mathbf{N}$ and the torsion τ .

Since $\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t)$, it follows that $\mathbf{B}(t)$ is well defined, independent of the parameterization, up to a sign: If a reparameterization reverses the direction of travel, then \mathbf{T} but not \mathbf{N} changes sign, and hence \mathbf{B} changes sign. Otherwise, \mathbf{B} does not change. Then, consideration of the formula

$$\mathbf{B}'(t) = -v(t)\tau(t)\mathbf{N}(t)$$

shows that like the curvature, the torsion is completely independent of the parameterization. To see that it does not change sign, fix t_0 , and define the curve $\mathbf{y}(t) = \mathbf{x}(t - t_0)$ and $\tilde{\mathbf{y}}(t) = \mathbf{x}(t_0 - t)$; the second curve is the “time reversal” of the first about $t = t_0$. Let $\mathbf{N}(t)$ and $\mathbf{B}(t)$ be the normal and binormal for $\mathbf{y}(t)$. Likewise, let $\tilde{\mathbf{N}}(t)$ and $\tilde{\mathbf{B}}(t)$ be the normal and binormal for $\tilde{\mathbf{y}}(t)$. By what we have seen above, since $\tilde{\mathbf{y}}(t) = \mathbf{y}(-t)$,

$$\tilde{\mathbf{N}}(t) = \mathbf{N}(-t) \quad \text{and} \quad \tilde{\mathbf{B}}(t) = -\mathbf{B}(-t).$$

Differentiating this last equation, $\tilde{\mathbf{B}}'(0) = \mathbf{B}'(0)$. Therefore, with $v(0)$ and $\tau(0)$ being computed for the $\mathbf{y}(t)$ curve,

$$-\tilde{v}(0)\tilde{\tau}(0)\tilde{\mathbf{N}}(0) = \tilde{\mathbf{B}}'(0) = \mathbf{B}'(0) = -v(0)\tau(0)\mathbf{N}(0) .$$

Since $\tilde{v}(0) = v(0)$ and $\tilde{\mathbf{N}}(0) = \mathbf{N}(0)$, we conclude $\tilde{\tau}(0) = \tau(0)$. This shows that the torsion does not change sign under time reversal. The conclusion is that the torsion, like the curvature, is determined by the geometry of the path itself, and not how fast or slow we move along it, or even the direction of motion.

2.1.6 Speed and arc length

The speed $v(t)$ represents the rate of change of the distance traveled with time. Given some reference time t_0 , define

$$s(t) = \int_{t_0}^t v(u)du . \quad (2.39)$$

Then by the Fundamental Theorem of Calculus,

$$\frac{d}{dt}s(t) = v(t)$$

and clearly $s(t_0) = 0$. Hence the rate of change of $s(t)$ is $v(t)$, which is the rate of change of the distance traveled with time, as one has moved along the path traced out by $\mathbf{x}(t)$.

Definition 32 (Arc length). *The function $s(t)$ defined by (2.39) is called the arc length along the path traced out by $\mathbf{x}(t)$ since time t_0 .*

Example 36 (Computation of arc length). *Let $\mathbf{x}(t)$ be given by $\mathbf{x}(t) = (t, 2^{3/2}t^{3/2}/3, t^2/2)$ for $t > 0$ as in Example 10. Then, as we have seen, for all $t > 0$, $v(t) = 1 + t$. Therefore,*

$$s(t) = \int_0^t (1+u)du = t + \frac{t^2}{2} .$$

If you took a piece of string, and cut it so it can be run along the path from the starting point to the position at time t , the length of the string would be $t + t^2/2$ units of distance.

By definition, $v(t) \geq 0$, and so $s(t)$ has a non-negative derivative. This means that it is an increasing function. As long as $v(t) > 0$; i.e., as long as the particle never comes to even an instantaneous rest, $s(t)$ is strictly monotone increasing. Let us suppose that for some $t_1 > t_0$, $v(t) > 0$ for all $t_0 < t < t_1$. Then $s(t)$ is strictly monotone increasing on the interval $[t_0, t_1]$.

Then for each $s \in [s(t_0), s(t_1)]$, there is exactly one value of $t \in [t_0, t_1]$ so that

$$s(t) = s . \quad (2.40)$$

This value of t , considered as a function of s , is the inverse function to the arc length function:

$$t(s) = t . \quad (2.41)$$

It answers a very simple question, namely: *How much time will have gone by when the distance travelled is s units of length?*

If you can compute an explicit expression for $s(t)$, such as the result $s(t) = t + t^2/2$ that we found in Example 9, what you then need to do to answer the question is to find the inverse function $t(s)$; i.e., to solve (2.40) to find t in terms of s :

Example 37 (Time as a function of arc length). *Let $\mathbf{x}(t)$ be given by*

$\mathbf{x}(t) = (t, 2^{3/2}t^{3/2}/3, t^2/2)$ as in Example 36. Then, as we have seen, for all $t > 0$, $s(t) = t + (t^2/2)$.

To find t as a function of s , write this as

$$s = t + \frac{t^2}{2}$$

and solve for t in terms of s . In this case,

$$t + \frac{t^2}{2} = \frac{1}{2}((t+1)^2 - 1)$$

so $t = \sqrt{2s+1} - 1$. That is,

$$t(s) = \sqrt{2s+1} - 1 .$$

This function tells you how long it took to travel a given distance s when moving along the curve.

We can then get a new parameterization of our curve by defining $\mathbf{x}(s)$ by

$$\mathbf{x}(s) = \mathbf{x}(t(s)) .$$

This is called the *arc length parameterization*. We have changed our habits of notation somewhat: Now we use the same letter \mathbf{x} for both parameterizations to emphasize that they are two parameterizations of the same curve.

Example 38 (Arc length parameterization). *Let $\mathbf{x}(t) = (t, 2^{3/2}t^{3/2}/3, t^2/2)$ as in Example 37. Then, as we have seen, for all $t > 0$, $t(s) = \sqrt{2s+1} - 1$. Therefore,*

$$\mathbf{x}(s) = \mathbf{x}(t(s)) = (\sqrt{2s+1} - 1, 2^{3/2}(\sqrt{2s+1} - 1)^{3/2}/3, (\sqrt{2s+1} - 1)^2/2) .$$

The arc length parameterization generally is complicated to work out explicitly. Even when you can work it out, it often looks a lot more complicated than whatever t parameterization you started with. So what is it good for?

The point about the arc length parameterization is that it is purely geometric, so that it helps us to understand the geometry of the path that a parameterized curve traces out. If we compute the rate of change of the unit tangent vector \mathbf{T} as a function of s , we are computing the rate of turning per unit distance along the curve. This is an intrinsic property of the curve itself. If we compute rate of change of the unit tangent vector \mathbf{T} as a function of t , we are computing something that depends on how fast we are moving on the curve, and not just on the curve itself. Indeed, if we use the arc length parameterization, $v(s) = 1$ for all s , and so the factors involving speed drop out of all of our formulas. For example,

$$\frac{d}{ds}\mathbf{x}(s) = \mathbf{T}(s) \quad \text{and} \quad \frac{d}{ds}\mathbf{T}(s) = \kappa(s)\mathbf{N}(s) .$$

Often, this last formula is taken as the definition of the normal vector \mathbf{N} and curvature κ . The advantage of this definition is that it is manifestly geometric, so that the normal vector \mathbf{N} and

curvature κ do not depend on the parameterization of the curve. The disadvantage is that it is generally very difficult to explicitly work out the arc length parameterization. In order to more quickly arrive at computational examples, we have chosen the form of the definition that is convenient for computation.

2.1.7 Speed, curvature and torsion are independent of the choice of a right-handed coordinate system

Consider a parametrized curve $\mathbf{x}(t)$ in \mathbb{R}^3 , and let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be any right handed orthonormal basis in \mathbb{R}^3 , and let \mathbf{x}_0 be any given vector in \mathbb{R}^3 . Define the functions $y_1(t)$, $y_2(t)$ and $y_3(t)$ by

$$y_j(t) = (\mathbf{x}(t) - \mathbf{x}_0) \cdot \mathbf{u}_j \quad j = 1, 2, 3 .$$

Then these are the coordinate of $\mathbf{x}(t)$ with respect to a coordinate system that has the origin at \mathbf{x}_0 , and such that the directions of the three coordinate axes are given by \mathbf{u}_1 , \mathbf{u}_2 and \mathbf{u}_3 .

Define the curve $\mathbf{y}(t) = (y_1(t), y_2(t), y_3(t))$. This is the same curve as the original curve *only described in a new right handed coordinate system*. We have:

$$\mathbf{x}(t) = \mathbf{x}_0 + \sum_{j=1}^3 y_j(t) \mathbf{u}_j .$$

We now show that speed, curvature and torsion of $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are the same.

We compute

$$\mathbf{x}'(t) = \sum_{j=1}^3 y'_j(t) \mathbf{u}_j \quad \text{and hence} \quad \|\mathbf{x}'(t)\| = \|\mathbf{y}'(t)\| .$$

showing that the speed $v(t)$ is the same for $\mathbf{x}(t)$ and $\mathbf{y}(t)$. Next, by Theorem 11, and the same sort of computations of $\mathbf{x}''(t)$ and $\mathbf{x}'''(t)$,

$$\|\mathbf{x}'(t) \times \mathbf{x}''(t)\| = \|\mathbf{y}'(t) \times \mathbf{y}''(t)\| \quad \text{and} \quad \mathbf{x}'''(t) \cdot \mathbf{x}''(t) \times \mathbf{x}'(t) = \mathbf{y}'''(t) \cdot \mathbf{y}''(t) \times \mathbf{y}'(t) .$$

Since the speed is the same, the first identity together with (2.36) shows that the curvature $\kappa(t)$ is the same for $\mathbf{x}(t)$ and $\mathbf{y}(t)$. Then, since the speed and curvature are the same, the second identity together with (2.37) shows that the torsion $\tau(t)$ is the same for $\mathbf{x}(t)$ and $\mathbf{y}(t)$.

For example, for any given \mathbf{x}_0 in \mathbb{R}^3 and any right handed orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ in \mathbb{R}^3 , and numbers r , b and c with $r > 0$, the curve

$$\mathbf{x}(t) = \mathbf{x}_0 + r \cos(ct) \mathbf{u}_1 + r \sin(ct) \mathbf{u}_2 + b(ct) \mathbf{u}_3 \tag{2.42}$$

has

$$\mathbf{y}(t) = (r \cos(ct), r \sin(ct), bct) \tag{2.43}$$

as its coordinate vector. (Note that in (2.42), $\mathbf{x}(0) = \mathbf{x}_0 + r\mathbf{u}_1$; here \mathbf{x}_0 does not stand for $\mathbf{x}(0)$.)

The right hand side of (2.43) is exactly the helix we studied in Example 34. By what we have just explained, to compute the speed, curvature and torsion of $\mathbf{x}(t)$, we may as well compute using the coordinate vector $\mathbf{y}(t)$ – the results are the same. Hence the computations of speed, curvature

and torsion in Example 34 apply to the more general class of helices with a parameterization of the form (2.42).

We now show that if $\mathbf{x}(t)$ is *any* parameterized curve in \mathbb{R}^3 with constant speed curvature and torsion, there are $\mathbf{x}_0 \in \mathbb{R}^3$, a right handed orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ in \mathbb{R}^3 , and constants c, r and b and so that $\mathbf{x}(t) = \mathbf{x}_0 + r \cos(ct)\mathbf{u}_1 + r \sin(ct)\mathbf{u}_2 + b(ct)\mathbf{u}_3$. That is, up to a change of the time scale, it has a parameterization of the form (2.42), and hence is a helix. Of course for the torsion to be defined, the curve must be thrice differentiable:

Theorem 28 (Curvature, torsion and helices). *A thrice differentiable curve with non-zero speed and curvature is a helix if and only if it has constant curvature and torsion.*

We now prove this. We begin with a lemma on the Darboux vector.

Lemma 8 (Constant curvature and torsion). *Let $\mathbf{x}(t)$ be a thrice differentiable curve with non-zero speed and curvature for $a < t < b$. Then the Darboux vector $\boldsymbol{\omega}(t)$ is constant on the interval (a, b) if and only if the curvature $\kappa(t)$ and the torsion $\tau(t)$ are constant on (a, b) .*

Proof. Suppose first that the curvature κ and torsion τ are constant. Then $\boldsymbol{\omega}(t) = \tau \mathbf{T}(t) + \kappa \mathbf{B}(t)$. Differentiating, and using the Frenet-Serret formulae (2.38),

$$\begin{aligned}\boldsymbol{\omega}'(t) &= \tau \mathbf{T}'(t) + \kappa \mathbf{B}'(t) \\ &= v(t)\tau \boldsymbol{\omega}(t) \times \mathbf{T}(t) + v(t)\kappa \boldsymbol{\omega}(t) \times \mathbf{B}(t) \\ &= v(t)\boldsymbol{\omega}(t) \times (\tau \mathbf{T}(t) + \kappa \mathbf{B}(t)) \\ &= v(t)\boldsymbol{\omega}(t) \times \boldsymbol{\omega}(t) = \mathbf{0}.\end{aligned}$$

Thus, when the curvature and torsion are constant, so is the Darboux vector.

For the converse, suppose that the Darboux vector is constant. Then $\tau(t) = \boldsymbol{\omega} \cdot \mathbf{T}(t)$, and so

$$\tau'(t) = \boldsymbol{\omega} \cdot \mathbf{T}'(t) = v(t)\kappa(t)\boldsymbol{\omega} \cdot \mathbf{N}(t) = 0$$

since the Darboux vector is always orthogonal to \mathbf{N} . A similar calculation shows that $\kappa'(t) = 0$. \square

Proof of Theorem 28. Consider any thrice differentiable curve with constant curvature and torsion. Since this property is independent of parameterization, we may as well suppose that the curve is parameterized by arc length. Therefore, consider any thrice differentiable curve parameterized by arc length, or, what is the same thing, a thrice differentiable curve $\mathbf{x}'(t)$ such that $v(t) = 1$ for all t . Then $\mathbf{x}'(t) = v(t)\mathbf{T}(t) = \mathbf{T}(t)$, and suppose that the curvature and torsion are non-zero constants κ and τ respectively. Then by Lemma 8, the Darboux vector $\boldsymbol{\omega}$ is constant, and is orthogonal to $\mathbf{N}(t)$ for all t . Define a right handed orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ by

$$\mathbf{u}_3 = \frac{1}{\sqrt{\kappa^2 + \tau^2}}\boldsymbol{\omega}, \mathbf{u}_1 = \mathbf{N}(0) \quad \text{and} \quad \mathbf{u}_2 = \mathbf{u}_3 \times \mathbf{u}_1.$$

Computing we find

$$\mathbf{u}_2 = \frac{1}{\sqrt{\kappa^2 + \tau^2}}(\tau \mathbf{T}(0) + \kappa \mathbf{B}(0)) \times \mathbf{N}(0) = \frac{1}{\sqrt{\kappa^2 + \tau^2}}(\tau \mathbf{B}(0) - \kappa \mathbf{T}(0)),$$

and hence Therefore,

$$\mathbf{T}(0) = -\frac{\kappa}{\sqrt{\kappa^2 + \tau^2}} \mathbf{u}_2 + \frac{\tau}{\sqrt{\kappa^2 + \tau^2}} \mathbf{u}_3 . \quad (2.44)$$

We will now use the Frenet-Serret formula to find $\mathbf{T}(t)$ for all t . Then since $\mathbf{x}'(t) = \mathbf{T}(t)$, under our unit speed assumption, we can integrate this obtain the curve itself. We start with $\mathbf{N}(t)$ which is simpler.

Since $\mathbf{N}(t)$ is orthogonal to the Darboux vector, and hence \mathbf{u}_3 , fo all t , we have

$$\mathbf{N}(t) = \cos \theta(t) \mathbf{u}_1 + \sin \theta(t) \mathbf{u}_2 ,$$

for some function $\theta(t)$. All we are using here is the the sum of the squares of the coefficients of \mathbf{u}_1 and \mathbf{u}_2 must be 1 for all t . Differentiating, we find

$$\mathbf{N}'(t) = \theta'(t)(-\sin \theta(t) \mathbf{u}_1 + \cos \theta(t) \mathbf{u}_2) ,$$

By the Frenet-Seret formulae, and the fact that $v(t) = 1$,

$$\mathbf{N}'(t) = \sqrt{\kappa^2 + \tau^2} \mathbf{u}_3 \times \mathbf{N}(t) = \sqrt{\kappa^2 + \tau^2} (\cos \theta(t) \mathbf{u}_2 - \sin \theta(t) \mathbf{u}_1) .$$

Comparing expression for $\mathbf{N}'(t)$ we see that $\theta'(t) = \sqrt{\kappa^2 + \tau^2}$, and then since $\mathbf{N}(0) = \mathbf{u}_1$, $\cos \theta(0) = 1$. Therefore, we may take $\theta(0) = 0$, and then have $\theta(t) = \sqrt{\kappa^2 + \tau^2}t$. Hence, we have an explicit formula for $\mathbf{N}(t)$:

$$\mathbf{N}(t) = \cos(\sqrt{\kappa^2 + \tau^2}t) \mathbf{u}_1 + \sin(\sqrt{\kappa^2 + \tau^2}t) \mathbf{u}_2 .$$

Next, again using the fact that $v(t) = 1$, $\mathbf{T}'(t) = \kappa \mathbf{N}(t)$, and so we have an explicit form formula for $\mathbf{T}'(t)$:

$$\mathbf{T}'(t) = \kappa \cos(\sqrt{\kappa^2 + \tau^2}t) \mathbf{u}_1 + \kappa \sin(\sqrt{\kappa^2 + \tau^2}t) \mathbf{u}_2 .$$

Therefore, by the Fundamental Theorem of Calculus, for all t ,

$$\begin{aligned} \mathbf{T}(t) &= \mathbf{T}(0) + \left[\int_0^t \kappa \cos(\sqrt{\kappa^2 + \tau^2}s) ds \right] \mathbf{u}_1 + \left[\int_0^t \kappa \sin(\sqrt{\kappa^2 + \tau^2}s) ds \right] \mathbf{u}_2 \\ &= \mathbf{T}(0) + \left[\frac{\kappa}{\sqrt{\kappa^2 + \tau^2}} \sin(\sqrt{\kappa^2 + \tau^2}t) \right] \mathbf{u}_1 - \left[\frac{\kappa}{\sqrt{\kappa^2 + \tau^2}} (\cos(\sqrt{\kappa^2 + \tau^2}t) - 1) \right] \mathbf{u}_2 \end{aligned}$$

Combining this with (2.44), we obtain

$$\mathbf{T}(t) = \left[\frac{\kappa}{\sqrt{\kappa^2 + \tau^2}} \sin(\sqrt{\kappa^2 + \tau^2}t) \right] \mathbf{u}_1 - \left[\frac{\kappa}{\sqrt{\kappa^2 + \tau^2}} \cos(\sqrt{\kappa^2 + \tau^2}t) \right] \mathbf{u}_2 + \frac{\tau}{\sqrt{\kappa^2 + \tau^2}} \mathbf{u}_3 .$$

Then, since $\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{T}(s) ds$ integrating one more we obtain that

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(0) + \left[\frac{\kappa}{\kappa^2 + \tau^2} (1 - \cos(\sqrt{\kappa^2 + \tau^2}t)) \right] \mathbf{u}_1 - \left[\frac{\kappa}{\kappa^2 + \tau^2} \sin(\sqrt{\kappa^2 + \tau^2}t) \right] \mathbf{u}_2 + t \frac{\tau}{\sqrt{\kappa^2 + \tau^2}} \mathbf{u}_3 \\ &= \left[\mathbf{x}(0) + \frac{\kappa}{\kappa^2 + \tau^2} \mathbf{u}_1 \right] \\ &\quad - \left[\frac{\kappa}{\kappa^2 + \tau^2} \cos(\sqrt{\kappa^2 + \tau^2}t) \right] \mathbf{u}_1 - \left[\frac{\kappa}{\kappa^2 + \tau^2} \sin(\sqrt{\kappa^2 + \tau^2}t) \right] \mathbf{u}_2 + t \frac{\tau}{\sqrt{\kappa^2 + \tau^2}} \mathbf{u}_3 . \end{aligned}$$

This has the form (2.42) with

$$r = \frac{\kappa}{\kappa^2 + \tau^2} , \quad c = \sqrt{\kappa^2 + \tau^2} , \quad \text{and} \quad b = \frac{\tau}{\kappa^2 + \tau^2} ,$$

and hence it is a helix. \square

2.1.8 Geodesics in \mathbb{R}^n and on the unit sphere

Let \mathbf{u} and \mathbf{w} be two unit vectors, where $\mathbf{w} \neq \pm\mathbf{u}$. The intersection of the unit sphere with the plane passing through \mathbf{u} , \mathbf{v} and $\mathbf{0}$ is a circle. Since the intersection is the solution set of the system of equations

$$\begin{aligned}\|\mathbf{x}\|^2 &= 1 \\ (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{x} &= 0\end{aligned}$$

it is a circle of the sort we have parameterized in Example 28. Such a circle, produced by intersecting a plane through the origin and the unit sphere is called a *great circle* on the unit sphere.

As we shall see, when $\mathbf{w} \neq \pm\mathbf{u}$, the great circle passing through \mathbf{u} and \mathbf{w} consists of two circular arcs that may be parameterized using the method of Example 28. The one that passes from \mathbf{u} to \mathbf{w} without passing through $-\mathbf{u}$ will have the lesser arc length of the two. In fact, this curve will have *less arc length than any other piecewise continuously differentiable curve on the unit sphere that runs from \mathbf{u} to \mathbf{w}* . Such curves that minimize arclength are called *geodesics*.

In mathematical writing, it is usual to write S^2 to denote the unit sphere in \mathbb{R}^3 , which is a “smooth” surface in \mathbb{R}^3 , and as such is “two dimensional” in an obvious sort of way.

Here is the problem to be considered: Given two points in S^2 ; i.e., two unit vectors \mathbf{u} and \mathbf{w} in \mathbb{R}^3 , we seek to find a continuous curve $\mathbf{u}(t)$, defined for $0 \leq t \leq T$, for some $T > 0$ that is piecewise continuously differentiable for $0 < t < T$, and such that:

(i) $\mathbf{u}(0) = \mathbf{u}$ and $\mathbf{u}(T) = \mathbf{w}$.

(ii) $\mathbf{u}(t) \in S^2$ for all $0 < t < T$.

(iii) The arc length of the curve as it runs from \mathbf{u} to \mathbf{w} is less than or equal to the arc length along any other curve of this same kind.

The requirement (ii) says that the curve $\mathbf{u}(t)$ must stay in the sphere S^2 . If we dropped this requirement, it would be valid to consider the curve

$$\mathbf{u}(t) = (1-t)\mathbf{u} + t\mathbf{w}$$

for $T = 1$. This is the straight line segment joining \mathbf{u} and \mathbf{w} , and since $\mathbf{u}'(t) = \mathbf{w} - \mathbf{u}$, the speed along this path is $v(t) = \|\mathbf{u}'(t)\| = \|\mathbf{w} - \mathbf{u}\|$. Thus, the arc length is

$$\int_0^1 v(t) dt = \int_0^1 \|\mathbf{w} - \mathbf{u}\| dt = \|\mathbf{w} - \mathbf{u}\| .$$

As you probably know, this straight line path from \mathbf{u} to \mathbf{w} has the least arc length among *all* piecewise continuously differentiable curves $\tilde{\mathbf{u}}(t)$ with $\tilde{\mathbf{u}}(0) = \mathbf{u}$ and $\tilde{\mathbf{u}}(T) = \mathbf{w}$, i.e., with the condition (ii) dropped:

Theorem 29 (Shortest paths in \mathbb{R}^n). *Let \mathbf{x} and \mathbf{y} be any two distinct points in \mathbb{R}^n . Let $\mathbf{x}(t)$ be any curve in \mathbb{R}^n that is continuous on $[0, T]$ for some $T > 0$, and piecewise continuously differentiable on $(0, T)$ with $\mathbf{x}(0) = \mathbf{x}$ and $\mathbf{x}(T) = \mathbf{y}$. Then the arc length of $\mathbf{x}(t)$ for $0 \leq t \leq T$ is at least $\|\mathbf{y} - \mathbf{x}\|$, and the arc length is exactly $\|\mathbf{y} - \mathbf{x}\|$ if and only if $\mathbf{x}(t)$ traverses the straight line segment from \mathbf{x} to \mathbf{y} without ever reversing the direction of travel.*

Proof. By the Fundamental Theorem of Calculus

$$\mathbf{y} - \mathbf{x} = \int_0^T \mathbf{x}'(t) dt \quad \text{and consequently} \quad \|\mathbf{y} - \mathbf{x}\|^2 = \int_0^T (\mathbf{y} - \mathbf{x}) \cdot \mathbf{x}'(t) dt .$$

By the Cauchy-Schwarz inequality

$$(\mathbf{y} - \mathbf{x}) \cdot \mathbf{x}'(t) \leq \|\mathbf{y} - \mathbf{x}\| \|\mathbf{x}'(t)\| , \quad (2.45)$$

and so

$$\|\mathbf{y} - \mathbf{x}\|^2 \leq \|\mathbf{y} - \mathbf{x}\| \left(\int_0^T \|\mathbf{x}'(t)\| dt \right) .$$

Dividing through by $\|\mathbf{y} - \mathbf{x}\|$, we have

$$\|\mathbf{y} - \mathbf{x}\| \leq \int_0^T \|\mathbf{x}'(t)\| dt , \quad (2.46)$$

and the quantity on the right is the arclength of the curve. There is equality in (2.46) if and only if there is equality in (2.45) for each t , and this means that the angle between $\mathbf{x}'(t)$ and $\mathbf{y} - \mathbf{x}$ is zero for each t . That is, for each t , $\mathbf{x}'(t)$ is a positive multiple of $\mathbf{y} - \mathbf{x}$, which means that $\mathbf{x}(t)$ lies on the straight line segment joining \mathbf{x} and \mathbf{y} , and never reverses direction. \square

Now we return to the sphere S^2 , and let us consider only paths that stay on the sphere. This is a natural constraint: If you are looking for the shortest path from New York to Beijing, the straight line segment is not really relevant: You would have to dig an impressive tunnel to travel along it. So let us try to find a shortest path from \mathbf{u} to \mathbf{w} where \mathbf{u} and \mathbf{w} are on S^2 , and where the path stays at all times on S^2 .

For any fixed, distinct $\mathbf{u}, \mathbf{w} \in S^2$, we define $\mathcal{P}_{\mathbf{u}, \mathbf{w}}$ to be the set of all continuous curves $\mathbf{u}(t)$ staying on S^2 , that are defined on some interval $[0, T]$ for some $T > 0$, and that are piecewise continuously differentiable on $(0, T)$, and such that $\mathbf{u}(0) = \mathbf{u}$ and $\mathbf{u}(T) = \mathbf{w}$.

The arc length function, which assigns the value

$$\int_0^T \|\mathbf{u}'(t)\| dt$$

to $\mathbf{u}(t) \in \mathcal{P}_{\mathbf{u}, \mathbf{w}}$, is a real valued function on $\mathcal{P}_{\mathbf{u}, \mathbf{w}}$. We seek that paths in $\mathcal{P}_{\mathbf{u}, \mathbf{w}}$, if any, that minimize the arc length function on $\mathcal{P}_{\mathbf{u}, \mathbf{w}}$. We shall initially suppose that $\mathbf{w} \neq -\mathbf{u}$, and come back to this special case later.

Theorem 30 (Geodesics on S^2). *Let \mathbf{u} and \mathbf{w} be any two distinct points in S^2 with $\mathbf{w} \neq -\mathbf{u}$. Then the arc length of any path $\mathbf{u}(t) \in \mathcal{P}_{\mathbf{u}, \mathbf{w}}$ is at least as large as*

$$\arccos(\mathbf{u} \cdot \mathbf{w}) ,$$

and the arc length is exactly $\arccos(\mathbf{u} \cdot \mathbf{w})$ if and only $\mathbf{u}(t)$ traverses the arc of the great circle through \mathbf{u} and \mathbf{w} that does not pass through $-\mathbf{u}$, and without ever reversing the direction of travel.

Proof. Decompose \mathbf{w} into its components orthogonal and parallel to \mathbf{u} : $\mathbf{w} = \mathbf{w}_\perp + \mathbf{w}_\parallel$. Since $\mathbf{w} \neq \pm \mathbf{u}$, $\mathbf{w}_\perp \neq \mathbf{0}$, and so we may define a unit vector \mathbf{z} by

$$\mathbf{z} = \frac{1}{\|\mathbf{w}_\perp\|} \mathbf{w}_\perp .$$

Then define an angle ϕ_1 by

$$\phi_1 = \arccos(\mathbf{w} \cdot \mathbf{u}) .$$

Because $\mathbf{w} \neq \pm \mathbf{u}$, $0 < \phi_1 < \pi$, and $\|\mathbf{w}_\parallel\|^2 = \cos^2 \phi_1$, and $\|\mathbf{w}_\perp\|^2 = 1 - \cos^2 \phi_1 = \sin^2 \phi$. Since $0 < \phi_1 < \pi$, $\sin \phi_1 > 0$, and so $\mathbf{w} = \sin \phi_1 \mathbf{z} + \cos \phi_1 \mathbf{u}$. We now define the curve

$$\mathbf{u}(t) := \sin(t\phi_1) \mathbf{z} + \cos(t\phi_1) \mathbf{u} .$$

Evidently, $\mathbf{u}(0) = \mathbf{u}$, and by what we have seen just above, $\mathbf{u}(1) = \mathbf{w}$.

We compute

$$\mathbf{u}'(t) = \phi_1 [-\cos(t\phi_1) \mathbf{z} + \sin(t\phi_1) \mathbf{u}] ,$$

and since \mathbf{u} and \mathbf{z} are orthonormal, $\|\mathbf{u}'(t)\| = \phi_1$. Therefore the arc length of this path is

$$\int_0^1 \|\mathbf{u}'(t)\| dt = \phi_1 = \arccos(\mathbf{w} \cdot \mathbf{u}) .$$

Notice that every point on this path lies on the plane through \mathbf{z} , \mathbf{u} and $\mathbf{0}$, and so it is an arc of a great circle, and is *the* arc of this great circle that does not pass through $-\mathbf{u}$ on the way to \mathbf{w} . Next we shall show that no other path does better.

Let us consider any path in $\mathcal{P}_{\mathbf{u}, \mathbf{w}}$. Without loss of generality, we may assume that $\mathbf{u}(t) \neq \mathbf{u}$ and $\mathbf{u}(t) \neq \mathbf{w}$ for any $t \in (0, T)$, for if $\mathbf{u}(t) = \mathbf{u}$ for any $t > 0$, we may as well start over, and forget about the part of the path traveled so far, which was wasted travel. Likewise, if $\mathbf{u}(t) = \mathbf{w}$ for any $t < T$, we may as well stop the path already.

Next, define an angle $\phi(t)$ by

$$\phi(t) = \arccos(\mathbf{u}(t) \cdot \mathbf{u}) .$$

Since $\phi(0) = 0$ and $\phi(T) = \arccos(\mathbf{w} \cdot \mathbf{u})$, there is a least value of t for which $\phi(t) = \arccos(\mathbf{w} \cdot \mathbf{u})$, and $0 < T_* \leq T$. Since the function $\arccos(s)$ is continuously differentiable on $(0, 1)$ and since $\mathbf{u}(t) \cdot \mathbf{u} \in (0, 1)$ for $t \in (0, T_*)$, by the chain rule, $\phi(t) = \arccos(\mathbf{u}(t) \cdot \mathbf{u})$ is piecewise continuously differentiable on $(0, T_*)$, and $0 < \phi(t) < \pi$ on this interval.

Now decompose $\mathbf{u}(t)$ into its components parallel and orthogonal to \mathbf{u} : $\mathbf{u}(t) = \mathbf{u}_\parallel(t) + \mathbf{u}_\perp(t)$.

We have

$$\mathbf{u}_\parallel(t) = (\mathbf{u}(t) \cdot \mathbf{u}) \mathbf{u} = \cos \phi(t) \mathbf{u} .$$

Since $\|\mathbf{u}_\perp\|^2 = 1 - \|\mathbf{u}_\parallel\|^2 = 1 - \cos^2 \phi(t) = \sin^2 \phi(t)$ and since $0 < \phi(t) < \pi$ for $0 < t < T_*$, $\|\mathbf{u}_\perp(t)\| = \sin \phi(t) > 0$ for all $0 < t < T_*$. Thus we can define a time dependent unit vector $\mathbf{z}(t)$ by

$$\mathbf{z}(t) = \frac{1}{\|\mathbf{u}_\perp(t)\|} \mathbf{u}_\perp(t) .$$

Then $\mathbf{u}_\perp(t) = \sin \phi(t) \mathbf{z}(t)$ and we have already noted that $\mathbf{u}_\parallel(t) = \cos \phi(t) \mathbf{u}$. Therefore,

$$\mathbf{u}(t) = \sin \phi(t) \mathbf{z}(t) + \cos \phi(t) \mathbf{u} .$$

We compute

$$\mathbf{u}'(t) = \phi'(t)[\cos \phi(t)\mathbf{z}(t) - \sin \phi(t)\mathbf{u}] + \sin \phi(t)\mathbf{z}'(t).$$

Since $\|\mathbf{z}(t)\| = 1$ for all $0 < t < T^*$, $\mathbf{z}'(t) \cdot \mathbf{z}(t) = 0$ for all such t . Likewise, since $\mathbf{z}(t) \cdot \mathbf{u} = 0$, and \mathbf{u} is constant, differentiating yields $\mathbf{z}'(t) \cdot \mathbf{u} = 0$ for all t . Thus $\mathbf{z}'(t)$ is orthogonal to both $\mathbf{z}(t)$ and $\mathbf{u}(t)$. Therefore,

$$\begin{aligned}\|\mathbf{u}'(t)\|^2 &= (\phi'(t))^2\|\cos \phi(t)\mathbf{z}(t) - \sin \phi(t)\mathbf{u}\|^2 + \sin^2 \phi(t)\|\mathbf{z}'(t)\|^2 \\ &= (\phi'(t))^2[\cos^2 \phi(t) + \sin^2 \phi(t)] + \sin^2 \phi(t)\|\mathbf{z}'(t)\|^2 \\ &= (\phi'(t))^2 + \sin^2 \phi(t)\|\mathbf{z}'(t)\|^2 \\ &\geq (\phi'(t))^2.\end{aligned}$$

Hence, the arc length along the curve for $0 < t < T_*$ is

$$\int_0^{T_*} \|\mathbf{u}'(t)\| dt \geq \phi(T_*) = \int_0^{T_*} |\phi'(t)| dt \geq \int_0^{T_*} \phi'(t) dt = \arccos(\mathbf{u} \cdot \mathbf{w}),$$

and there is equality if and only if $\phi'(t) \geq 0$ for all t $\mathbf{z}'(t) = 0$ for all t , meaning that $\mathbf{z}(t)$ is a constant unit vector \mathbf{z} orthogonal to \mathbf{u} . In this case,

$$\mathbf{u}(t) = \sin \phi(t)\mathbf{z} + \cos \phi(t)\mathbf{u}$$

for $0 < t < T_*$ with $\phi(t)$ monotone increasing from 0 to ϕ_1 . Notice that for each such t , $\mathbf{u}(t)$ lies in the plane through \mathbf{z} , \mathbf{u} and $\mathbf{0}$, and so is on the great circle through which this plane slices the sphere.

Next, the arc length traversed between $0 < t < T^*$ is less than the arc length traversed between $0 < t < T$ unless $T_* = T$, so that if the arc length of our path is ϕ_1 , then $T_* = T$ and

$$\mathbf{u}(T_*) = \mathbf{w} = \sin \phi(T_*)\mathbf{z} + \cos \phi(T_*)\mathbf{u},$$

Then the plane through \mathbf{z} , \mathbf{u} and $\mathbf{0}$ is also the plane through \mathbf{w} , \mathbf{u} and $\mathbf{0}$. Thus, for the arc length of the path to equal ϕ_1 , it must traverse the arc of the great circle \mathbf{z} , \mathbf{u} and $\mathbf{0}$ that does not pass through $-\mathbf{u}$, and the angle between $\mathbf{u}(t)$ and \mathbf{u} must be monotone increasing. \square

There was nothing particularly three dimensional about the proof of Theorem 30. Indeed, it can be extended to arbitrary dimensions. Define S^n to be the set of all unit vectors in \mathbb{R}^{n+1} . The geometry of these higher dimensional spheres turns out to be important in many questions concerning physics and engineering. Indeed, the three dimensional sphere S^3 in four dimensional space \mathbb{R}^4 has a direct connection with rotations in the three dimensional space \mathbb{R}^3 that is important in many applications.

Finally, we come to the case $\mathbf{w} = -\mathbf{u}$. To reach $-\mathbf{u}$ starting from \mathbf{u} , one must first arrive at some point $\tilde{\mathbf{w}}$ that is very close, but not equal to $-\mathbf{u}$. By what we have seen above, the length of this part of the path is at least $\arccos(\tilde{\mathbf{w}} \cdot \mathbf{u})$, hence the length of the whole path to \mathbf{u} is at least this large. Taking $\tilde{\mathbf{w}}$ closer and closer to $-\mathbf{u}$, we see that the arclength is at least ϕ for any $\phi < \pi$, and hence it is at least π . There are infinitely many planes through $-\mathbf{u}$, \mathbf{u} and $\mathbf{0}$, which are collinear, and so there are infinitely many great circles connecting \mathbf{u} to $-\mathbf{u}$. The arc length along any of them is π .

Definition 33 (The geodesic distance function on S^2). Define the function d_{S^2} on the Cartesian product $S^2 \times S^2$ by

$$d_{S^2}(\mathbf{u}, \mathbf{w}) = \arccos(\mathbf{u} \cdot \mathbf{w})$$

for \mathbf{u}, \mathbf{w} in S^2 . This is the geodesic distance function on S^2

The function $d_{S^2}(\mathbf{u}, \mathbf{w})$ is a metric on S^2 . That is,

- (1) For all $\mathbf{u}, \mathbf{w} \in S^2$, $d_{S^2}(\mathbf{u}, \mathbf{w}) \geq 0$ and $d_{S^2}(\mathbf{u}, \mathbf{w}) = 0$ if and only if $\mathbf{u} = \mathbf{w}$.
- (2) For all $\mathbf{u}, \mathbf{w} \in S^2$, $d_{S^2}(\mathbf{u}, \mathbf{w}) = d_{S^2}(\mathbf{w}, \mathbf{u})$.
- (3) For all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in S^2$, $d_{S^2}(\mathbf{u}, \mathbf{w}) \leq d_{S^2}(\mathbf{u}, \mathbf{v}) + d_{S^2}(\mathbf{v}, \mathbf{w})$.

Property (1) follows from the fact that $\mathbf{u} \cdot \mathbf{w} < 1$ for $\mathbf{u} \neq \mathbf{w}$ and with $\mathbf{u}, \mathbf{w} \in S^2$. Likewise, (2) is a consequence of the fact that $\mathbf{u} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{u}$.

The inequality (3) is the *triangle inequality* on S^2 . Here is one way to see this using Theorem 30.

Build path from \mathbf{u} to \mathbf{w} as follows: Let $\mathbf{u}_1(t)$, $t \in [0, 1]$, be a shortest path from \mathbf{u} to \mathbf{v} . Let $\mathbf{u}_2(t)$, $t \in [0, 1]$, be a shortest path from \mathbf{v} to \mathbf{w} . Define a path $\mathbf{u}(t)$, $t \in [0, 2]$, from \mathbf{u} to \mathbf{w} by

$$\mathbf{u}(t) = \begin{cases} \mathbf{u}_1(t) & 0 \leq t \leq 1 \\ \mathbf{u}_2(t-1) & 1 \leq t \leq 2 \end{cases}$$

This path is continuous and piecewise continuously differentiable. Therefore, by Theorem 30, $d_{S^2}(\mathbf{u}, \mathbf{w}) = \arccos(\mathbf{u} \cdot \mathbf{w})$ is less than or equal to the length of this composite path. But by construction, the length of the composite paths is the sum of the two lengths, namely $d_{S^2}(\mathbf{u}, \mathbf{v}) + d_{S^2}(\mathbf{v}, \mathbf{w})$.

The proof we have given of the triangle inequality for the geodesic distance function on S^2 uses the rather sophisticated Theorem 30. But the triangle inequality simply says that for any three unit vectors \mathbf{u} , \mathbf{v} and \mathbf{w} in \mathbb{R}^3 ,

$$\arccos(\mathbf{u} \cdot \mathbf{w}) \leq \arccos(\mathbf{u} \cdot \mathbf{v}) + \arccos(\mathbf{v} \cdot \mathbf{w}) . \quad (2.47)$$

In fact, it is possible to prove this directly, without considering paths.

Given three such unit vectors \mathbf{u} , \mathbf{v} and \mathbf{w} , define

$$\theta = \arccos(\mathbf{u} \cdot \mathbf{v}) \quad \text{and} \quad \phi = \arccos(\mathbf{v} \cdot \mathbf{w}) .$$

Write $\mathbf{u} = \mathbf{u}_{\parallel} + \mathbf{u}_{\perp}$ and $\mathbf{w} = \mathbf{w}_{\parallel} + \mathbf{w}_{\perp}$ where parallel and perpendicular components are taken with respect to \mathbf{v} . Then $\mathbf{u} \cdot \mathbf{w} = \mathbf{u}_{\parallel} \cdot \mathbf{w}_{\parallel} + \mathbf{u}_{\perp} \cdot \mathbf{w}_{\perp}$, and

$$\mathbf{u}_{\parallel} \cdot \mathbf{w}_{\parallel} = \cos \theta \cos \phi \quad \text{and} \quad \mathbf{u}_{\perp} \cdot \mathbf{w}_{\perp} \geq -\|\mathbf{u}_{\perp}\| \|\mathbf{w}_{\perp}\| = -\sin \theta \sin \phi ,$$

where we used the Cauchy-Schwarz inequality. By the angle addition formula,

$$\cos(\theta + \phi) = \cos \theta \cos \phi - \sin \theta \sin \phi \leq \mathbf{u} \cdot \mathbf{w} .$$

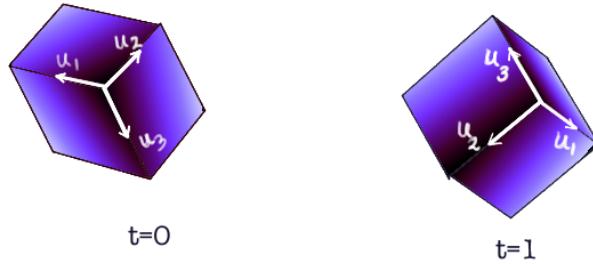
Since the cosine function is monotone decreasing on $[0, \pi]$, $\arccos(\mathbf{u} \cdot \mathbf{w}) \leq \theta + \phi$, and this proves the inequality (2.47).

Again, in this proof, we did not use any cross products or anything specific to \mathbb{R}^3 . Thus, this proof shows that (2.47) is valid for any here unit vectors in \mathbb{R}^n , for any n , and thus we can define a metric; i.e., a distance function, on the n dimensional sphere in \mathbb{R}^{n+1} , which is the set of all unit vectors in \mathbb{R}^{n+1} , by

$$d_{S^n}(\mathbf{u}, \mathbf{w}) = \arccos(\mathbf{u} \cdot \mathbf{w}) .$$

2.1.9 Rotations, continuity and the right hand rule

We now apply some of what we have learned recently to the study of rigid body motion. Imagine a rigid cubical box moving in three dimensional space. Here is a picture showing the box shaped object at two times: $t = 0$ and $t = 1$:



As it moves, the box carries with it a “reference frame” of three unit vectors \mathbf{u}_1 , \mathbf{u}_2 and \mathbf{u}_3 . Since physical motions are continuous, rigid body motion involves a continuous time dependent orthonormal frame $\{\mathbf{u}_1(t), \mathbf{u}_2(t), \mathbf{u}_3(t)\}$

The orthonormal basis $\{\mathbf{u}_1(t), \mathbf{u}_2(t), \mathbf{u}_3(t)\}$ is right handed in case $\mathbf{u}_1(t) \times \mathbf{u}_2(t) \cdot \mathbf{u}_3(t) = 1$, and is left handed in case $\mathbf{u}_1(t) \times \mathbf{u}_2(t) \cdot \mathbf{u}_3(t) = -1$, and ± 1 are the only possible values for this triple product.

Now, if $\mathbf{u}_j(t)$ is continuous for each $j = 1, 2, 3$, then $\mathbf{u}_1(t) \times \mathbf{u}_2(t) \cdot \mathbf{u}_3(t)$ is a continuous function of t . Since it only has two possible values, and cannot jump from one to the other, it must be constant. That is, under our continuity assumption,

- Let $\{\mathbf{u}_1(t), \mathbf{u}_2(t), \mathbf{u}_3(t)\}$ be a continuously time dependent orthonormal basis of \mathbb{R}^3 . Then if $\{\mathbf{u}_1(0), \mathbf{u}_2(0), \mathbf{u}_3(0)\}$ is right-handed, so is $\{\mathbf{u}_1(t), \mathbf{u}_2(t), \mathbf{u}_3(t)\}$ for every t .

In particular, it is impossible to “continuously interpolate” between a right-handed orthonormal basis and a left-handed orthonormal basis: If $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right-handed orthonormal basis and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a left-handed orthonormal basis, *there does not exist any* continuously time dependent orthonormal basis $\{\mathbf{u}_1(t), \mathbf{u}_2(t), \mathbf{u}_3(t)\}$, $0 \leq t \leq 1$ with

$$\mathbf{u}_j(0) = \mathbf{u}_j \quad \text{and} \quad \mathbf{u}_j(1) = \mathbf{v}_j$$

for $j = 1, 2, 3$.

However, as we shall now show, if $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ are both right-handed (or both left-handed), then there is a continuous interpolation between them, and one such interpolation is

through a “rotation about a fixed axis at constant angular velocity”. The following lemma concerning Householder reflections is the key to see why this is so.

Lemma 9 (Householder reflections and the cross product). *Let \mathbf{u} be any unit vector in \mathbb{R}^3 and let \mathbf{a} and \mathbf{b} be any two vectors in \mathbb{R}^3 . Then*

$$\mathbf{h}_{\mathbf{u}}(\mathbf{a} \times \mathbf{b}) = -\mathbf{h}_{\mathbf{u}}(\mathbf{a}) \times \mathbf{h}_{\mathbf{u}}(\mathbf{b}) ,$$

Proof. Direct computation shows that

$$\begin{aligned} \mathbf{h}_{\mathbf{u}}(\mathbf{a} \times \mathbf{b}) + \mathbf{h}_{\mathbf{u}}(\mathbf{a}) \times \mathbf{h}_{\mathbf{u}}(\mathbf{b}) &= 2[\mathbf{a} \times \mathbf{b} - (\mathbf{a} \times \mathbf{b} \cdot \mathbf{u})\mathbf{u} - (\mathbf{a} \cdot \mathbf{u})\mathbf{u} \times \mathbf{b} - (\mathbf{b} \cdot \mathbf{u})\mathbf{a} \times \mathbf{u}] \\ &= 2[(\mathbf{a} \times \mathbf{b})_{\perp} - (\mathbf{a} \cdot \mathbf{u})\mathbf{u} \times \mathbf{b} + (\mathbf{b} \cdot \mathbf{u})\mathbf{u} \times \mathbf{a}] \end{aligned} \quad (2.48)$$

where $(\mathbf{a} \times \mathbf{b})_{\perp}$ is the component of $\mathbf{a} \times \mathbf{b}$ orthogonal to \mathbf{u} .

However, since \mathbf{u} is a unit vector, Lagrange’s identity gives us:

$$\begin{aligned} (\mathbf{a} \times \mathbf{b})_{\perp} &= -\mathbf{u} \times [\mathbf{u} \times (\mathbf{a} \times \mathbf{b})] \\ &= -\mathbf{u} \times [(\mathbf{u} \cdot \mathbf{b})\mathbf{a} - (\mathbf{u} \cdot \mathbf{a})\mathbf{b}] \\ &= -(\mathbf{u} \cdot \mathbf{b})\mathbf{u} \times \mathbf{a} + (\mathbf{u} \cdot \mathbf{a})\mathbf{u} \times \mathbf{b} \end{aligned}$$

Using this in (2.48), one obtains $\mathbf{h}_{\mathbf{u}}(\mathbf{a} \times \mathbf{b}) + \mathbf{h}_{\mathbf{u}}(\mathbf{a}) \times \mathbf{h}_{\mathbf{u}}(\mathbf{b}) = \mathbf{0}$. □

Since Householder reflections preserve dot products, and hence lengths and angles, we know that whenever $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right-handed orthonormal basis, then

$$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} := \{\mathbf{h}_{\mathbf{u}}(\mathbf{u}_1), \mathbf{h}_{\mathbf{u}}(\mathbf{u}_2), \mathbf{h}_{\mathbf{u}}(\mathbf{u}_3)\} \quad (2.49)$$

an orthonormal basis. By the lemma,

$$\mathbf{v}_1 \times \mathbf{v}_2 = \mathbf{h}_{\mathbf{u}}(\mathbf{u}_1) \times \mathbf{h}_{\mathbf{u}}(\mathbf{u}_2) = -\mathbf{h}_{\mathbf{u}}(\mathbf{u}_1 \times \mathbf{u}_3) = -\mathbf{h}_{\mathbf{u}}(\mathbf{u}_3) = -\mathbf{v}_3 ,$$

so that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is left-handed. Likewise if $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is left-handed, (2.49) defines a right-handed orthonormal basis.

Now we are ready to draw some important conclusions.

Theorem 31 (Right handed orthonormal bases and reflection). *Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ be two distinct right-handed orthonormal bases. Then there are unit vectors \mathbf{u} and \mathbf{v} such that*

$$\mathbf{h}_{\mathbf{v}}(\mathbf{h}_{\mathbf{u}}(\mathbf{u}_j)) = \mathbf{v}_j \quad \text{for } j = 1, 2, 3 . \quad (2.50)$$

Proof. Since the bases are distinct, we must have $\mathbf{u}_j \neq \mathbf{v}_j$ for some j . By cyclicly permuting the indices, we may suppose that $\mathbf{u}_1 \neq \mathbf{v}_1$.

Let $\mathbf{u} = \|\mathbf{u}_1 - \mathbf{v}_1\|^{-1}(\mathbf{u}_1 - \mathbf{v}_1)$. Then $\mathbf{h}_{\mathbf{u}}(\mathbf{u}_1) = \mathbf{v}_1$, and we then define $\mathbf{w}_j := \mathbf{h}_{\mathbf{u}}(\mathbf{u}_j)$ for $j = 1, 2$, so that we have the left handed orthonormal basis

$$\{\mathbf{v}_1, \mathbf{w}_2, \mathbf{w}_3\} = \{\mathbf{h}_{\mathbf{u}}(\mathbf{u}_1), \mathbf{h}_{\mathbf{u}}(\mathbf{u}_2), \mathbf{h}_{\mathbf{u}}(\mathbf{u}_3)\} .$$

Now suppose that $\mathbf{w}_2 = \mathbf{v}_2$. Then we must have $\mathbf{w}_3 = -\mathbf{v}_3$. In this case, we take $\mathbf{v} := \mathbf{w}_3$. Then since this vector is orthogonal to both \mathbf{v}_1 and $\mathbf{w}_2 = \mathbf{v}_2$, $\mathbf{h}_v(\mathbf{v}_1) = \mathbf{v}_1$ and $\mathbf{h}_v(\mathbf{w}_2) = \mathbf{w}_2 = \mathbf{v}_2$. Finally, $\mathbf{h}_v(\mathbf{w}_3) = -\mathbf{w}_3 = \mathbf{v}_3$. That is,

$$\{\mathbf{h}_v(\mathbf{v}_1), \mathbf{h}_b(\mathbf{w}_2), \mathbf{h}_v(\mathbf{w}_3)\} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} .$$

Thus, in this case, successively applying \mathbf{h}_u and then \mathbf{h}_v transforms $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ into $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$

On the other hand, if $\mathbf{w}_2 \neq \mathbf{v}_2$, we define $\mathbf{v} = \|\mathbf{v}_2 - \mathbf{w}_2\|^{-1}(\mathbf{v}_2 - \mathbf{w}_2)$, so that $\mathbf{h}_v(\mathbf{w}_2) = \mathbf{v}_2$. Note that \mathbf{w}_2 and \mathbf{v}_2 are both orthogonal to \mathbf{v}_1 since $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $\{\mathbf{v}_1, \mathbf{w}_2, \mathbf{w}_3\}$ are both orthonormal bases. But then \mathbf{v} is orthogonal to \mathbf{v}_1 , and so $\mathbf{h}_v(\mathbf{v}_1) = \mathbf{v}_1$.

Now since $\{\mathbf{v}_1, \mathbf{w}_2, \mathbf{w}_3\}$ is a left handed orthonormal basis,

$$\{\mathbf{h}_v(\mathbf{v}_1), \mathbf{h}_b(\mathbf{w}_2), \mathbf{h}_v(\mathbf{w}_3)\} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{h}_v(\mathbf{w}_3)\}$$

is a right handed orthonormal basis. Since any two vectors determine the third, and since $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is right handed, it must be that $\mathbf{h}_v(\mathbf{w}_2) = \mathbf{v}_3$. Either way, we have proved (2.50). \square

In what follows, let us fix two distinct right-handed orthonormal bases $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, and let us define \mathbf{f} by

$$\mathbf{f}(\mathbf{x}) = \mathbf{h}_v(\mathbf{h}_u(\mathbf{x}))$$

where \mathbf{h}_v and \mathbf{h}_u are the Householder reflections provided by Theorem 31 so that \mathbf{f} transforms $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ into $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$.

If it were the case that $\mathbf{v} = \pm \mathbf{u}$, then we would have $\mathbf{h}_v = \mathbf{h}_u$, and \mathbf{f} would be the identity transformation. Since the two orthonormal bases are distinct, $\mathbf{v} \neq \pm \mathbf{u}$, Decompose \mathbf{v} into its components parallel and orthogonal to \mathbf{u} . Since $\mathbf{v} \neq \pm \mathbf{u}$, $\mathbf{v}_\perp \neq \mathbf{0}$. Define $\mathbf{z} = \|\mathbf{v}_\perp\|^{-1}\mathbf{v}_\perp$, which is a unit vector orthogonal to \mathbf{u} . Then

$$\mathbf{v} = \mathbf{v}_\parallel + \mathbf{v}_\perp = (\mathbf{v} \cdot \mathbf{u})\mathbf{u} + \|\mathbf{v}_\perp\|\mathbf{z} .$$

Define $\Theta \in [0, \pi]$ by

$$\Theta := \arccos(\mathbf{v} \cdot \mathbf{u}) . \quad (2.51)$$

and then since $\|\mathbf{v}_\perp\|^2 = 1 - \|\mathbf{v}_\parallel\|^2 = 1 - \cos^2(\Theta)$, we have $\mathbf{v} = \cos \Theta \mathbf{u} + \sin \Theta \mathbf{z}$. We now define a curve $\mathbf{u}(t)$ by

$$\mathbf{u}(t) := \cos(t\Theta)\mathbf{u} + \sin(t\Theta)\mathbf{z} , \quad (2.52)$$

By construction $\mathbf{u}(t)$ is a unit vector for each t , $\mathbf{u}(0) = \mathbf{u}$ and $\mathbf{u}(1) = \mathbf{v}$.

Given this interpolation between \mathbf{u} and \mathbf{v} , define the t dependent orthogonal transformation \mathbf{f}_t by

$$\mathbf{f}_t(\mathbf{x}) := \mathbf{h}_{\mathbf{u}(t)}(\mathbf{h}_u(\mathbf{x})) . \quad (2.53)$$

Since $\mathbf{u}(0) = \mathbf{u}$, and since $\mathbf{h}_u \circ \mathbf{h}_u$ is the identity, \mathbf{f}_0 is the identity transformation, and by construction \mathbf{f}_1 transforms $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ into $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. Consequently, if we define

$$\mathbf{u}_j(t) = \mathbf{f}_t(\mathbf{u}_j) \quad j = 1, 2, 3 \quad \text{and} \quad 0 \leq t \leq 1 ,$$

$\{\mathbf{u}_1(t), \mathbf{u}_2(t), \mathbf{u}_3(t)\}$ interpolates between $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$.

We now claim that the transformation of \mathbb{R}^3 sending the vector \mathbf{w} to the vector $\mathbf{f}_t(\mathbf{w})$ is a *right handed rotation about the line with direction vector $\mathbf{u} \times \mathbf{v}$ through an angle 2Θ* .

To see this, consider the orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ where $\mathbf{u}_1 = \mathbf{u}$, $\mathbf{u}_2 = \mathbf{z}$ and $\mathbf{u}_3 = \|\mathbf{u} \times \mathbf{v}\|^{-1}\mathbf{u} \times \mathbf{v}$. This is right handed since $\mathbf{u} \times \mathbf{z} = \|\mathbf{u} \times \mathbf{v}\|^{-1}\mathbf{u} \times \mathbf{v}$. It is easy to compute the action of \mathbf{f}_t in this basis: For any coordinates $\tilde{x}_0, \tilde{y}_0, \tilde{z}_0$,

$$\mathbf{f}_t(\tilde{x}_0\mathbf{u}_1 + \tilde{y}_0\mathbf{u}_2 + \tilde{z}_0\mathbf{u}_3) = (\cos(2t\Theta)x_0 + \sin(2t\Theta)y_0)\mathbf{u}_1 + (-\sin(2t\Theta)x_0 + \cos(2t\Theta)y_0)\mathbf{u}_2 + \tilde{z}_0\mathbf{u}_3 .$$

The coordinate vector of $\mathbf{f}_t(\tilde{x}_0\mathbf{u}_1 + \tilde{y}_0\mathbf{u}_2 + \tilde{z}_0\mathbf{u}_3)$ is therefore given by

$$(\cos(2t\Theta)x_0 + \sin(2t\Theta)y_0, -\sin(2t\Theta)x_0 + \cos(2t\Theta)y_0, \tilde{z}_0) .$$

That is, the coordinate vector rotates at a constant angular velocity around the \mathbf{u}_3 axis. Looking down on the \tilde{x}, \tilde{y} -plane from above, the rotation in the plane is counterclockwise when $\Theta > 0$; i.e., when the angle between \mathbf{u} and \mathbf{v} is acute. We summarize:

Theorem 32 (Rotations and reflections). *Given two unit vectors \mathbf{u} and \mathbf{v} in \mathbb{R}^3 with $\mathbf{v} \neq \pm\mathbf{u}$, define*

$$\theta := \frac{1}{2}\arccos(\mathbf{v} \cdot \mathbf{u}) \quad \text{and} \quad \mathbf{a} := \frac{1}{\|\mathbf{u} \times \mathbf{v}\|}\mathbf{u} \times \mathbf{v} .$$

Then the transformation \mathbf{f} defined by $\mathbf{f}(\mathbf{x}) := \mathbf{h}_{\mathbf{v}}(\mathbf{h}_{\mathbf{u}}(\mathbf{x}))$ is rotation by an angle θ about the axis along \mathbf{a} . Thus, every rotation in \mathbb{R}^3 can be written as the composition product of two Householder reflections.

By differentiating $\mathbf{f}_t(\mathbf{x}_0)$, we will gain new insight into the Frenet-Serret equations and the meaning of the Darboux vector.

Fix a vector $\mathbf{x}_0 \in \mathbb{R}^3$, and define a curve $\mathbf{x}(t)$ by $\mathbf{x}(t) = \mathbf{f}_t(\mathbf{x}_0)$. Defining $\mathbf{y}_0 = \mathbf{h}_{\mathbf{u}}(\mathbf{x}_0)$, we have the equivalent formula

$$\mathbf{x}(t) = \mathbf{h}_{\mathbf{u}(t)}(\mathbf{y}_0) = \mathbf{x}_0 - 2(\mathbf{y}_0 \cdot \mathbf{u}(t))\mathbf{u}(t) .$$

Differentiating, and using the fact that reflections are their own inverses so that $\mathbf{x}_0 = \mathbf{h}_{\mathbf{u}(t)}(\mathbf{x}(t))$

$$\begin{aligned} \mathbf{x}'(t) &= -2[\mathbf{y}_0 \cdot \mathbf{u}'(t)]\mathbf{u}(t) - 2[\mathbf{y}_0 \cdot \mathbf{u}(t)]\mathbf{u}'(t) \\ &= -2[\mathbf{h}_{\mathbf{u}(t)}(\mathbf{x}(t)) \cdot \mathbf{u}'(t)]\mathbf{u}(t) - 2[\mathbf{h}_{\mathbf{u}(t)}(\mathbf{x}(t)) \cdot \mathbf{u}(t)]\mathbf{u}'(t) \end{aligned}$$

Now note that $\mathbf{h}_{\mathbf{u}(t)}(\mathbf{u}(t)) = -\mathbf{u}(t)$, but since $\mathbf{u}'(t)$ is orthogonal to $\mathbf{u}(t)$, $\mathbf{h}_{\mathbf{u}(t)}(\mathbf{u}'(t)) = \mathbf{u}'(t)$. Hence, by the identity $\mathbf{h}_{\mathbf{u}(t)}(\mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{h}_{\mathbf{u}(t)}(\mathbf{b})$, which is valid for any \mathbf{a} and \mathbf{b} ,

$$\mathbf{x}'(t) = -2[\mathbf{x}(t) \cdot \mathbf{u}'(t)]\mathbf{u}(t) + 2[\mathbf{x}(t) \cdot \mathbf{u}(t)]\mathbf{u}'(t) = 2\mathbf{x}(t) \times (\mathbf{u}'(t) \times \mathbf{u}(t)) = 2(\mathbf{u}(t) \times \mathbf{u}'(t)) \times \mathbf{x}(t) ,$$

where we have used Lagrange's identity. Computing $\mathbf{u}(t) \times \mathbf{u}'(t)$, we find

$$(\cos(t\Theta)\mathbf{u} + \sin(t\Theta)\mathbf{z}) \times \Theta(-\sin(t\Theta)\mathbf{u} + \cos(t\Theta)\mathbf{z}) = \Theta\mathbf{u} \times \mathbf{z} = \Theta \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} .$$

Therefore, if we define the *rotation vector $\boldsymbol{\omega}$* by

$$\boldsymbol{\omega} = 2\Theta \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} , \tag{2.54}$$

Thus, $\mathbf{x}(t)$ satisfies the equation $\mathbf{x}'(t) = \boldsymbol{\omega} \times \mathbf{x}(t)$. That is, in terms of $\mathbf{f}_t(\mathbf{x}_0)$,

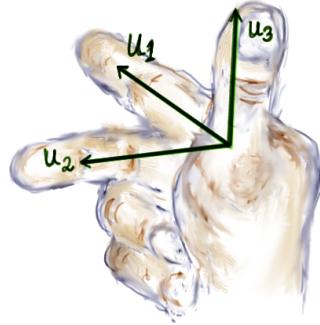
$$\frac{d}{dt} \mathbf{f}_t(\mathbf{x}_0) = \boldsymbol{\omega} \times \mathbf{f}_t(\mathbf{x}_0) . \quad (2.55)$$

You will recognize this as having the form of the Frenet-Serret equations written in terms of the Darboux vector. Therefore, what the Frenet-Serret equations describe is the instantaneous rotation process that carries the moving frame $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ along as t advances: At each time t , it is the change in $\{\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)\}$ is the same as if these vectors were undergoing a right-handed rotation along the direction of the Darboux vector $\boldsymbol{\omega}$ and a constant angular speed $\|\boldsymbol{\omega}\|$. This explains the general meaning of the Darboux vector.

We can finally explain the terminology “right handed orthonormal basis”. We begin by making an identification of \mathbb{R}^3 with the physical three dimension space around us. This requires us to identify the standard basis vectors \mathbf{e}_1 , \mathbf{e}_2 and \mathbf{e}_3 in \mathbb{R}^3 with three orthogonal directions in physical space.

To do this, fix three orthogonal directions in physical space – for instance, East, North and “straight up” might be good choices for somebody standing anywhere on the Earth except the North or South Poles. Next, take your right hand, and arrange you thumb and fingers so that your thumb, index finger and middle finger each point in one of these three orthogonal directions, as in the picture below. *At this stage of the process*, we number the directions: Identify \mathbf{e}_1 with the direction in which your index finger points, identify \mathbf{e}_2 with the direction in which your middle finger points, and identify \mathbf{e}_3 with the direction in which your index thumb points.

Now let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be some other set of three orthogonal directions. Try to rigidly rotate your right hand (keeping the index finger, middle finger and thumb orthogonal) around so that your index finger points in the direction of \mathbf{u}_1 , your middle finder points in the direction of \mathbf{u}_2 , and your thumb points in the direction of \mathbf{u}_3 , as in the picture:



- If this is possible, then the basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is right handed, and otherwise, it is not.

Indeed, if this motion of your hand *is possible*, then the motion of your hand provides a continuous interpolation between the reference basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ and $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. By what we have seen at the beginning, this means that $(\mathbf{u}_1 \times \mathbf{u}_2) \cdot \mathbf{u}_3 = (\mathbf{e}_1 \times \mathbf{e}_2) \cdot \mathbf{e}_3 = 1$ and hence $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is right handed.

Conversely, if $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is right-handed, like $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, then there is a rotation process that carries $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ over to $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. Therefore, if you arrange your right hand so that your index finger points in the \mathbf{e}_1 direction, your middle finger in the \mathbf{e}_2 direction, and your thumb in the \mathbf{e}_3

direction, and you then rotate your right hand about the corresponding axis of rotation, through the corresponding angle of rotation, your right hand will indeed be oriented as in the picture.

We may now also explain the “*right-hand rule*”: Let \mathbf{a} and \mathbf{b} be two vectors in \mathbb{R}^3 such that neither is a multiple of the other. Let \mathbf{b}_\perp be the component of \mathbf{b} that is orthogonal to \mathbf{a} , and define a right-handed orthonormal basis by

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{a}\|} \mathbf{a}, \quad \mathbf{u}_2 = \frac{1}{\|(\mathbf{b})_\perp\|} (\mathbf{b})_\perp \quad \text{and} \quad \mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2.$$

Then

$$\mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{b}_\perp = \|\mathbf{a}\| \|\mathbf{b}_\perp\| \mathbf{u}_1 \times \mathbf{u}_3 = \|\mathbf{a}\| \|\mathbf{b}_\perp\| \mathbf{u}_3.$$

That is, $\mathbf{a} \times \mathbf{b}$ is a positive multiple of \mathbf{u}_3 .

This means that if you configure your right hand as in the picture, with your thumb pointing in the direction of \mathbf{a} , and your index finger in the plane containing \mathbf{a} and \mathbf{b} , then your middle finger points in the direction of \mathbf{u}_3 ; i.e., in the direction of $\mathbf{a} \times \mathbf{b}$. This is commonly called the *right-hand rule* for the direction of $\mathbf{a} \times \mathbf{b}$.

2.2 Exercises

2.1 Let $\mathbf{x}(t) = (t+1, t^2)$. This is a parameterization of the parabola $y = (x-1)^2$.

- (a) Compute $\mathbf{v}(t) = \mathbf{x}'(t)$ and $\mathbf{a}(t) = \mathbf{x}''(t)$.
- (b) Compute $v(t)$ and $\mathbf{T}(t)$.
- (c) Find the tangent line to this curve at $t = 1$.

2.2 Let $\mathbf{x}(t) = (t^{-2}, 4/\sqrt{t}, t)$ for $t > 0$.

- (a) Compute $\mathbf{v}(t) = \mathbf{x}'(t)$ and $\mathbf{a}(t) = \mathbf{x}''(t)$.
- (b) Compute $v(t)$ and $\mathbf{T}(t)$.
- (c) Find the tangent line to this curve at $t = 1$.

2.3 Let $\mathbf{x}(t)$ and $\mathbf{y}(t)$ be two continuous curves in \mathbb{R}^n . Show that $f(t) := \mathbf{x}(t) \cdot \mathbf{y}(t)$ is a continuous real valued function of t . Also for $n = 3$, show that

$$\mathbf{z}(t) := \mathbf{x}(t) \times \mathbf{y}(t)$$

is a continuous curve in \mathbb{R}^3 .

2.4 Let $\mathbf{x}(t) = (\cos(t), \sin(t), t/r)$ where $r > 0$. The curve $\mathbf{x}(t)$ is a helix in \mathbb{R}^3 .

- (a) Compute $\mathbf{v}(t)$ and $\mathbf{a}(t)$.
- (b) Compute $v(t)$ and $\mathbf{T}(t)$.
- (c) Compute the curvature $\kappa(t)$ and the torsion $\tau(t)$, as well as $\mathbf{N}(t)$ and $\mathbf{B}(t)$.
- (d) Compute the Darboux vector $\boldsymbol{\omega}(t)$.
- (e) Find the tangent line to this curve at $t = \pi/4$, and the equation of the osculating plane to the curve at $t = \pi/2$. Find the intersection of this line and plane.

2.5: Let $\mathbf{x}(t)$ be the curve given by

$$\mathbf{x}(t) = (e^t \cos t, e^t \sin t, e^t) .$$

(a) Compute the arc length $s(t)$ as a function of t , measured from the starting point $\mathbf{x}(0)$, and find an arc-length parameterization of this curve

(b) Compute curvature $\kappa(t)$ and torsion $\tau(t)$ as a function of t .

(c) Find an equation for the osculating plane at time $t = 0$

2.6: Let $\mathbf{x}(t)$ be the curve given by

$$\mathbf{x}(t) = (t^{3/2}, 3t, 6t^{1/2})$$

for $t > 0$.

(a) what is the arc length along the curve between $\mathbf{x}(1)$ and $\mathbf{x}(4)$?

(b) Compute curvature $\kappa(t)$ and torsion $\tau(t)$ as a function of t .

(c) Find an equation for the osculating plane at $t = 1$, and find a parameterization of the tangent line to the curve at $t = 1$.

2.7: Let $\mathbf{x}(t)$ be the curve given by $\mathbf{x}(t) = (t, t^2/2, t^3/3)$.

(a) Find the equation of the osculating plane at $t = 1$.

(b) Compute the distance from the origin to the osculating plane at $t = 1$.

2.8: Let $\mathbf{x}(t)$ be the curve given by

$$\mathbf{x}(t) = (2t, t^2, t^3/3) .$$

(a) Compute the arc length $s(t)$ as a function of t , measured from the starting point $\mathbf{x}(0)$.

(b) Compute curvature $\kappa(t)$ and torsion $\tau(t)$ as a function of t .

(c) Find equations for the osculating planes at time $t = 0$ and $t = 1$, and find a parameterization of the line formed by the intersection of these planes.

2.9 Consider the ellipse in \mathbb{R}^2 given by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

where $a, b > 0$.

(a) Show that the path traced out by the parameterized curve $\mathbf{x}(t) = (a \cos(t), b \sin(t))$ is this ellipse. In other words, $\mathbf{x}(t) = (a \cos(t), b \sin(t))$ is a parameterization of this ellipse.

(b) Compute the curvature $\kappa(t)$, and find the minimum and maximum values of curvature on the ellipse, and the places where the curvature takes on these values.

2.10 Let $\mathbf{x}(t)$ be the curve given by $\mathbf{x}(t) = (t, \sqrt{2} \ln(t), 1/t)$ for $t > 0$.

(a) Find the arc length along the curve from $\mathbf{x}(1)$ to $\mathbf{x}(3)$.

- (b) Find the arc length along the curve from $\mathbf{x}(1)$ to $\mathbf{x}(t)$ as a function of t .
(c) Find the arc length parameterization $\mathbf{x}(s)$ of this curve.

2.11 Find the arc length along the parabola $y = (x - 1)^2$ from the point $(0, 1)$ to the point $(1, 0)$.
(See Exercise 2.1.)

2.12 Find the arc length parameterization of the curve given by $\mathbf{x}(t) = (t^{-2}, 4/\sqrt{t}, t)$ for $t > 0$.
(See Exercise 2.2.) What is the arc length along the segment of the curve joining $\mathbf{x}(1)$ and $\mathbf{x}(4)$?

2.13 Let $\mathbf{b} = (2, 1, 2)$. Let $\mathbf{x}(t)$ be the curve given satisfying the initial value problem

$$\mathbf{x}'(t) = \mathbf{b} \times \mathbf{x}(t) \quad \text{and} \quad \mathbf{x}(0) = (1, 1, 1).$$

- (a) Compute $\mathbf{x}(\pi)$ and find the arc length along the curve from $\mathbf{x}(0)$ to $\mathbf{x}(\pi)$.

- (b) Compute the curvature and torsion for this curve as a function of t .

2.14 Let $\mathbf{b} = (4, 7, 4)$. Let $\mathbf{x}(t)$ be the curve given satisfying the initial value problem

$$\mathbf{x}'(t) = \mathbf{b} \times \mathbf{x}(t) \quad \text{and} \quad \mathbf{x}(0) = (2, 2, 1).$$

- (a) Compute $\mathbf{x}(\pi)$ and find the arc length along the curve from $\mathbf{x}(0)$ to $\mathbf{x}(\pi)$.

- (b) Compute the curvature and torsion for this curve as a function of t .

2.15 Show that for $b > 0$ and $0 \leq a < 1$, the set of points (x, y) that satisfy (??) is an ellipse with one focus at the origin, and the other at $(-2f, 0)$, and semi-major axis R_+ where f and R_+ are given in terms of a and b by (??).

2.16 Let $\mathbf{x}(t)$ be the curve given by $\mathbf{x}(t) = (\cos t + 1, \cos t + \sin t, \sin t + 1)$.

- (a) Compute curvature $\kappa(t)$ and torsion $\tau(t)$ as a function of t .

- (b) Find an equation for the osculating plane at time $t = 0$

- (c) Find the distance between the plane given by $x - y + z = 0$ to $\mathbf{x}(t)$ as a function of t .

2.17 Consider the helix whose Darboux vector is $(3, 0, 4)$, with $\mathbf{x}(0) = \mathbf{0}$, and $\{\mathbf{T}(0), \mathbf{N}(0), \mathbf{B}(0)\} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. Find a formula for $\mathbf{x}(s)$, the arc-length parameterization of the helix.

2.18 The latitude and longitude of Milan Italy is $45^\circ 27'' N$ $9^\circ 10'' E$. The latitude and longitude of Cairo Egypt is $30^\circ 2'' N$ $31^\circ 21'' E$. Using this information, and the value of 6371 kilometers for the radius of the Earth, and the assumption that the Earth is spherical, compute the length of the shortest route on the surface of the Earth from Milan to Cairo.

2.19 Consider the vectors

$$\mathbf{u} = \frac{1}{3}(2, 1, 0, 2) \quad \text{and} \quad \mathbf{w} = \frac{1}{15}(10, -5, 8, 6).$$

These vectors both belong to S^3 , the unit sphere in \mathbb{R}^4 . Find a continuous curve $\mathbf{u}(t)$ defined on some interval $[0, T]$, some $T > 0$, that is continuously differentiable on $(0, T)$, with each $\mathbf{u}(t) \in S^3$, $\mathbf{u}(0) = \mathbf{u}$, $\mathbf{u}(T) = \mathbf{w}$, and whose arc length is minimal among all such curves.

2.20 Let $\mathbf{a} = \frac{1}{3}(2, 1, 2)$. Let $\mathbf{u} = \frac{1}{3}(1, 2, -2)$, and note that this is a unit vector orthogonal to \mathbf{a} . Find a unit vector \mathbf{v} so that $\mathbf{f}(\mathbf{x}) := \mathbf{h}_{\mathbf{v}}(\mathbf{h}_{\mathbf{u}}(\mathbf{x}))$ is the rotation of \mathbf{x} through the angle $\theta = \pi/3$ about the axis along \mathbf{a} , and then compute $\mathbf{f}((1, 1, 1))$.

2.21 Let $\mathbf{x}(t)$ be a twice differentiable curve in \mathbb{R}^3 such that $v(t_0) > 0$. Let \mathbf{v} and \mathbf{a} denote the velocity and acceleration at time t_0 . Let v and κ denote the speed and curvature at time t_0 . Prove that

$$\kappa = \frac{\|\mathbf{v} \times \mathbf{a}\|}{v^3}. \quad (2.56)$$

2.22 Let $\mathbf{x}(t)$ be a thrice differentiable curve in \mathbb{R}^3 such that $v(t_0) > 0$ and $\kappa(t_0) > 0$. Let \mathbf{v} and \mathbf{a} denote the velocity and acceleration at time t_0 . Let v and κ denote the speed and curvature at time t_0 . Prove that τ , the torsion at time t_0 , is given by

$$\tau = \frac{\mathbf{a}' \cdot \mathbf{v} \times \mathbf{a}}{v^6 \kappa^2} = -\frac{\mathbf{x}''' \cdot \mathbf{x}'' \times \mathbf{x}'}{v^6 \kappa^2}. \quad (2.57)$$

Chapter 3

CONTINUOUS FUNCTIONS

3.1 Continuity in several variables

3.1.1 Functions of several variables

Consider a function \mathbf{f} from \mathbb{R}^n to \mathbb{R}^m . Such a function takes a vector variable \mathbf{x} as input, and returns a vector $\mathbf{f}(\mathbf{x})$ as output. For example, consider the function \mathbf{f} from \mathbb{R}^2 to \mathbb{R}^3 given by

$$\mathbf{f}(x, y) = (x^2 + y^2, xy, x^2 - y^2). \quad (3.1)$$

(We will usually use the notation $\mathbf{f}(\mathbf{x})$ or $\mathbf{f}(x, y)$ instead of the more cumbersome $\mathbf{f}((x, y))$, but they all mean the same thing.) Introducing the functions

$$f_1(\mathbf{x}) = x^2 + y^2 \quad f_2(\mathbf{x}) = xy \quad \text{and} \quad f_3(\mathbf{x}) = x^2 - y^2,$$

we can write \mathbf{f} in (3.1) as a vector whose entries are functions from \mathbb{R}^2 to \mathbb{R} :

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}))$$

Often, the questions we ask about $\mathbf{f}(\mathbf{x})$ can be answered by considering the entry functions f_1 , f_2 and f_3 one at a time.

What kinds of questions will we be asking about such functions? Many of the questions have to do with *solving equations involving* \mathbf{f} . For example, consider the equation

$$\mathbf{f}(\mathbf{x}) = (2, 1, 0). \quad (3.2)$$

We can rewrite this as a *system of equations* using the entry functions f_1 , f_2 and f_3 :

$$\begin{aligned} f_1(x, y) &= 2 \\ f_2(x, y) &= 1 \\ f_3(x, y) &= 0. \end{aligned} \quad (3.3)$$

More explicitly,

$$\begin{aligned} x^2 + y^2 &= 2 \\ xy &= 1 \\ x^2 - y^2 &= 0. \end{aligned} \tag{3.4}$$

To solve a system of equations in several variables, one has to eliminate variables. In this case, elimination is not hard: Adding the first and third equation, we find $2x^2 = 2$, or $x^2 = 1$. The third equation now tells us $y^2 = x^2 = 1$. So $x = \pm 1$ and $y = \pm 1$. Going to the second equation, we see that if $x = 1$, then $y = 1$ also, and if $x = -1$, then $y = -1$ also. Hence the equation (3.2) has exactly two solutions:

$$\mathbf{x}_1 = (1, 1) \quad \text{and} \quad \mathbf{x}_2 = (-1, -1).$$

That is, for these vector \mathbf{x}_1 and \mathbf{x}_2 ,

$$\mathbf{f}(\mathbf{x}_1) = \mathbf{f}(\mathbf{x}_2) = (2, 1, 0),$$

and no other input vectors \mathbf{x} yield the desired output.

In general, it is not easy to solve vector equations in vector variables of the form $\mathbf{f}(\mathbf{x}) = \mathbf{b}$ except in the special case that $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is linear. In the linear case, as we shall see, there are very effective algorithms for computing the solution in a finite number of operations. Sometimes, as in our example just above, one can also do this for non-linear functions \mathbf{f} as well.

However, it is not always possible to arrive at the solution of a non-linear equation in finitely many steps. The way forward is to use, in principle, *infinitely many steps*. This is not as bad as it may sound. There are good methods, such as *Newton's method* for producing a sequence of vectors $\{\mathbf{x}_k\}$ with the property that

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{z} \tag{3.5}$$

where \mathbf{z} is an *exact* solution of the equation $\mathbf{f}(\mathbf{x}) = \mathbf{b}$; i.e.,

$$\mathbf{f}(\mathbf{z}) = \mathbf{b}. \tag{3.6}$$

Since human beings cannot do infinite computations, we can never carry out all of the computations needed to arrive at \mathbf{z} ; we must stop at the k th stage for some k , and be satisfied with the approximation \mathbf{x}_k . But this is not so bad. Even with familiar numbers like π , which is one solution to the equation $\sin(x) = 0$, we can only compute – exactly – a finite number of digits in its decimal representation. It will be the same here when we apply Newton's method to solving equations involving functions of vector variables: We will generate a *sequence of approximate solutions of rapidly improving accuracy*, and from this sequence, we can *exactly* compute any desired number of decimals in each of the entries of the solution to which the sequence converges.

- In this sense, methods of successive approximations yield methods of exact computation.

As soon as we contemplate methods of successive approximation for the solution of equations like $\mathbf{f}(\mathbf{x}) = \mathbf{b}$, we are led to the concept of continuity for functions of a vector variable. Suppose you

have a sequence $\{\mathbf{x}_k\}$ of vectors in the domain of \mathbf{f} satisfying (3.5) where \mathbf{z} satisfies (3.6). We would like to think of the vectors \mathbf{x}_k , at least for large k , as *approximate solutions* of the equation $\mathbf{f}(\mathbf{x}) = \mathbf{b}$. Thus we would hope that for large k , $\|\mathbf{f}(\mathbf{x}_k) - \mathbf{b}\|$ would be very small. More precisely, we would hope that:

$$\lim_{k \rightarrow \infty} \mathbf{f}(\mathbf{x}_k) = \mathbf{b} \quad (3.7)$$

In this case, we are justified in considering $\{\mathbf{x}_k\}$ as a sequence of approximate solutions of the equation $\mathbf{f}(\mathbf{x}) = \mathbf{b}$. However, there are functions \mathbf{f} , even of one variable, for which (3.5) and (3.6) do not imply (3.7).

Example 39 (Bad behavior under limits). *Let $f(x)$ be the function on \mathbb{R} defined by*

$$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Consider the sequence $\{x_k\}$ given by $x_k = 1/k$. Let $z = 0$, Then

$$\lim_{k \rightarrow \infty} x_k = z \quad \text{and} \quad f(z) = 0 ,$$

however

$$\lim_{k \rightarrow \infty} f(x_k) = 1 \neq 0 .$$

The function f in this example is discontinuous: It has a “jump” at $x = 0$.

The important class of functions for which (3.5) and (3.6) *do* imply (3.7) is the class of continuous functions. However, when the input variable is a vector in \mathbb{R}^n , $n \geq 2$, the notion of continuity is somewhat more subtle than it is when the input variable is in \mathbb{R}^1 : *A function on \mathbb{R}^2 can be discontinuous without having any jumps*, as we explain next.

3.1.2 Continuity in several variables

In plain words, but a bit roughly, this is what continuity at \mathbf{x}_0 means for a function \mathbf{f} from \mathbb{R}^n to \mathbb{R}^m :

- *We are guaranteed that $\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}_0)$ up to any given small margin of error provided that $\mathbf{x} \approx \mathbf{x}_0$ up to some other small enough margin of error: The output of the function will be close to $\mathbf{f}(\mathbf{x}_0)$ provided the input is sufficiently close to \mathbf{x}_0 .*

Let us remove the roughness, and make this precise. There are two margins of error involved – one on the input, and one on the output. Let $\delta > 0$ denote the margin of error on the input, and let $\epsilon > 0$ be the margin of error on the output. We are looking for a guarantee that if $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, then $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| < \epsilon$, and we want there to be such a guarantee with $\delta > 0$ no matter how small $\epsilon > 0$ may be. Of course δ , the margin of error on the input, will depend on ϵ . But continuity means such a guarantee is possible for all $\epsilon > 0$.

It is not be possible to make such a guarantee for all functions. Even for the single variable function in Example 39, such a guarantee is not possible when $x_0 = 0$ and ϵ is any number less than 1: There exist x arbitrarily close to $x_0 = 0$ with $|f(x) - f(x_0)| = 1$. This function f is somewhat hypersensitive: You change the input ever so slightly, and the response changes completely.

Definition 34 (Continuity). A function \mathbf{f} from \mathbb{R}^n to \mathbb{R}^m is continuous at \mathbf{x}_0 in case for every $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \delta(\epsilon) \quad \Rightarrow \quad \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| \leq \epsilon . \quad (3.8)$$

for all \mathbf{x} in the domain of \mathbf{f} . The function \mathbf{f} is continuous if it is continuous at each \mathbf{x}_0 in its domain.

Whether a function is continuous or not is a matter of considerable practical importance.

- If a function \mathbf{f} from \mathbb{R}^n to \mathbb{R}^m is not continuous at a solution \mathbf{x}_0 of $\mathbf{f}(\mathbf{x}) = \mathbf{b}$, it is no use at all to find a vector \mathbf{x}_1 with even $\|\mathbf{x}_1 - \mathbf{x}_0\| < 10^{-300}$ since without continuity, there is no guarantee that $\mathbf{f}(\mathbf{x}_1)$ is at all close to $\mathbf{f}(\mathbf{x}_0) = \mathbf{b}$.

Without continuity, only exact solutions are meaningful. But these will often involve irrational numbers that cannot be exactly represented on a computer. Therefore, whether a function is continuous or not is a serious practical matter.

How do we tell if a function is continuous? Sometimes one can check this *directly* from the definition:

Example 40 (Continuity of linear functions from \mathbb{R}^n to \mathbb{R}). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be given by

$$f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} .$$

We may assume $\mathbf{a} \neq \mathbf{0}$, or else f is constant, and hence obviously continuous.

The basis question before us, the question of continuity of f , amounts to the question: How large is $|f(\mathbf{x}) - f(\mathbf{x}_0)|$ compared to $\|\mathbf{x} - \mathbf{x}_0\|$? The Cauchy-Schwarz inequality provides the means to make the comparison. (Comparisons are exactly what inequalities are all about.) We have:

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{x}_0)| &= |\mathbf{a} \cdot \mathbf{x} - \mathbf{a} \cdot \mathbf{x}_0| \\ &= |\mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0)| \\ &\leq \|\mathbf{a}\| \|\mathbf{x} - \mathbf{x}_0\| . \end{aligned}$$

It follows that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \frac{1}{\|\mathbf{a}\|} \epsilon \Rightarrow |f(\mathbf{x}) - f(\mathbf{x}_0)| \leq \epsilon .$$

Thus, with $\delta = \epsilon / \|\mathbf{a}\|$, we have the guarantee we seek, and f is continuous at \mathbf{x}_0 . Since \mathbf{x}_0 was any vector in \mathbb{R}^n , f is continuous on \mathbb{R}^n .

A very important special case is that of the coordinate functions

$$c_j(x_1, \dots, x_n) = x_j .$$

Note that

$$c_j(\mathbf{x}) = \mathbf{e}_j \cdot \mathbf{x} ,$$

and so the coordinate functions are continuous, as a spacial case of the example treated here.

The next theorem says that all questions about the continuity of functions from \mathbb{R}^n to \mathbb{R}^m reduce to questions about continuity of functions from \mathbb{R}^n to \mathbb{R} .

Theorem 33 (Continuity and continuity of the entry functions). *Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ where*

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) .$$

Then \mathbf{f} is continuous if and only if each f_j is continuous as a function from \mathbb{R}^n to \mathbb{R} .

Proof. Suppose that $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous. Then for each $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ so that (3.8) is true. Then since for each j , $|f_j(\mathbf{x}) - f_j(\mathbf{x}_0)| \leq \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\|$,

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \delta(\epsilon) \quad \Rightarrow \quad |f_j(\mathbf{x}) - f_j(\mathbf{x}_0)| \leq \epsilon ,$$

and thus each f_j is continuous.

Conversely, suppose that each f_j is continuous. Then for any $\epsilon > 0$, there is a $\delta_j(\epsilon/\sqrt{n})$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \delta_j(\epsilon/\sqrt{n}) \quad \Rightarrow \quad |f_j(\mathbf{x}) - f_j(\mathbf{x}_0)| \leq \frac{\epsilon}{\sqrt{n}} .$$

But since

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| = \left(\sum_{j=1}^n |f_j(\mathbf{x}) - f_j(\mathbf{x}_0)|^2 \right)^{1/2} \leq \sqrt{n} \max_{j=1,\dots,n} \{|f_j(\mathbf{x}) - f_j(\mathbf{x}_0)|\} ,$$

if we define

$$\delta(\epsilon) = \min_{j=1,\dots,n} \{\delta_1(\epsilon/\sqrt{n}), \dots, \delta_n(\epsilon/\sqrt{n})\} ,$$

then $\delta(\epsilon) > 0$ and (3.8) is true, so that \mathbf{f} is continuous. \square

Example 41 (Continuity of linear functions from \mathbb{R}^n to \mathbb{R}^m). *Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ have the form $\mathbf{f}(\mathbf{x}) = (\mathbf{a}_1 \cdot \mathbf{x}, \dots, \mathbf{a}_m \cdot \mathbf{x})$ for some set of vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$ in \mathbb{R}^n . As we shall see, the general linear transformation from \mathbb{R}^n to \mathbb{R}^m has this form.*

Since each of the entry functions in \mathbf{f} is continuous, as shown in Example 40, it follows from Theorem 33 that \mathbf{f} is continuous.

The next theorem provides convenient means for checking continuity of functions from \mathbb{R}^n to \mathbb{R} . Before presenting the theorem, let us look at an important example to which it pertains.

Example 42. *Let $f(x, y)$ be given by*

$$f(x, y) = xy .$$

This function is a second order polynomial in the two variables x and y . As we shall soon see, all polynomials, in any finite number of variables, are continuous. The reasons this is true can be readily grasped by examining this simple example.

To show that f is continuous, pick any $\mathbf{x}_0 = (x_0, y_0) \in \mathbb{R}^2$. We must then compare $|f(x, y) - f(x_0, y_0)|$ with $\|\mathbf{x} - \mathbf{x}_0\| = \sqrt{(x - x_0)^2 + (y - y_0)^2}$. There are two variables involved, and the basic idea is to add and subtract so that one writes $f(x, y) - f(x_0, y_0)$ as a sum of differences in which only one variable at a time is changing. For example,

$$xy - x_0y_0 = (x - x_0)y + x_0(y - y_0) \tag{3.9}$$

so that

$$|xy - x_0y_0| \leq |x - x_0||y| + |x_0||(y - y_0)| . \quad (3.10)$$

As we have observed a number of times,

$$|x - x_0| \leq \sqrt{(x - x_0)^2 + (y - y_0)^2} = \|\mathbf{x} - \mathbf{x}_0\|$$

and, likewise, $|y - y_0| \leq \|\mathbf{x} - \mathbf{x}_0\|$. Therefore,

$$|f(x, y) - f(x_0, y_0)| = |xy - x_0y_0| \leq (|x_0| + |y|)\|\mathbf{x} - \mathbf{x}_0\| .$$

We are getting there, but there is still a factor of $|y|$ on the right hand side, and we would like the right hand side to be expressed only in terms of constants, such as x_0 and y_0 , and in terms of $\|\mathbf{x} - \mathbf{x}_0\|$. However, the factor of $|y|$ is readily dealt with: By the triangle inequality and what we have said above,

$$|y| = |y_0 + (y - y_0)| \leq |y_0| + |y - y_0| \leq |y_0| + \|\mathbf{x} - \mathbf{x}_0\| .$$

Putting it all together, we have

$$|f(x, y) - f(x_0, y_0)| = |xy - x_0y_0| \leq (|x_0| + |y_0| + \|\mathbf{x} - \mathbf{x}_0\|)\|\mathbf{x} - \mathbf{x}_0\| .$$

Whenever $\|\mathbf{x} - \mathbf{x}_0\| \leq 1$, this means that

$$|f(x, y) - f(x_0, y_0)| = |xy - x_0y_0| \leq (|x_0| + |y_0| + 1)\|\mathbf{x} - \mathbf{x}_0\| .$$

Therefore, if

$$\delta(\epsilon) := \min \left\{ \frac{\epsilon}{|x_0| + |y_0| + 1}, 1 \right\} ,$$

then

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \delta(\epsilon) \Rightarrow |f(\mathbf{x}) - f(\mathbf{x}_0)| \leq \epsilon .$$

This shows that f is continuous at \mathbf{x}_0 . Since \mathbf{x}_0 is any vector in \mathbb{R}^2 , f is continuous on \mathbb{R}^2 . The same “divide and conquer” strategy that was used in (3.9) can be used to show that the product of any two continuous functions f and g from \mathbb{R}^n to \mathbb{R} is continuous. This is part of the next theorem.

Theorem 34 (Building continuous functions from \mathbb{R}^n to \mathbb{R}). *Let f and g be continuous functions from some domain U in \mathbb{R}^n to \mathbb{R} . Define the functions fg and $f + g$ by $fg(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x})$ and $(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. Then fg and $f + g$ are continuous on U . furthermore, if $g \neq 0$ anywhere in U , then f/g defined by $(f/g)(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$ is continuous in U . Finally, if h is a continuous function from \mathbb{R} to \mathbb{R} , then the composition $h \circ f$ is continuous on U .*

Proof. Consider the case of $f + g$. Fix any $\epsilon > 0$, and any \mathbf{x}_0 in U . Since f and g are continuous there is a $\delta_f(\epsilon/2) > 0$ and a $\delta_g(\epsilon/2) > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \delta_f(\epsilon/2) \quad \Rightarrow \quad |f(\mathbf{x}) - f(\mathbf{x}_0)| \leq \epsilon/2$$

and

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \delta_g(\epsilon/2) \quad \Rightarrow \quad |g(\mathbf{x}) - g(\mathbf{x}_0)| \leq \epsilon/2$$

Now define

$$\delta(\epsilon) := \max\{\delta_f(\epsilon/2), \delta_g(\epsilon/2)\}.$$

Then, whenever $|\mathbf{x} - \mathbf{x}_0| \leq \delta(\epsilon)$,

$$|(f + g)(\mathbf{x}) - (f + g)(\mathbf{x}_0)| \leq |f(\mathbf{x}) - f(\mathbf{x}_0)| + |g(\mathbf{x}) - g(\mathbf{x}_0)| \leq \epsilon/2 + \epsilon/2 = \epsilon$$

so that

$$|\mathbf{x} - \mathbf{x}_0| \leq \delta(\epsilon) \quad \Rightarrow \quad |(f + g)(\mathbf{x}) - (f + g)(\mathbf{x}_0)| \leq \epsilon.$$

This proves the continuity of $f + g$. The other cases are similar. In fact, the proof for products very closely follows the treatment of the special case in Example 42, and the proof for composition is exactly like the proof in the corresponding single variable theorem. Finally, composition with $h(t) = 1/t$ allows one to reduce the analysis of division to that of multiplication. Therefore, the proofs for these remaining cases are left as exercises. \square

Example 43 (Continuity piece by piece). *To apply Theorem 34, try to recognize a function as being built out of known continuous pieces. For example, consider*

$$z(x, y) = \cos((1 + x^2 + y^2)^{-1}).$$

This is built out of the continuous building blocks

$$f(x, y) = x \quad g(x, y) = y \quad \text{and} \quad h(x, y) = 1.$$

Indeed,

$$z(\mathbf{x}) = \cos\left(\frac{h}{h + ff + gg}\right)(\mathbf{x}).$$

Repeated application of Theorem 34 then shows $z(\mathbf{x})$ is continuous.

Example 44 (Continuity of polynomials and rational functions). *A polynomial in several variables is a sum of monomials, which are functions of the form*

$$f(x_1, \dots, x_n) = a \prod_{j=1}^n x_j^{p_j}$$

where each p_j is a non-negative integer, and a is a constant. For example, on \mathbb{R}^3 with coordinates x , y , and z ,

$$3x^2y^3z \quad \text{and} \quad -2x^4z^2$$

are monomials. Since the coordinate functions are continuous, as shown in Example 40, and since products of continuous functions are continuous by Theorem 34, it follows that monomials are continuous. Then, by Theorem 34 once more, and sum of finitely many monomials is continuous, and hence every polynomial, e.g.,

$$f(x, y, z) = 3x^2y^3z - 2x^4z^2$$

is continuous.

A rational function is a ratio of polynomials; e.g.

$$f(x, y, z) = \frac{3x^2y^3z + x - yz}{1 + x^4y^2} .$$

By Theorem 34 once more, a rational function is continuous at all points \mathbf{x}_0 where the denominator is not zero. In the example at hand, the denominator is never zero, and the function is continuous on all of \mathbb{R}^3 .

3.1.3 Continuity and limits

When we wish to determine whether a given function is continuous or not, one way is to make direct use of the ϵ and δ definition. This is a good strategy for many problems, and it is what we have done until now.

The main theorem of this subsection gives us an alternative to the the ϵ and δ analysis of continuity. It provides a characterization of continuity in terms of limits: A function is continuous if and only if one can “pass limits to the inside of the function”.

Theorem 35 (Characterization of continuity in terms of limits). *Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Then \mathbf{f} is continuous at $\mathbf{x}_0 \in \mathbb{R}^n$ if and only whenever $\{\mathbf{x}_k\}$ is a sequence in \mathbb{R}^n with*

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}_0 ,$$

then

$$\lim_{k \rightarrow \infty} \mathbf{f}(\mathbf{x}_k) = \mathbf{f}(\mathbf{x}_0) .$$

Proof. Suppose that \mathbf{f} is continuous, and that $\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}_0$. We must show that $\lim_{k \rightarrow \infty} \mathbf{f}(\mathbf{x}_k) = \mathbf{f}(\mathbf{x}_0)$.

For this purpose, pick $\epsilon > 0$. Since \mathbf{f} is continuous, there exists $\delta(\epsilon) > 0$ so that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| \leq \epsilon$ whenever $\|\mathbf{x} - \mathbf{x}_0\| < \delta(\epsilon)$. Since $\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}_0$, there is an $N_{\delta(\epsilon)}$ so that $\|\mathbf{x}_k - \mathbf{x}_0\| \leq \delta(\epsilon)$ whenever $k \geq N_{\delta(\epsilon)}$. Therefore,

$$k \geq N_{\delta(\epsilon)} \quad \Rightarrow \quad \|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}_0)\| \leq \epsilon .$$

This shows that $\lim_{k \rightarrow \infty} \mathbf{f}(\mathbf{x}_k) = \mathbf{f}(\mathbf{x}_0)$.

On the other hand, suppose that \mathbf{f} is not continuous at \mathbf{x}_0 . Then for some $\epsilon > 0$, there is no $\delta(\epsilon) > 0$ such that every point \mathbf{x} with $\|\mathbf{x} - \mathbf{x}_0\|$ satisfies $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| < \epsilon$. In particular, for each natural number k , there exists \mathbf{x}_k such that $\|\mathbf{x}_k - \mathbf{x}_0\| < 1/k$ but $\|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}_0)\| \geq \epsilon$. Then $\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}_0$, but $\{\mathbf{f}(\mathbf{x}_k)\}$ cannot possibly converge to $\mathbf{f}(\mathbf{x}_0)$. \square

Example 45 (Analysis of continuity via limits). *Let $f(x, y)$ be given by*

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

Is f continuous?

To solve this problem, note that away from $(0, 0)$ f is a continuous rational function. Hence, the only question is whether it is continuous at $\mathbf{0} = (0, 0)$. The function is continuous if and only if for every sequence $\{\mathbf{x}_n\}$ with $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{0}$, it is the case that $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = f(\mathbf{0}) = 0$.

Let us try some sequences $\{\mathbf{x}_n\}$ that approach $\mathbf{0}$, starting with some obvious ones. Let the sequence approach $\mathbf{0}$ from along the y -axis, taking $\mathbf{x}_n = (1/n, 0)$. In this case, we get $f(\mathbf{x}_n) = 0$ for all n , so certainly $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = 0$. Since the function is symmetric in x and y , the same thing happens if we take the sequence to approach $\mathbf{0}$ from along the y -axis; i.e., taking $\mathbf{x}_n = (0, 1/n)$. But if we take the sequence to approach $\mathbf{0}$ from along the line $y = x$; that is with $\mathbf{x}_n = (1/n, 1/n)$, we find

$$f(1/n, 1/n) = \frac{1/n^2}{1/n^2 + 1/n^2} = \frac{1}{2}$$

for all n , and hence for this choice of $\{\mathbf{x}_n\}$, $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{0}$, but

$$\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = \frac{1}{2} \neq 0 = f(\mathbf{0}) .$$

Hence, this function is discontinuous.

In general, as soon as we have found one sequence for which $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{x}_0$, but $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) \neq f(\mathbf{x}_0)$, the analysis is over: The function is definitely not continuous at \mathbf{x}_0 .

As we have seen in the last example, a sequence $\{\mathbf{x}_n\}$ can approach a point \mathbf{x}_0 in a variety of ways. If the function f is to be continuous at \mathbf{x}_0 , it must be the case that $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = f(\mathbf{x}_0)$ no matter how the terms \mathbf{x}_n in the sequence approach \mathbf{x}_0 , horizontally, vertical along some angle or along some other more complicated curve.

It is therefore helpful to factor $\mathbf{x}_n - \mathbf{x}_0$ into its magnitude and direction: Given a sequence $\{\mathbf{x}_n\}$ and a point \mathbf{x}_0 , for each n , define

$$r_n = \|\mathbf{x}_n - \mathbf{x}_0\| \quad \text{and} \quad \mathbf{u}_n = \frac{1}{r_n}(\mathbf{x}_n - \mathbf{x}_0) .$$

Then \mathbf{u}_n is a unit vector and we have $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{x}_0$ if and only if $\lim_{n \rightarrow \infty} r_n = 0$. In particular, whether $\{\mathbf{x}_n\}$ converges to \mathbf{x}_0 does not depend at all on the sequence of direction vectors $\{\mathbf{u}_n\}$. For this reason, it is often useful to write $\mathbf{x}_n - \mathbf{x}_0 =: r_n \mathbf{u}_n$ so that

$$f(\mathbf{x}_n) = f(\mathbf{x}_0 + (\mathbf{x}_n - \mathbf{x}_0)) = f(\mathbf{x}_0 + r_n \mathbf{u}_n)$$

and then to see whether or not

$$|f(\mathbf{x}_0 + r \mathbf{u}) - f(\mathbf{x}_0)|$$

can be seen to be small for small values of r , independent of \mathbf{u} .

Example 46 (Analysis of continuity via limits). Let $f(x, y)$ be given by

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

Is f continuous?

To solve this problem, note that away from $(0, 0)$ f is a continuous rational function. Hence, the only question is whether it is continuous at $\mathbf{0} = (0, 0)$. The function is continuous if and only if for every sequence $\{\mathbf{x}_n\}$ with $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{0}$, it is the case that $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = f(\mathbf{0}) = 0$. And since $f(x, y) = 0$ everywhere along the x and y axes, if the function is to be continuous, it can only be the case that $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = 0$ for every sequence $\{\mathbf{x}_n\}$ that converges to $\mathbf{0}$.

Therefore, consider any such sequence $\{\mathbf{x}_n\} = \{(x_n, y_n)\}$. Define $r_n = \|\mathbf{x}_n\| = \sqrt{x_n^2 + y_n^2}$, and $\mathbf{u}_n = r_n^{-1} \mathbf{x}_n$. Since \mathbf{u}_n is a unit vector, we can write it as $\mathbf{u}_n = (\cos \theta_n, \sin \theta_n)$ for some angle θ . Then $(x_n, y_n) = r_n (\cos \theta_n, \sin \theta_n)$. (That is, we are using polar coordinates.)

Now compute

$$f(x_n, y_n) = \frac{r_n^3 \cos^2 \theta_n \sin \theta_n}{r_n^4 \cos^4 \theta_n + r_n^2 \sin^2 \theta_n}.$$

If θ_n is an integer multiple of π , this is zero, otherwise, we may divide through by $r_n^2 \sin^2 \theta_n$ to obtain

$$f(x_n, y_n) = \left(\frac{r_n \cos \theta_n \cot \theta_n}{r_n^2 \cos^2 \theta_n \cot^2 \theta_n + 1} \right).$$

Let us do a worst case analysis: For fixed r_n , what choice of θ_n makes $|f(x_n, y_n)|$ as large as possible?

To simplify the analysis, define $s = r_n \cos \theta_n \cot \theta_n$. As θ_n varies over the interval $(0, 2\pi)$, s takes on all values in $(-\infty, \infty)$. It is easy to see that the function $\varphi(s) = s/(1+s^2)$ has the maximum value $1/2$, achieved at $s = 1/2$, and the minimum value $-1/2$ achieved at $s = -1$. Thus, for any sequence $\{r_n\}$ converging to 0, there is a sequence of unit vectors $\{\mathbf{u}_n\}$ such that $f(r_n \mathbf{u}_n) = 1/2$ for all n . Since there is another sequence (running along the x -axis, say) for which the corresponding limit is 0, f is not continuous at $\mathbf{0}$.

We can even make our worst case analysis more explicit. We found that in the worst case, r_n and θ_n are related by $r_n \cos \theta_n \cot \theta_n = 1$, and multiplying through by $r_n \sin \theta_n$, this is equivalent to $(r_n \cos \theta_n)^2 = r_n \sin \theta_n$, which is the same as $x_n^2 = y_n$. That is, we see the worst case behavior when the sequence $\{\mathbf{x}_n\}$ approaches the origin along the parabola $y = x^2$. (One also has $f(x, y) = -1/2$, an equally bad case, along the parabola $y = -x^2$.)

Note that one cannot see the discontinuity of f at the origin by testing only with sequences that approach the origin along a line with a fixed direction. Suppose for example that we choose our sequence $\{\mathbf{x}_n\}$ along the line $x = ay$ for some number a . If we take $x_n = 1/n$, then $y_n = a/n$. Clearly with $\mathbf{x}_n = (1/n, a/n)$, $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{0}$, however,

$$f(1/n, a/n) = \frac{a/n^3}{1/n^4 + a^2/n^2} = \frac{a/n}{1/n^2 + a^2}$$

for all n , and hence also for this choice of $\{\mathbf{x}_n\}$, $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{0} = f(\mathbf{0})$. The same happens if our sequence approaches along a line of the form $y = ax$. The lack of continuity is only revealed when we consider a sequence that approaches the origin along an appropriate curve; lines will not suffice.

In the last two examples, we have shown that functions were discontinuous by finding a sequence $\{\mathbf{x}_n\}$ that converges to $\mathbf{0}$, but for which $\{f(\mathbf{x}_n)\}$ does not converge to $f(\mathbf{0})$. To show that a function is not continuous, you only need to display one such sequence.

On the other hand, to prove that a function is continuous at \mathbf{x}_0 , you must show that for every sequence $\{\mathbf{x}_n\}$ that converges to \mathbf{x}_0 , it is always the case that $\{f(\mathbf{x}_n)\}$ converges to $f(\mathbf{x}_0)$. We have

even seen that it is not enough to look only at sequences that approach \mathbf{x}_0 from along all lines; it can be that problems only show up for sequences that approach \mathbf{x}_0 along some curve, as in Example, 46 or perhaps even in a more complicated way.

The worst case analysis that we used in Example, 46 provides a way forward.

3.1.4 The Squeeze Principle in several variables

As in the single variable calculus, we can use the *Squeeze Principle* to make comparisons with simple functions.

Lemma 10 (Squeeze principle). *Let f be a given real valued function defined on some set $U \subset \mathbb{R}^n$ that contains \mathbf{x}_0 . Suppose also that there exists some $R > 0$ so that if $\|\mathbf{x} - \mathbf{x}_0\| < R$, then $\mathbf{x} \in U$. Suppose there is a continuous function $g(r)$ defined on $[0, R]$ with values $[0, \infty)$ and with $g(0) = 0$ such that for all unit vectors $\mathbf{u} \in \mathbb{R}^n$ and all $r \in (0, R]$,*

$$|f(\mathbf{x}_0 + r\mathbf{u}) - f(\mathbf{x}_0)| \leq g(r) . \quad (3.11)$$

Then f is continuous at \mathbf{x}_0 .

Proof. Consider any sequence $\{\mathbf{x}_n\}$ in U (where f is defined) such that $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{x}_0$. Define $r_n = \|\mathbf{x}_n - \mathbf{x}_0\|$ and $\mathbf{u}_n = r_n^{-1}(\mathbf{x}_n - \mathbf{x}_0)$ so that $\mathbf{x}_n = \mathbf{x}_0 + (\mathbf{x}_n - \mathbf{x}_0) = \mathbf{x}_0 + r_n \mathbf{u}_n$. Then for all n such that $r_n < R$,

$$0 \leq |f(\mathbf{x}_n) - f(\mathbf{x}_0)| = |f(\mathbf{x}_0 + r_n \mathbf{u}_n) - f(\mathbf{x}_0)| \leq g(r_n)$$

. Since g is continuous with $g(0) = 0$, $\lim_{n \rightarrow \infty} g(r_n) = g(\lim_{n \rightarrow \infty} r_n) = g(0) = 0$. Hence, by the Squeeze Principle for sequences of real numbers, $\lim_{n \rightarrow \infty} |f(\mathbf{x}_n) - f(\mathbf{x}_0)| = 0$. Since the convergent sequence was arbitrary, this proves the continuity. \square

Example 47 (Analysis of continuity via the Squeeze Principle, 1). *Let $f(x, y)$ be given by*

$$f(x, y) = \begin{cases} \frac{x^2 + y^2}{\sqrt{x^2 + y^2 + 1} - 1} & (x, y) \neq (0, 0) \\ 2 & (x, y) = (0, 0) . \end{cases}$$

Is f continuous?

To solve this problem, note that away from $(0, 0)$, f is a continuous function. Hence, the only question is whether it is continuous at $\mathbf{0}$.

To apply the Squeeze Principle, write $(x, y) = \mathbf{x} = r\mathbf{u}$, and note that

$$f(x, y) - 2 = \frac{r^2}{\sqrt{r^2 + 1} - 1} - 2 := g(r) .$$

In this example, the angular dependence drops out effortlessly! We need to check that if we define $g(0) = 0$, then g is continuous. By l'Hospital's rule

$$\lim_{r \rightarrow 0} \frac{r^2}{\sqrt{r^2 + 1} - 1} = \lim_{r \rightarrow 0} \frac{2r}{r/\sqrt{r^2 + 1}} = 2 ,$$

and so $\lim_{r \rightarrow 0} g(r) = 0$. This proves that f is continuous at the origin.

In our next example, the angular dependence does not drop out entirely, but the approach still leads to a simple proof of continuity.

Example 48 (Analysis of continuity via the Squeeze Principle, 2). *Let $f(x, y)$ be given by*

$$f(x, y) = \begin{cases} \frac{x^5}{x^4 + y^6} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0) \end{cases}.$$

Is f continuous?

To solve this problem, note that away from $(0, 0)$, f is a continuous function. Hence, the only question is whether it is continuous at $\mathbf{0}$.

To apply the Squeeze Principle, write $(x, y) = \mathbf{x} = r\mathbf{u}$, and note that with $\mathbf{u} = (\cos \theta, \sin \theta)$,

$$0 \leq |f(r\mathbf{u})| = r \frac{|\cos^5 \theta|}{\cos^4 \theta + r^2 \sin^6 \theta} = r \frac{|\cos \theta|}{1 + r^2 \sin^2 \theta \tan^4 \theta}.$$

In this case we can get away with doing a separate worst case analysis of the numerator and denominator in $\frac{|\cos^5 \theta|}{\cos^4 \theta + r^2 \sin^6 \theta}$. (Such a separate analysis is not always possible; See Example 46). No matter what θ is, the numerator is bounded above by 1. Also no matter what θ is, the denominator is bounded below by 1. Hence the ratio is bounded above by 1. Therefore,

$$0 \leq |f(r\mathbf{u})| \leq r.$$

Hence we may apply the Squeeze Principle with $g(r) = r$, which is continuous and has $g(0) = 0$. Therefore f is continuous at the origin.

It is not always necessary to write things out in terms of r and \mathbf{u} . Often one can see what is going on by making simple comparisons. The following inequalities are often helpful in this regard.

Lemma 11 (Sums and powers). *For all numbers $a, b \geq 0$, and all $p > 0$*

$$\frac{1}{2}(a^p + b^p) \leq (a + b)^p \leq 2^p(a^p + b^p). \quad (3.12)$$

Also

$$ab \leq \frac{a^2 + b^2}{2} \quad (3.13)$$

and the is equality in (3.13) if and only if $a = b$.

Proof. We may assume $b \leq a$. Then $a/2 \leq (a + b)/2 \leq a$. Since for all $p > 0$, the p th power function is monotone increasing,

$$\left(\frac{a}{2}\right)^p \leq \left(\frac{a+b}{2}\right)^p \leq a^p.$$

But since $(a^p + b^p)/2 \leq a^p$, and $a^p \leq a^p + b^p$, we have

$$\frac{1}{2} \left(\left(\frac{a}{2}\right)^p + \left(\frac{b}{2}\right)^p \right) \leq \left(\frac{a+b}{2}\right)^p \leq a^p + b^p.$$

Multiplying through by 2^p we obtain (3.12).

Finally,

$$0 \leq \frac{(a-b)^2}{2} = \frac{a^2 + b^2}{2} - ab,$$

which proves the last part. □

Before turning to our next example, let us explain how (3.13) can motivate the choice of the sequence $(1/n, 1/n^2)$ that we used in Example 46.

In studying the continuity of ratios, we must compare the sizes of the numerator and denominator. For the function in Example 46, the denominator is $x^4 + y^2$ and the numerator is x^2y . By (3.13),

$$x^4 + y^2 \geq 2x^2y$$

with equality if and only if $x^2 = y$. Hence (3.13) this ratio is maximal along the parabola $x^2 = y$, and this motivates examining the behavior of f along the sequence $(1/n, 1/n^2)$.

Example 49 (Analysis of continuity via the Squeeze Principle, 3). *Let $f(x, y)$ be given by*

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

where $p > 0$. For which values of p is f continuous at $\mathbf{0} = (0, 0)$?

To answer this question, focus first on the denominator – that is where the complexity lies. Note that

$$|x|^p + |y|^p = (x^2)^{p/2} + (y^2)^{p/2}.$$

Hence by (3.12), applied with $p/2$ in place of p and some rearranging of terms,

$$2^{-p/2}(x^2 + y^2)^{p/2} \leq (x^2)^{p/2} + (y^2)^{p/2} \leq 2(x^2 + y^2)^{p/2}.$$

That is,

$$2^{-p/2}\|\mathbf{x}\|^p \leq |x|^p + |y|^p \leq 2\|\mathbf{x}\|^p.$$

Let us write $x = \|\mathbf{x}\| \cos \theta$ and $y = \|\mathbf{x}\| \sin \theta$ where θ is the angle between \mathbf{x} and the x -axis. (This amounts to using polar coordinates). We then have that for $\mathbf{x} \neq \mathbf{0}$,

$$2^{-p/2}\|\mathbf{x}\|^{2-p}2\sin \theta \cos \theta \leq f(\mathbf{x}) \leq 2\|\mathbf{x}\|^{2-p}2\sin \theta \cos \theta.$$

It is now clear that for $p \geq 2$, f does not have a limit at $\mathbf{x} = 0$, while for $0 < p < 2$, we have that

$$|f(\mathbf{x}) - 0| \leq g(\|\mathbf{x}\|)$$

where $g(t) = 2t^{2-p}$, and $\lim_{t \rightarrow 0} g(t) = 0$. Hence, for these values of p , f is continuous by the Squeeze Principle.

3.1.5 Continuity versus separate continuity

Many questions about the behavior of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ can be answered by examining “single variable slices” of the function. We now explain this fundamental strategy in some detail.

Given any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, and any parameterized line $\mathbf{x}_0 + t\mathbf{v}$ in \mathbb{R}^n , define the function $g : \mathbb{R} \rightarrow \mathbb{R}$ by

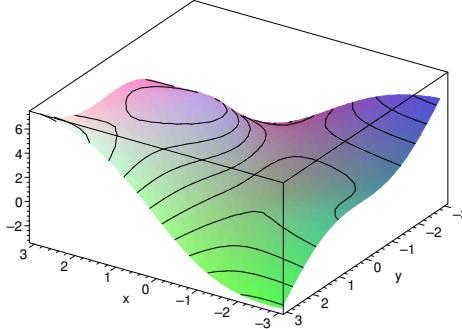
$$g(t) = f(\mathbf{x}_0 + t\mathbf{v}).$$

The function g is a garden variety single variable function, and it describes a “slice” of the function f . Studying all such slices, we can learn many useful things about f .

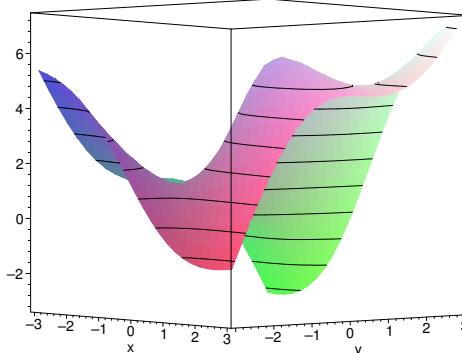
For example, consider the function $f(x, y)$ given by

$$f(x, y) = \frac{3(1+x)^2 + xy^3 + y^2}{1+x^2+y^2} .$$

Here is a plot of the graph of $z = f(x, y)$ for $-3 \leq x, y \leq 3$:



Here is another picture of the same graph, but from a different angle that give more of a side view:

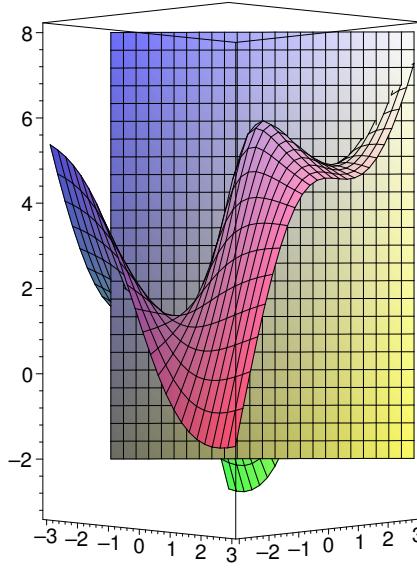


In both graphs, the curves drawn on the surface show points that have the same z value, which we can think of as representing “altitude”. Drawing them in helps us get a good visual understanding of the “landscape” in the graph. They are called *contour curves*, and are drawn in on any topographic map. (The formula for $f(x, y)$ was chosen to produce a graph that looks like a bit of a mountain landscape.)

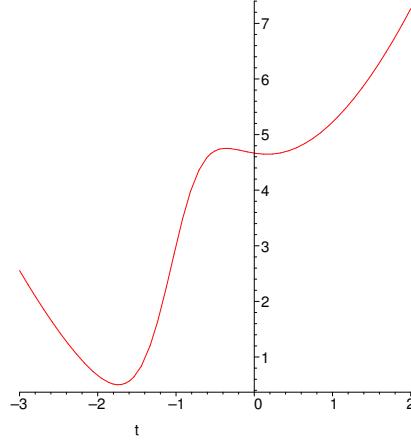
Now that we understand what the graph of $z = f(x, y)$ looks like, let’s *slice it along a line*. Suppose for example that you are walking on a path in this landscape, and that the projection of your path on the surface down into the x, y plane is the line $\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{v}$ with

$$\mathbf{x}_0 = (1, 1) \quad \text{and} \quad \mathbf{v} = (1, 1) .$$

Consider the vertical plane over this line; i.e., the plane in \mathbb{R}^3 that contain this line and the z -axis. Here is a picture of this vertical plane slicing through the graph of $z = f(x, y)$:



The next graph shows the “altitude profile” as we walk along the graph; this curve is where the surface intersects our vertical plane:



Compare the last two graphs, and make sure you see how the curve in the second one corresponds to the intersection of the plane and the surface in the first one.

In this example, we have illustrated the slicing strategy with a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ only so we could plot graphs and produce visual aids to our understanding. But the basic formula

$$g(t) = f(\mathbf{x}_0 + t\mathbf{v})$$

can be used in any number of dimensions.

The slicing strategy turns out to be a very useful method for studying functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$, as we shall see. However it has its limitations. For instance, one might hope that a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous if and only if for each choice of \mathbf{x}_0 and \mathbf{v} in \mathbb{R}^n , the function $g(t) = f(\mathbf{x}_0 + t\mathbf{v})$ is continuous. This is not the case.

In fact, we have already seen this in Example 46. While the function f in Example 46 is discontinuous at the origin, its slice along every line through the origin (and therefore every line, since the only discontinuity is at the origin) is continuous. You could walk along the “landscape”

described by f on any straight line path through the origin, but you would never run into a cliff. Otherwise put, there are no jumps in any linear slice.

Here is an even more dramatic example.

Example 50 (Continuous on every line through $\mathbf{0}$, but unbounded near $\mathbf{0}$). Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$f(x, y) = \begin{cases} \frac{1}{x^2 + y^2} \frac{2x^4 y}{x^8 + y^2} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0). \end{cases}$$

Since f is the product of rational functions whose denominators vanish only at $(0, 0)$, f itself is continuous away from $(0, 0)$. Hence the slice of f along any line that does not pass through the origin is continuous. We will now show that the slice of f along **every** line is continuous.

By what we have said above, it remains to consider lines through the origin. Notice that f is identically equal to zero along the y -axis. (There is a factor of x in the numerator.) Constant functions are certainly continuous, so this slice of f is continuous.

Now consider any other line through the origin. This has the equation $y = ax$ for some $a \in \mathbb{R}$, $a \neq 0$. On the line $y = ax$, for $x \neq 0$,

$$f(x, ax) = \frac{1}{x^2 + a^2 x^2} \frac{2x^5 a}{x^8 + a^2 x^2} = \frac{1}{1 + a^2} \frac{2xa}{x^6 + a^2}$$

which is a continuous – and even differentiable – function of x . Therefore, for each non-zero $a \in \mathbb{R}$, the function $g(t)$ defined by

$$g(t) := f(\mathbf{0} + t(1, a))$$

is continuous. Hence, the slice of f along every line through $\mathbf{0}$ is continuous.

However, f itself is **not** continuous. To see this, consider the sequence $\{\mathbf{x}_n\}$ given by $\mathbf{x}_n := (1/n, 1/n^4)$. Evidently, $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{0}$, while for each n ,

$$f(\mathbf{x}_n) = \frac{1}{n^{-2} + n^{-8}} \frac{2n^{-8}}{2n^{-8}} = \frac{n^2}{1 + n^{-6}}.$$

Hence $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = \infty$, and so not only is f discontinuous, there are points arbitrarily close to the origin at which f takes on arbitrarily large positive values.

If you consider the sequence $\mathbf{x}_n := (1/n, -1/n^4)$, you can see that there are points arbitrarily close to the origin at which f takes on arbitrarily large negative values.

Yet as you walk along the slice over any straight line through the origin, you never encounter any jumps or singularities of any sort.

At this point you might wonder how wise we have been in choosing our definition of continuity. We are excluding functions f that vary continuously on the line segment connecting any two points from the class of continuous functions. Are we not being too restrictive?

Let us consider an alternate definition:

Definition 35 (Separate continuity). A function $f(x, y)$ on \mathbb{R}^2 is **separately continuous** in case for each x_0 , the function $y \mapsto f(x_0, y)$ is a continuous function of y , and if for y_0 , the function $x \mapsto f(x, y_0)$ is a continuous function of x .

The function f in Example 50 is separately continuous. Moreover, separate continuity is easier to check than continuity – it can be done one variable at a time. Could it be that the n dimensional analog of this is actually a better generalization of continuity to several variables? Unfortunately, no.

Mathematical definitions are made the way they are because of what can be done with them. They are discarded unless they capture concepts that are useful in problem solving.

Part of the problem-solving value of the concept of continuity lies in its relevance to *minimum–maximum problems*. You know from the theory of functions of a single variable that if g is any *continuous* function of x , then g *attains its maximum* on any closed interval $[a, b]$. That is, there is a point x_0 with $a \leq x_0 \leq b$ so that for every x with $a \leq x \leq b$,

$$g(x) \leq g(x_0) .$$

In this case, we say that x_0 is a *maximizer* of g on $[a, b]$. Finding maximizers is one of the important applications of the differential calculus.

In the next section, we show that continuity is the right hypothesis for proving a multi-variable version of this important theorem, and that separate continuity is not enough. Separate continuity is easier to check, but alas, it is just not that useful.

3.2 Continuity, compactness and maximizers

3.2.1 Open and closed sets in \mathbb{R}^n

In this subsection we introduce the class of subsets of \mathbb{R}^n that generalizes the class of bounded, closed intervals in \mathbb{R} .

Definition 36 (Open ball of radius r , and bounded sets). *For each $\mathbf{x} \in \mathbb{R}^n$, and each $r > 0$, the open ball of radius r about \mathbf{x} is the subset $B_r(\mathbf{x}_0)$ of \mathbb{R}^n defined consisting of all $\mathbf{y} \in \mathbb{R}^n$ such that $\|\mathbf{x} - \mathbf{y}\| < r$. That is,*

$$B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{y}\| < r\} .$$

A set $A \subset \mathbb{R}^n$ is bounded in case $A \subset B_r(\mathbf{0})$ for some $r > 0$.

Definition 37 (Open and closed sets). *A set $C \subset \mathbb{R}^n$ is closed in case whenever $\{\mathbf{x}_k\}$ is a sequence of points belonging to C , and the limit $\mathbf{z} = \lim_{k \rightarrow \infty} \mathbf{x}_k$ exists, then $\mathbf{z} \in C$. A set $U \subset \mathbb{R}^n$ is open in case whenever $\mathbf{x} \in U$, there is an $r > 0$ so that $B_r(\mathbf{x}) \subset U$. The empty set \emptyset is defined to be both open and closed.*

It is a good exercise to use the triangle inequality to show that for each $\mathbf{x} \in \mathbb{R}^n$ and each $r > 0$, $B_r(\mathbf{x})$ is open. You may be wondering how one could every prove a set with infinitely many points is closed. Are there not infinitely many sequences to be dealt with in the proof? Yes, and so we had better do that implicitly. We shall prove a theorem that is useful in this regard. First, we state a useful definition.

Definition 38 (Level sets). *Given any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, and any $a \in \mathbb{R}$, the set consisting of all solutions of the equation $f(\mathbf{x}) = a$ is called the level set of f at height a . It is written as*

$$\{ \mathbf{x} : f(\mathbf{x}) = a \}.$$

Likewise, the sub-level set of f at height a and the super-level set of f at height a , respectively, are the sets

$$\{ \mathbf{x} : f(\mathbf{x}) \leq a \} \quad \text{and} \quad \{ \mathbf{x} : f(\mathbf{x}) \geq a \}.$$

Example 51 (Spheres as level sets). *Consider the function $f(\mathbf{x}) = \|\mathbf{x}\|^2 = \sum_{j=1}^n x_j^2$. For each $r > 0$, the sphere of radius r is the level set of f through r^2 .*

Theorem 36 (Continuity and closed level sets). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous. Then for each $a \in \mathbb{R}$, the level set of f through a is closed, as are the sub-level set of f through a and the super-level set of f through a .*

Proof. Consider the level set $\{ \mathbf{x} : f(\mathbf{x}) = a \}$. If this is empty, we are finished, since the empty set is closed by definition. Otherwise, consider any convergent sequence $\{\mathbf{x}_k\}$ such that each x_k belongs to the level set. Since $\{\mathbf{x}_k\}$ is convergent, there exists $\mathbf{z} \in \mathbb{R}^n$ with $\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{z}$. Since f is continuous, we have

$$\lim_{k \rightarrow \infty} f(\mathbf{x}_k) = f(\mathbf{z}).$$

But since each \mathbf{x}_k lies in the level set, $f(\mathbf{x}_k) = a$ for all k . Thus $f(\mathbf{z}) = a$, and hence the limit point \mathbf{z} also belongs to the level set. Since the convergent sequence was an arbitrary convergent sequence in the level set, this proves that the level set is closed.

The proofs for sub-level sets and super-level sets are very much the same, and are left to the reader. \square

Example 52 (Spheres are closed). *The function $f(\mathbf{x}) = \|\mathbf{x}\|^2 = \sum_{j=1}^n x_j^2$ is continuous. Indeed, by Example 40, each of the coordinate functions; i.e., the functions sending \mathbf{x} to x_j , $j = 1, \dots, n$ are continuous. Then by Theorem 34, the function sending \mathbf{x} to $\sum_{j=1}^n x_j^2$ is continuous. Thus, by Theorem 36, for each $r > 0$, the sphere of radius r is closed.*

We conclude this subsection with one more relation between open and closed sets that shall be useful to us later on.

Definition 39 (Complementary sets). *For any set $A \subset \mathbb{R}^n$, the complement of A , A^c , is the subset of all $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} \notin A$.*

Theorem 37 (Complementarity of open and closed sets). (1) *Let $U \subset \mathbb{R}^n$ be open. Then U^c , the complement of U , is closed. (2) Let $C \subset \mathbb{R}^n$ be closed. Then C^c , the complement of C , is open.*

Proof. Let $U \subset \mathbb{R}^n$ be open, and consider any convergent sequence $\{\mathbf{x}_k\}$ in U^c . We must show that $\mathbf{z} = \lim_{k \rightarrow \infty} \mathbf{x}_k \in U^c$.

Suppose not. Then $\mathbf{z} = \lim_{k \rightarrow \infty} \mathbf{x}_k \in U$. Since U is open, for some $r > 0$, $B_r(\mathbf{z}) \subset U$. But since $\mathbf{z} = \lim_{k \rightarrow \infty} \mathbf{x}_k$, for all sufficiently large k , $\|\mathbf{x}_k - \mathbf{z}\| < r$, and this means that all such \mathbf{x}_k belong to U . But this is impossible, since each $\mathbf{x}_k \in U^c$. Hence, $\mathbf{z} = \lim_{k \rightarrow \infty} \mathbf{x}_k \in U^c$.

Next, let $C \subset \mathbb{R}^n$ be closed. We must show that for each $\mathbf{x} \in C^c$, there is some $r > 0$, $B_r(\mathbf{x}) \subset C^c$. Suppose not. Then for each $k \in \mathbb{N}$, there is some $\mathbf{x}_k \in B_{1/k}(\mathbf{x}) \cap C$. By construction, for all $\epsilon > 0$

$$k > 1/\epsilon \quad \Rightarrow \quad \|\mathbf{x}_k - \mathbf{x}\| < \epsilon.$$

Thus, $\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}$. Since C is closed and since each \mathbf{x}_k belongs to C , it would have to be the case that $\mathbf{x} \in C$. However, this is impossible since $\mathbf{x} \in C^c$. Therefore, there is some $r > 0$ with $B_r(\mathbf{x}) \subset C^c$. \square

3.2.2 Minimizers and maximizers

Definition 40 (Minimizers and maximizers). *Let f be a function defined on a closed set $C \subset \mathbb{R}^n$ with values in \mathbb{R} . Then $\mathbf{x}_0 \in C$ is a maximizer of f in C if and only if*

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) \quad \text{for all } \mathbf{x} \in C. \quad (3.14)$$

Likewise, $\mathbf{x}_0 \in C$ is a minimizer of f in C if and only if

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) \quad \text{for all } \mathbf{x} \in C. \quad (3.15)$$

Notice that \mathbf{x}_0 is a minimizer of f in C if and only if \mathbf{x}_0 is a maximizer of $-f$ in C . Therefore, it suffices to prove theorems on the existence of maximizers: Each such theorem implies a corresponding theorem for minimizers.

We shall show in the next subsection that every continuous real valued function defined on a bounded closed set $C \subset \mathbb{R}^n$ has a maximizer in C . First, we show by example that this is not true if one replaces continuity by separate continuity.

Example 53 (Separately continuous, but no maximizer). *Consider the function f from Example 50. As we have seen there, this function is separately continuous, and even more, its slice along every line is continuous. Let $C = \{(x, y) \mid x^2 + y^2 \leq 1\}$ which is plainly bounded, and is closed by Theorem 36.*

We will now show that f has neither a maximizer nor a minimizer on C , and in fact, is unbounded above and below on C , all despite being separately continuous. Indeed, we have already done the work. We have seen in Example 50 that

$$f(1/n, 1/n^4) = \frac{n^2}{2} \quad \text{and} \quad f(1/n, -1/n^4) = -\frac{n^2}{2}.$$

For $n > 1$ both $(1/n, 1/n^4)$ and $(1/n, -1/n^4)$ belong to C . Hence f is unbounded above and below on C .

However, for continuous functions of several variables, there is an analog of the single variable theorem. This alone makes continuity a much more useful concept than separate continuity.

3.2.3 Compactness and existence of maximizers

Definition 41 (Compact sets). *A set $C \subset \mathbb{R}^n$ is compact in case it is closed and bounded.*

The key to the existence of maximizers, and many other problems as well, is the following theorem which says that a set is compact if and only if every infinite sequence of points in the set has a subsequence that converges to a point in C . It is one of the fundamental theorems of mathematical analysis. The theorem is powerful because often it is easy to check that a set is compact, and then you know something significant about *every* infinite sequence in that set. The very interesting proof relies on the “pigeonhole principle” and the completeness of the real numbers.

Theorem 38 (The Bolzano-Weierstrass Theorem). *Let C be a compact subset of \mathbb{R}^n . Then for every sequence $\{\mathbf{x}_n\}$ in C , there is a subsequence $\{\mathbf{x}_{n_k}\}$ and a $\mathbf{z} \in C$ such that*

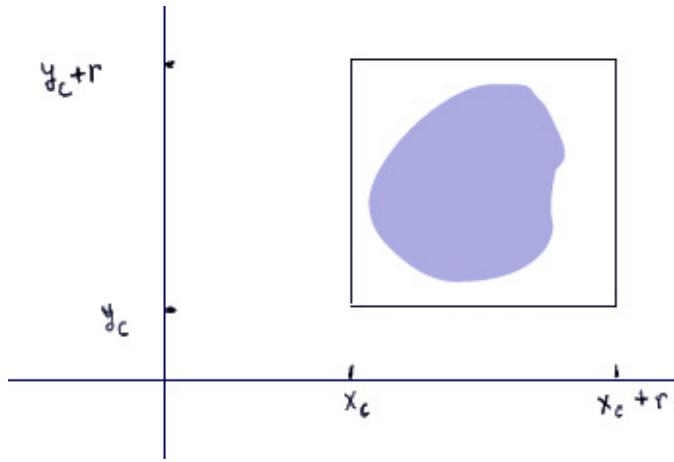
$$\lim_{k \rightarrow \infty} \mathbf{x}_{n_k} = \mathbf{z} .$$

On the other hand, if C is not compact, then there exists a sequence $\{\mathbf{x}_n\}$ in C that has no subsequence that converges to an element \mathbf{z} of C .

Proof. We will prove this for $n = 2$ so that we can draw pictures. Once you understand the idea for $n = 2$, you will see that it applies for all n .

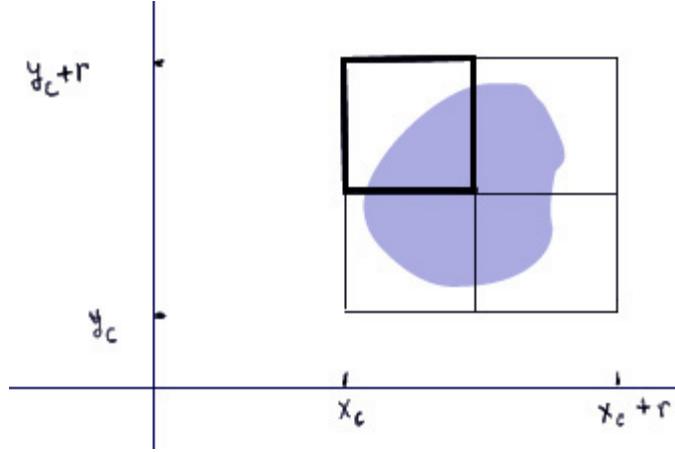
Since C is a closed and bounded set, there are numbers x_c, y_c and r so that C is contained in the square

$$x_c \leq x \leq x_c + r \quad \text{and} \quad y_c \leq y \leq y_c + r .$$



The shaded region is the closed, bounded set C .

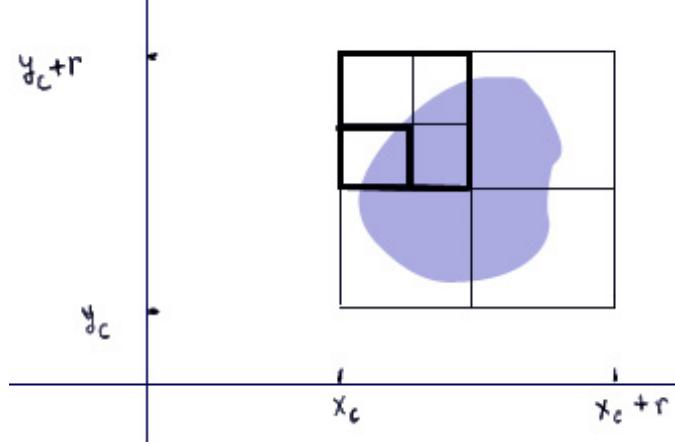
Now consider any infinite sequence $\{\mathbf{x}_n\}$ of points in C . To obtain a convergent subsequence, first divide the square into four congruent squares. Since the four squares cover C , by the pigeonhole principle, at least one of the four squares must be such that it contains infinitely many terms of the sequences $\{\mathbf{x}_n\}$.



Here, the upper left square contained infinitely many terms, and we chose it.

Now define \mathbf{x}_{n_1} to be the first term in $\{\mathbf{x}_n\}$ that belongs to the chosen square.

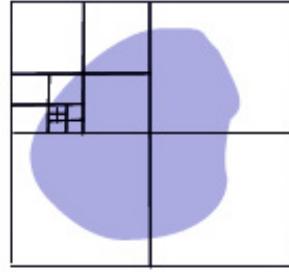
Next, subdivide the square in the first step into four smaller squares as in the diagram below. Again, by the pigeonhole principle, one of these must be such that it contains infinitely many terms of the sequence $\{\mathbf{x}_n\}$. Choose such a square.



Here we chose the lower left square in the previously selected square.

Now define \mathbf{x}_{n_2} to be the first term in $\{\mathbf{x}_n\}$ after \mathbf{x}_{n_1} that belongs to the chosen square.

Iterating this procedure produces a sequence of points $\{\mathbf{x}_{n_k}\}$ such that for each $m > 0$, the sequence $\{\mathbf{x}_{n_k} : k \geq m\}$ lies in a square of side length $r2^{-m}$. This follows from the fact that the procedure described above produces a nested set of squares.



Since the side length is reduced by a factor of 2 with each subdivision, and since it starts at $r/2$, at the m th stage we have a square of side length $2^{-m}r$. Thus, for all $k, \ell \geq m$, \mathbf{x}_{n_k} and \mathbf{x}_{n_ℓ} belong to a square of side length $2^{-m}r$. How far apart can they possibly be? No further than the length of the diagonal of the square, which is $\sqrt{2}$ times $2^{-m}r$; i.e., $2^{1/2-m}r$. Thus, for any $\epsilon > 0$, choose m so that $2^{1/2-m}r < \epsilon$, and then for this m ,

$$k, \ell \geq m \quad \Rightarrow \quad \| -\mathbf{x}_{n_k} - \mathbf{x}_{n_\ell} \| < \epsilon .$$

Since for any coordinate index j , $|(\mathbf{x}_\ell)_j - (\mathbf{x}_m)_j| \leq \|\mathbf{x}_\ell - \mathbf{x}_m\|$,

$$k, \ell \geq m \quad \Rightarrow \quad |(\mathbf{x}_{n_k})_j - (\mathbf{x}_{n_\ell})_j| < \epsilon .$$

This means that $\{(\mathbf{x}_{n_k})_j\}$, where j is fixed and k indexes the terms of the sequence, is a Cauchy sequence. By the completeness property of the real numbers, or what is the same thing, the construction of the real numbers out of the rational numbers, every Cauchy sequence has a limit, and so there exists a number $z_j \in \mathbb{R}$ so that

$$\lim_{k \rightarrow \infty} (\mathbf{x}_{n_k})_j = z_j . \tag{3.16}$$

Now define a vector \mathbf{z} by $(\mathbf{z})_j = z_j$ for $j = 1, \dots, n$. Then (3.16) implies

$$\lim_{k \rightarrow \infty} \mathbf{x}_{n_k} = \mathbf{z} .$$

Then, since C is closed, and $\{\mathbf{x}_{n_k}\}$ is in C , \mathbf{z} belongs to C .

For the final part, suppose C is not compact. Then either C is not closed, or not bounded, or both. If it is not closed, there is some sequence $\{\mathbf{x}_n\}$ in C that converges to some $\mathbf{z} \notin C$. In this case, every subsequence of $\{\mathbf{x}_n\}$ converges to $\mathbf{z} \notin C$, and so no subsequence of $\{\mathbf{x}_n\}$ can converge to an element of C . Likewise if C is unbounded, there is a sequence $\{\mathbf{x}_n\}$ in C with $\|\mathbf{x}_n\| \geq n$ for all n , and evidently no subsequence of this sequence converges at all. Thus, when C is not compact, there are sequences in C for which no subsequence converges to an element of C .

□

Our first application of the Bolzano-Weierstrass Theorem is to the existence of maximizers.

Theorem 39 (Continuity and Maximizers). *Let f be a continuous function defined on a compact set $C \subset \mathbb{R}^n$ with values in \mathbb{R} . Then there is point \mathbf{z} in C so that*

$$f(\mathbf{x}) \leq f(\mathbf{z}) \quad \text{for all } \mathbf{x} \in C . \tag{3.17}$$

Proof. Let B be the least upper bound of f on C . That is, B is either infinity if f is unbounded on C , or else it is the least number that is greater than or equal to $f(\mathbf{x})$ for all $\mathbf{x} \in C$. We aim to show that B is finite, and that there is an $\mathbf{z} \in C$ with $f(\mathbf{z}) = B$. Then \mathbf{z} is the maximizer we seek.

First, suppose that $B = \infty$. Then for each n , there is an $\mathbf{x}_n \in C$ such that $f(\mathbf{x}_n) \geq n$. By Theorem 38, there is a convergent subsequence $\{\mathbf{x}_{n_k}\}$ with limit $\mathbf{z} \in C$. Since f is continuous,

$$f(\mathbf{z}) = \lim_{k \rightarrow \infty} f(\mathbf{x}_{n_k}).$$

However, since $f(\mathbf{x}_{n_k}) \geq n_k \geq k$ for all k , $\lim_{k \rightarrow \infty} f(\mathbf{x}_{n_k})$ does not exist. This contradiction shows that B must be finite.

Now, by the definition of the Least Upper Bound, for each n , the set

$$\{ \mathbf{x} \in C : f(\mathbf{x}) \geq B - 1/n \}$$

is not empty. Therefore, for each n we may choose \mathbf{x}_n in this set. By the Squeeze Principle, since $B - 1/n \leq f(\mathbf{x}_n) \leq B$, $\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = B$.

By Theorem 38, the sequence $\{\mathbf{x}_n\}$ has a convergent subsequence $\{\mathbf{x}_{n_k}\}$ with limit $\mathbf{z} \in C$. Since f is continuous, $f(\mathbf{z}) = \lim_{k \rightarrow \infty} f(\mathbf{x}_{n_k}) = B$. Thus, $\mathbf{z} \in C$ and $f(\mathbf{z}) = B \geq f(\mathbf{x})$ for all $\mathbf{x} \in C$. \square

The Bolzano-Weierstrass Theorem has many consequences, as we shall see throughout this course. We close this section by explaining one concerning *isometries*:

Definition 42. Let \mathbf{f} be a function from $U \subset \mathbb{R}^n$ to \mathbb{R}^m with the property that for all $\mathbf{x}, \mathbf{y} \in U$,

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|.$$

Then \mathbf{f} is called an *isometry*. That is \mathbf{f} is an isometry in case it preserves the distances between points; the distance between the images equals the distance between the original points.

Observe that isometries are always continuous. It is worth writing down a proof with explicit reference to the definitions involved.

Theorem 40 (Isometries on compact sets are invertible). Let C be a compact set in \mathbb{R}^n . Let \mathbf{f} be an isometry defined on C with values in C . Then \mathbf{f} is an invertible function from C onto C .

Proof. First, \mathbf{f} is one-to-one: If for some $\mathbf{x}, \mathbf{y} \in C$, $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{y})$, then

$$0 = \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|,$$

and so $\mathbf{y} = \mathbf{x}$.

Showing that for each $\mathbf{y} \in C$ there is an $\mathbf{x} \in C$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ takes more effort; we shall use Theorems 38 and 39 for this.

Suppose there is some $\mathbf{y} \in C$ such that $\mathbf{y} \neq \mathbf{f}(\mathbf{x})$ for any $\mathbf{x} \in C$. The function

$$g(\mathbf{x}) := \|\mathbf{f}(\mathbf{x}) - \mathbf{y}\|$$

is a continuous function on C , being the composition of continuous functions. Thus, by Theorem 39, there is a $\mathbf{z} \in C$ with $g(\mathbf{z}) \leq g(\mathbf{x})$ for all $\mathbf{x} \in C$. Since $\mathbf{y} \neq \mathbf{f}(\mathbf{z})$, $g(\mathbf{z}) := r > 0$. and thus there is some $r > 0$ such that

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{y}\| \geq r \quad (3.18)$$

for all $\mathbf{x} \in C$. Now, inductively define the sequence $\{\mathbf{x}_n\}$ by defining $\mathbf{x}_1 = \mathbf{y}$, and then

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n)$$

for all $n \geq 1$.

By Theorem 38, there exists a subsequence $\{\mathbf{x}_{n_k}\}$ converging to some $\mathbf{z} \in C$. Since every convergent sequence is a Cauchy sequence, there are k and ℓ so that $k < \ell$ and

$$\|\mathbf{x}_{n_k} - \mathbf{x}_{n_\ell}\| < r/2 .$$

But with $\mathbf{f}^{(m)}$ denoting the m th composition power of \mathbf{f} ,

$$\mathbf{x}_{n_k} = \mathbf{f}^{(n_k)}(\mathbf{y}) \quad \text{and} \quad \mathbf{x}_{n_\ell} = \mathbf{f}^{(n_\ell)}(\mathbf{y}) .$$

By the isometry property,

$$r/2 \geq \|\mathbf{x}_{n_k} - \mathbf{x}_{n_\ell}\| = \|\mathbf{f}^{(n_k)}(\mathbf{y}) - \mathbf{f}^{(n_\ell)}(\mathbf{y})\| = \|\mathbf{y} - \mathbf{f}^{(n_\ell - n_k)}(\mathbf{y})\| .$$

Now define $\mathbf{x} := \mathbf{f}^{(n_\ell - n_k - 1)}(\mathbf{y})$ so that

$$\mathbf{f}^{(n_\ell - n_k)}(\mathbf{y}) = \mathbf{f}(\mathbf{x}) .$$

Then $\|\mathbf{y} - \mathbf{f}(\mathbf{x})\| < r/2$ which contradicts (3.18). This contradiction shows that there cannot exist any $\mathbf{y} \in C$ such that $\mathbf{y} \neq \mathbf{f}(\mathbf{x})$ for all $\mathbf{x} \in C$. Hence, \mathbf{f} maps C onto C . Since we have already seen that \mathbf{f} is one-to-one, this proves \mathbf{f} is invertible. \square

We can now use Theorem 40 to give a short proof of the fundamental Lemma 1, which says that there does not exist any set of $n+1$ orthonormal vectors in \mathbb{R}^n . Indeed, the unit sphere S , consisting of all unit vectors in \mathbb{R}^n is compact, as we have seen in this section.

Second Proof of Lemma 1. Given any set $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of n orthonormal vectors in \mathbb{R}^n , define a function \mathbf{f} from \mathbb{R}^n to \mathbb{R}^n by

$$\mathbf{f}(x_1, \dots, x_n) = \sum_{j=1}^n x_j \mathbf{u}_j .$$

Then, since $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is orthonormal,

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|^2 = \left\| \sum_{j=1}^n (x_j - y_j) \mathbf{u}_j \right\|^2 = \sum_{j=1}^n |x_j - y_j|^2 = \|\mathbf{x} - \mathbf{y}\|^2 .$$

Thus, \mathbf{f} is an isometry. Likewise, we see $\|\mathbf{f}(\mathbf{x})\| = \|\mathbf{x}\|$ so that if $\mathbf{x} \in S$, then also $\mathbf{f}(\mathbf{x}) \in S$. Thus, the restriction of \mathbf{f} to S is an isometry from S into S .

Now consider any other unit vector \mathbf{u}_{n+1} . By Theorem 40, \mathbf{f} not only maps S into S , it maps S onto S , and since $\mathbf{u}_{n+1} \in S$, there is an $\mathbf{x} = (x_1, \dots, x_n) \in S$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{u}_{n+1}$; i.e.,

$$\mathbf{u}_{n+1} = \sum_{j=1}^n x_j \mathbf{u}_j .$$

Taking the dot product of both sides with \mathbf{u}_{n+1} , $1 = \mathbf{u}_{n+1} \cdot \mathbf{u}_{n+1} = \sum_{j=1}^n x_j (\mathbf{u}_{n+1} \cdot \mathbf{u}_j)$. If \mathbf{u}_{n+1} were orthogonal to each \mathbf{u}_j , the right hand side would be zero, and this is impossible. Hence there does not exist any set of $n + 1$ orthonormal vectors in \mathbb{R}^n . \square

There is a strong contrast between the two proofs we have given of Lemma 1. The first proof involved the explicit construction of a product of Householder reflections taking $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ into $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. The second proof avoids this, and is much shorter – but this is because it makes use of the Bolzano-Weierstrass Theorem. Both kinds of proof have their place. Sometimes an explicit construction will not be possible, and an argument making use of the Bolzano-Weierstrass Theorem is the only way to proceed. However, when an explicit construction is possible, it is often worthwhile to make it: The explicit construction used in our first proof of Lemma 1 is what enabled us to prove that an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of \mathbb{R}^3 is right handed if and only if it is related to $\{v\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ by a rotation.

3.2.4 The Squeeze Principle revisited

Let f be a function on \mathbb{R}^n that is continuous away from some point \mathbf{x}_0 , and suppose we wish to determine continuity at \mathbf{x}_0 . We have encountered this situation many times. Fix $r > 0$, and for all unit vectors $\mathbf{u} \in \mathbb{R}^n$ define a function $h_r(\mathbf{u}) = |f(\mathbf{x}_0 + r\mathbf{u}) - f(\mathbf{x}_0)|$. Then h_r is continuous on the unit sphere S^{n-1} in \mathbb{R}^n , which is closed and bounded. Hence there exists a maximizer \mathbf{u}_* (depending on r). Let $g(r)$ be this maximal value $h_r(\mathbf{u}_*) = |f(\mathbf{x}_0 + r\mathbf{u}_*) - f(\mathbf{x}_0)|$. This gives us a well defined function. We are using the existence of maximizers to define it.

We claim that f is continuous at \mathbf{x}_0 if and only if $\lim_{r \rightarrow 0} g(r) = 0$. To see this we define $r_n = \|\mathbf{x}_n - \mathbf{x}_0\|$ and $\mathbf{u}_n = r_n^{-1}(\mathbf{x}_n - \mathbf{x}_0)$,

$$0 \leq |f(\mathbf{x}_n) - f(\mathbf{x}_0)| = |f(\mathbf{x}_0 + r_n \mathbf{u}_n) - f(\mathbf{x}_0)| \leq g(r_n) .$$

If $\lim_{r \rightarrow 0} g(r) = 0$, then by the Squeeze Principle for sequences of real numbers, $\lim_{n \rightarrow \infty} |f(\mathbf{x}_n) - f(\mathbf{x}_0)| = 0$.

On the other hand, if it is not the case that $\lim_{r \rightarrow 0} g(r) = 0$, there is some sequence $\{r_n\}$ of positive numbers convergent to zero such that $\lim_{n \rightarrow \infty} g(r_n) > 0$. Let \mathbf{u}_n be the maximizer of h_{r_n} considered above. Then

$$g(r_n) = |f(\mathbf{x}_0 + r_n \mathbf{u}_n) - f(\mathbf{x}_0)| ,$$

and so

$$\lim_{n \rightarrow \infty} |f(\mathbf{x}_0 + r_n \mathbf{u}_n) - f(\mathbf{x}_0)| > 0 .$$

This shows that f is discontinuous at \mathbf{x}_0 .

In other words, whenever f is continuous at \mathbf{x}_0 , there is an appropriate functions g such that the Squeeze Principle may be applied to prove the continuity of f , namely the function $g(r)$ defined here. We could not introduce this function when we proved our theorem on the Squeeze Principle since it relies on Theorem 39 on the existence of maximizers.

3.3 Exercises

3.1 Complete the proof of Theorem 34.

3.2 A function \mathbf{f} defined domain $U \subset \mathbb{R}^n$ with values in \mathbb{R} is called a *Lipschitz continuous* function in case there is some number M so that

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\| \quad (3.19)$$

for all \mathbf{x} and \mathbf{y} in U .

- (a) Show that a Lipschitz continuous function is continuous by finding a valid margin of error on the input; i.e., a valid $\delta(\epsilon)$ that has a very simple form: *proportional to ϵ* .
- (b) For $R > 0$, let U denote the ball of radius R about the origin; i.e., $U = B_R(\mathbf{0})$. Let $f(\mathbf{x})$ be defined on U by $f(\mathbf{x}) = \|\mathbf{x}\|^2$. Using the identity

$$\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y})$$

and the Cauchy-Schwarz inequality, show that f is Lipschitz on U with Lipschitz constant $2R$.

- (c) Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ have the form $\mathbf{f}(\mathbf{x}) = (\mathbf{a}_1 \cdot \mathbf{x}, \dots, \mathbf{a}_m \cdot \mathbf{x})$ for some set of vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$ in \mathbb{R}^n , as in Example 41. Show that \mathbf{f} is Lipschitz continuous on \mathbb{R}^n .

3.3 Consider the function f defined by

$$f(x_1, x_2) = \sin(x_1) \cos(x_2) .$$

Note that

$$f(x_1, x_2) - f(y_1, y_2) = (\sin(x_1) - \sin(y_1)) \cos(x_2) + \sin(y_1)(\cos(x_2) - \cos(y_2)) \quad (3.20)$$

Show that

$$|\sin(x_1) - \sin(y_1)| \leq |x_1 - y_1| \quad \text{and} \quad |\cos(x_2) - \cos(y_2)| \leq |x_2 - y_2| .$$

(This is a single variable problem, and the fundamental theorem of calculus can be applied). Combine this with the identity (3.20) to show that f satisfies (3.19) with $M = \sqrt{2}$.

3.4 Let $f(x, y)$ be given by

$$f(x, y) = \begin{cases} \frac{x^2 \sin(xy)}{x^6 + y^2} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0) \end{cases} .$$

(a) For any $a, b \in \mathbb{R}$, define the sequence $\{\mathbf{x}_n\}$ by $\mathbf{x}_n = (a/n, b/n)$. Compute $\lim_{n \rightarrow \infty} f(\mathbf{x}_n)$.

(b) For any $a, b \in \mathbb{R}$, define the sequence $\{\mathbf{x}_n\}$ by $\mathbf{x}_n = (a/n, b/n^3)$. Compute $\lim_{n \rightarrow \infty} f(\mathbf{x}_n)$.

(c) Is the function f continuous? Justify your answer.

3.5 Consider the function defined by

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

Is this function is continuous? Justify your answer.

3.6 Consider the function defined by

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

Is this function is continuous? Justify your answer. Hint: You make a ratio larger and simpler at the same time if you discard a positive term for the denominator.

3.7 Consider the function defined by

$$f(x, y) = \begin{cases} |x|^r \sin(y/x) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

where $r > 0$. For which values of r , if any, is f continuous?

3.8 Let \mathbf{a} and \mathbf{b} be given vectors in \mathbb{R}^3 such that neither is a multiple of the other. Define a function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ by

$$f(\mathbf{x}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{x}).$$

Define

$$\mathbf{x}_0 = \frac{1}{\|\mathbf{a} \times \mathbf{b}\|} \mathbf{a} \times \mathbf{b}.$$

Show that

$$f(\mathbf{x}) \leq f(\mathbf{x}_0)$$

for all unit vectors $\mathbf{x} \in \mathbb{R}^3$. In other words, show that \mathbf{x}_0 is the maximizer of f on the unit sphere in \mathbb{R}^3 .

3.9 Given m vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$ in \mathbb{R}^n , define the function f from \mathbb{R}^n to \mathbb{R} by

$$f(\mathbf{x}) = \sum_{j=1}^m (\mathbf{a}_j \cdot \mathbf{x})^2.$$

Show that f has both a maximizer and a minimizer on the closed unit ball

$\overline{B} = \{\mathbf{x} : \sum_{j=1}^n x_j^2 \leq 1\}$, but has only a minimizer, and no maximizer, on the open unit ball $B = \{\mathbf{x} : \sum_{j=1}^n x_j^2 < 1\}$. Hint: Show that the maximizer on \overline{B} lies on the boundary of \overline{B} .

3.10 Let \mathbf{f} be any function from \mathbb{R}^n to \mathbb{R}^m . For any set $A \subset \mathbb{R}^m$, define $f^{-1}(A)$ to be the set of all points \mathbf{x} , if any, in \mathbb{R}^n such that $\mathbf{f}(\mathbf{x}) \in A$. The set $f^{-1}(A)$, which may be the empty set, is called

the *preimage of A under f*. Do not be misled by the notation: $f^{-1}(A)$ is defined whether or not the function \mathbf{f} itself is invertible.

(a) Prove that \mathbf{f} is continuous if and only if whenever A is an open set in \mathbb{R}^m , then $f^{-1}(A)$ is an open set in \mathbb{R}^n . This result provides a way to talk about continuity without explicitly bringing ϵ and δ into the discussion. It also has other uses.

(b) Use the result of part **(a)** to give a short proof that whenever \mathbf{f} is a continuous function from \mathbb{R}^n to \mathbb{R}^m , and \mathbf{g} is a continuous function from \mathbb{R}^m to \mathbb{R}^ℓ , then $\mathbf{g} \circ \mathbf{f}$ is a continuous function from \mathbb{R}^n to \mathbb{R}^ℓ .

3.11 Let $K \subset \mathbb{R}^n$ be compact. Show that for each $\mathbf{x} \in \mathbb{R}^n$, there exists at least one point $\mathbf{y}_K(\mathbf{x}) \in K$ (depending on \mathbf{x} , as indicated in the notation) such that

$$\|\mathbf{x} - \mathbf{y}_K(\mathbf{x})\| \leq \|\mathbf{x} - \mathbf{z}\|$$

for all $\mathbf{z} \in K$. Define the *distance from \mathbf{x} to K* to be the number $\|\mathbf{x} - \mathbf{y}_K(\mathbf{x})\|$ for this choice of \mathbf{y} .

The function

$$d_K(\mathbf{x}) := \|\mathbf{x} - \mathbf{y}_K(\mathbf{x})\|$$

then gives the distance from \mathbf{x} to K . Show that the function d_K is a continuous function on \mathbb{R}^n . Is it Lipschitz continuous?

3.12 Let $K \subset \mathbb{R}^n$ be compact, and let \mathbf{f} be a continuous function from \mathbb{R}^n to \mathbb{R}^m . Define $L \subset \mathbb{R}^m$ by

$$L := \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{f}(\mathbf{x}) \text{ for some } \mathbf{x} \in K\}.$$

Is L necessarily compact? Justify your answer.

Chapter 4

DIFFERENTIABLE FUNCTIONS

4.1 Vertical slices and directional derivatives

4.1.1 Directional derivatives and partial derivatives

We now pick up with the “slicing” idea introduced in Chapter 3, and try to take derivatives along slices of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in order to understand the nature of the graph of $x_{n+1} = f(\mathbf{x})$ in \mathbb{R}^{n+1} . Of course, we can only actually plot the graph for $n \leq 2$, so we will be especially interested in the case $n = 2$ at the beginning.

Definition 43 (Directional derivatives). *Given a function $f(\mathbf{x})$ defined in an open subset U of \mathbb{R}^n , and some point $\mathbf{x}_0 \in \mathbb{R}^n$, and also a non zero vector \mathbf{v} in \mathbb{R}^n , the directional derivative of f at \mathbf{x}_0 in the direction \mathbf{v} is defined by*

$$\lim_{h \rightarrow 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h}, \quad (4.1)$$

provided this limit exists. If the limit does not exist, the directional derivative does not exist.

Given f , \mathbf{x}_0 and \mathbf{v} , the function

$$g(t) = f(\mathbf{x}_0 + t\mathbf{v}) \quad (4.2)$$

represents the “slice” of f over the line $\mathbf{x}_0 + t\mathbf{v}$ in \mathbb{R}^n . Then the directional derivative of f at \mathbf{x}_0 in the direction \mathbf{v} is just $g'(0)$. This means that directional derivatives can be computed by familiar single variable methods.

Example 54 (Slicing a function along a line). *For example, if $f(x, y) = \frac{xy^2}{1+x^2+y^2}$, $\mathbf{x}_0 = (1, 1)$ and $\mathbf{v} = (1, 2)$, then*

$$g(t) = f(1+t, 1+2t) = \frac{(1+t)(1+2t)^2}{1+(1+t)^2+(1+2t)^2} = \frac{1+5t+8t^2+4t^3}{3+6t+5t^2}.$$

The result is a familiar garden variety function of a single variable t . It is a laborious but straightforward task to now compute that $g'(0) = 1$. Please do the calculation; you will then appreciate the better way of computing directional derivatives that we shall soon explain!

Directional derivatives may exist for some directions, but not others:

Example 55 (Sometimes there are directional derivatives only in certain directions). *Let f be the function defined by*

$$f(x, y) = \begin{cases} \frac{x^2 - y^2}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}. \quad (4.3)$$

Let $\mathbf{x}_0 = (0, 0)$ and let $\mathbf{v} = (a, b)$ for some numbers a and b . The question we now ask is: For which values of a and b does there exists the directional derivative of f at \mathbf{x}_0 in direction \mathbf{v} ?

To answer this, let's compute $f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)$, divide by h , and try to take the limit $h \rightarrow 0$. We find that

$$f(\mathbf{x}_0) = 0 \quad \text{and} \quad f(\mathbf{x}_0 + h\mathbf{v}) = \frac{a^2 - b^2}{a^2 + b^2}.$$

(For $\mathbf{v} \neq \mathbf{0}$, as in Definition 43, we do not divide by zero on the right.) Therefore

$$\frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h} = \frac{1}{h} \left(\frac{a^2 - b^2}{a^2 + b^2} \right).$$

As $h \rightarrow 0$, this “blows up”, unless $a = \pm b$, in which case the the right hand side is zero for every $h \neq 0$, and so the limit does exist, and is zero. Therefore, for this “bad” function, the directional derivative exists if and only if the direction vector \mathbf{v} is is a non-zero multiple of either $(1, 1)$ or $(-1, 1)$.

There are two “special” direction to consider – the direction along the coordinate axes. These special directional derivatives are called *partial derivatives*:

Definition 44 (Partial derivatives). *Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ defined in a neighborhood of \mathbf{x}_0 , for each $1 \leq j \leq n$, the partial derivative of f with respect to x_j at \mathbf{x}_0 is denoted by $\frac{\partial}{\partial x_j} f(\mathbf{x}_0)$ and is defined by*

$$\frac{\partial}{\partial x_j} f(\mathbf{x}_0) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_j) - f(\mathbf{x}_0)}{h} \quad (4.4)$$

provided that the limit exists.

- Partial derivatives are special cases of directional derivatives – the cases in which the direction vector \mathbf{v} is one of the standard basis vectors \mathbf{e}_i .

Now, let us see how to compute partial derivatives and directional derivatives: This turns out to be easy! If $g(x)$ is related to $f(x, y)$ through

$$g(x) = f(x, y_0),$$

then

$$\frac{\partial}{\partial x} f(x_0, y_0) = \lim_{h \rightarrow 0} \frac{g(x_0 + h) - g(x_0)}{h} = g'(x_0). \quad (4.5)$$

This is great news: We will not need to make explicit use of the definition of partial derivatives very often to compute them. When computing a partial derivative, just treat all of the other variables as constants, and differentiate in the single “active” variable in the usual way. Thus, we can use everything we know about computing derivatives for functions of a single variable when we are computing partial derivatives.

Example 56 (Differentiating in one variable). Let $f(x, y) = \sqrt{1 + x^2 + y^2}$ and $(x_0, y_0) = (1, 2)$. Then with $g(x)$ defined by $g(x) = f(x, y_0)$, $g(x) = \sqrt{5 + x^2}$. By a simple computation,

$$g'(x) = \frac{x}{\sqrt{5 + x^2}}$$

and in particular $\frac{\partial}{\partial x} f(1, 2) = \frac{1}{\sqrt{6}}$.

In single variable calculus, the *derivative function* $g'(x)$ is the function assigning the “output” $g'(x_0)$ to the “input” x_0 .

In the same way, we let $\frac{\partial}{\partial x} f(x, y)$ denote the function of the two variables x and y that assigns the “output” $\frac{\partial}{\partial x} f(x_0, y_0)$ to the “input” (x_0, y_0) . The same conventions apply to other other functions and more variables.

Example 57 (Computing partial derivative functions). Let $f(x, y) = \sqrt{1 + x^2 + y^2}$. Holding y fixed – as a parameter instead of a variable – we differentiate with respect to x as in the single variable calculus, and find

$$\frac{\partial}{\partial x} f(x, y) = \frac{x}{\sqrt{x^2 + y^2}} .$$

Likewise,

$$\frac{\partial}{\partial y} f(x, y) = \frac{y}{\sqrt{x^2 + y^2}} .$$

Once more, because computing partial derivatives is just a matter of differentiating with respect to one chosen variable, everything we know about differentiating with respect to one variable can be applied – in particular the chain rule and the product rule.

Example 58 (Using the single variable chain rule). The function $f(x, y) = \sqrt{1 + x^2 + y^2}$ that we considered in Example 57 can be written as a composition $f(x, y) = g(h(x, y))$ where

$$g(z) = \sqrt{z + 1} \quad \text{and} \quad h(x, y) = x^2 + y^2 .$$

Since

$$g'(z) = \frac{1}{2\sqrt{1+z}} \quad \text{and} \quad \frac{\partial}{\partial x} h(x, y) = 2x ,$$

we have

$$\frac{\partial}{\partial x} f(x, y) = \frac{\partial}{\partial x} g(h(x, y)) = g'(h(x, y)) \frac{\partial}{\partial x} h(x, y) = \frac{1}{2\sqrt{1+h(x,y)}} 2x = \frac{x}{\sqrt{x^2 + y^2}} ,$$

as before.

What we saw in Example 58 is a generally useful fact about partial derivatives: If g is a differentiable function of a single variable, and h is a function of two (or more) variables with $\partial h / \partial x$ defined, then

$$\frac{\partial}{\partial x} g(h(x, y)) = g'(h(x, y)) \frac{\partial}{\partial x} h(x, y) ,$$

and similarly with y and any other variables. The validity of this identity need not be formulated as a theorem, and does not need a new proof: It is true because of the single variable chain rule, and because we are differentiating in the single variable x .

- In short, as far as computing partial derivatives goes, there is nothing much new: Just pay attention to one variable at a time, and differentiate with respect to it as usual.

Now let us see how to compute directional derivatives in terms of partial derivatives. The key is another chain rule, which is a genuinely multivariable chain rule.

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

We have already seen in our study of continuity that knowing the behavior of “slices” of a function f along lines does not tell us the whole story about the behavior of the function f : We need to look at the behavior along more general families of curves. It is the same with differentiability.

In this subsection we prove a chain rule for functions of a vector variable that is useful for understanding the behavior of f over slices along differentiable curves. That is, let $\mathbf{x}(t)$ be a differentiable vector valued function in \mathbb{R}^n . Let f be a function from \mathbb{R}^n to \mathbb{R} . Consider the composite function $g(t)$ defined by

$$g(t) = f(\mathbf{x}(t)) .$$

Here we ask the question:

- Under what conditions on f is g differentiable, and can we compute $g'(t)$ in terms of $\mathbf{x}'(t)$ and the partial derivatives of f ?

Before answering this question, we make a useful definition. We organize the partial derivatives of f into a vector. This definition will figure in most of what we do in the rest of this chapter.

Definition 45 (Gradient). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ have each of its partial derivatives well defined at \mathbf{x}_0 . Then the gradient of f at \mathbf{x}_0 is the vector $\nabla f(\mathbf{x}_0) \in \mathbb{R}^n$ given by*

$$\nabla f(\mathbf{x}_0) = \left(\frac{\partial}{\partial x_1} f(\mathbf{x}_0), \dots, \frac{\partial}{\partial x_n} f(\mathbf{x}_0) \right) .$$

Since you know how to compute partial derivatives, you know how to compute gradients: It is just a matter of organizing the partial derivatives, once you have computed them, into a vector. Here is an example:

Example 59 (Computing a gradient). *With $f(x) = \frac{xy^2}{1+x^2+y^2}$, we compute that*

$$\frac{\partial}{\partial x} f(\mathbf{x}) = \frac{y^2(1+y^2-x^2)}{(1+x^2+y^2)^2}$$

and

$$\frac{\partial}{\partial y} f(\mathbf{x}) = \frac{2xy(1+x^2)}{(1+x^2+y^2)^2} .$$

Therefore,

$$\nabla f(\mathbf{x}) = \frac{1}{(1+x^2+y^2)^2} (y^2(1+y^2-x^2), 2xy(1+x^2)) .$$

We are now ready to state our multivariable chain rule:

Theorem 41 (The chain rule for functions from \mathbb{R}^n to \mathbb{R}). *Let f be any function defined in an open set U of \mathbb{R}^n with values in \mathbb{R} . Suppose that each of the partial derivatives of f is defined and continuous at every point of U . Let $\mathbf{x}(t)$ be a differentiable function from \mathbb{R} to \mathbb{R}^n . Then, for all values of t so that $\mathbf{x}(t)$ lies in U ,*

$$\lim_{h \rightarrow 0} \frac{f(\mathbf{x}(t+h)) - f(\mathbf{x}(t))}{h} = \mathbf{x}'(t) \cdot \nabla f(\mathbf{x}(t)) . \quad (4.6)$$

This is an important theorem, and before proving it, we make some remarks. First, the chain rule in Theorem 41 applies to the composition of functions from \mathbb{R} to \mathbb{R}^n and then from \mathbb{R}^n to \mathbb{R} . Both functions involved in this composition are multivariable functions on one end or the other. The chain rule in Example 58 applies to the composition of functions from \mathbb{R}^n to \mathbb{R} and then from \mathbb{R} to \mathbb{R} . In the latter case, as we have seen, we are really only using the single variable chain rule. But the chain rule of Theorem 41 describes rates of change when all of the variables x_1, \dots, x_n are changing at once. In proving it, we will make essential use of the assumption of continuity of the partial derivatives. Without this assumption, the theorem would not be true.

Second, Theorem 41 has a simple corollary that gives us a formula for computing directional derivatives in terms of partial derivatives.

Corollary 4 (Directional derivatives and gradients). *Let f be any function defined on an open set U of \mathbb{R}^n with values in \mathbb{R} . Suppose that each partial derivative of f is defined and continuous at every point of U . Then for any \mathbf{x}_0 in U , and any direction vector \mathbf{v} in \mathbb{R}^n ,*

$$\lim_{h \rightarrow 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h} = \mathbf{v} \cdot \nabla f(\mathbf{x}_0) . \quad (4.7)$$

Proof: Simply consider the case in which $\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{v}$, and apply Theorem 41. □

If you worked through all the calculations in Example 54, you know that computing directional derivatives “straight from the definition” as we did there can be pretty laborious. The good news is that Corollary 4 provides a *much better way!*

Example 60 (Directional derivatives via gradients). *Consider $f(x) = \frac{xy^2}{1+x^2+y^2}$, $\mathbf{x}_0 = (1, 1)$ and $\mathbf{v} = (1, 2)$ as in Example 54. In that example, we computed (the hard way) that the corresponding directional derivative is 1.*

But now, from Example 59, we have that

$$\nabla f(\mathbf{x}) = \frac{1}{(1+x^2+y^2)^2} (y^2(1+y^2-x^2), 2xy(1+x^2)) ,$$

and hence, substituting $x = 1$ and $y = 1$, we have $\nabla f(\mathbf{x}_0) = \frac{1}{9}(1, 4)$. Therefore, $\mathbf{v} \cdot \nabla f(\mathbf{x}_0) = 1$ is the directional derivative, as we found before with more labor.

The reason that we did not already introduce a special notation for the directional derivative of f at \mathbf{x}_0 in the direction \mathbf{v} is that Corollary 4 provides one, namely $\mathbf{v} \cdot \nabla f(\mathbf{x}_0)$. We couldn’t use it in the last subsection because we hadn’t yet defined gradients, but now that we have, this will be

our standard notation for directional derivatives, at least when we are dealing with “nice” functions whose partial derivatives are continuous.

Corollary 4 provides an efficient means for computing directional derivatives: Once you have computed the gradient, you can take the dot product with many different direction vectors and compute many different directional derivatives without doing any more serious work. In the approach used in Example 54, you would have to start from scratch each time you considered a new direction vector.

We are finally ready for the proof of Theorem 41. The key to the proof, in which we finally explain the importance of continuity for the partial derivatives is the *Mean Value Theorem* from single variable calculus:

The Mean Value Theorem says that if $g(s)$ is continuous on the closed interval $a \leq s \leq b$, and has a derivative $g'(s)$ for all s in the open interval $a < s < b$, then there is a value of c with $a < c < b$ such that

$$\frac{g(b) - g(a)}{b - a} = g'(c) .$$

The principle expressed here is the one by which the police know that if you drove 100 miles in one hour, then at some point on your trip, you were driving at *exactly* 100 miles per hour.

Proof of Theorem 41: We give the proof for $n = 2$ to keep the notation simple. Once this case is understood, the general case will be clear.

Fix some t , and some $h > 0$. To simplify the notation, define the numbers x_0, y_0, x_1 and y_1 by

$$(x_0, y_0) = \mathbf{x}(t) \quad \text{and} \quad (x_1, y_1) = \mathbf{x}(t + h) .$$

Using this notation, note that

$$\begin{aligned} f(\mathbf{x}(t + h)) - f(\mathbf{x}(t)) &= f(x_1, y_1) - f(x_0, y_0) \\ &= [f(x_1, y_1) - f(x_0, y_1)] + [f(x_0, y_1) - f(x_0, y_0)] . \end{aligned}$$

In going from the first line to the second, we have subtracted and added back in the quantity $f(x_0, y_1)$, and grouped the terms in brackets. Why add and subtract the same thing? The point is that in the first group, only the x variable is varying, and in the second group, only the y variable is varying. *Thus, we can use single variable methods on these groups.*

To do this for the first group, define the function $g(s)$ by

$$g(s) = f(x_0 + s(x_1 - x_0), y_1) .$$

Notice that

$$g(1) - g(0) = f(x_1, y_1) - f(x_0, y_1) .$$

Then, if g is continuously differentiable, the Mean Value Theorem tells us that

$$f(x_1, y_1) - f(x_0, y_1) = \frac{g(1) - g(0)}{1 - 0} = g'(c)$$

for some c between 0 and 1.

But by the definition of $g(s)$, we can compute $g'(s)$ by taking a partial derivative of f , since as s varies, only the x component of the input to f is varied. Thus,

$$g'(s) = \frac{\partial}{\partial x} f(x_0 + s(x_1 - x_0), y_1)(x_1 - x_0).$$

Therefore, for some c between 0 and 1,

$$[f(x_1, y_1) - f(x_0, y_1)] = \left[\frac{\partial}{\partial x} f(x_0 + c(x_1 - x_0), y_1) \right] (x_1 - x_0).$$

In the exact same way, we deduce that for some \tilde{c} between 0 and 1,

$$[f(x_0, y_1) - f(x_0, y_0)] = \left[\frac{\partial}{\partial y} f(x_0, y_0 + \tilde{c}(y_1 - y_0)) \right] (y_1 - y_0).$$

Therefore,

$$\begin{aligned} \frac{f(\mathbf{x}(t+h)) - f(\mathbf{x}(t))}{h} &= \left[\frac{\partial}{\partial x} f(x_0 + c(x_1 - x_0), y_1) \right] \frac{x_1 - x_0}{h} \\ &\quad + \left[\frac{\partial}{\partial y} f(x_0, y_0 + \tilde{c}(y_1 - y_0)) \right] \frac{y_1 - y_0}{h}. \end{aligned}$$

Up to now, h has been fixed. But having derived this identity, it is now easy to analyze the limit $h \rightarrow 0$.

First, as $h \rightarrow 0$, $x_1 \rightarrow x_0$ and $y_1 \rightarrow y_0$. Therefore,

$$\lim_{h \rightarrow 0} \frac{\partial}{\partial x} f(x_0 + c(x_1 - x_0), y_1) = \frac{\partial}{\partial x} f(x_0, y_0) = \frac{\partial}{\partial x} f(\mathbf{x}(t)),$$

and

$$\lim_{h \rightarrow 0} \frac{\partial}{\partial y} f(x_0, y_0 + \tilde{c}(y_1 - y_0)) = \frac{\partial}{\partial y} f(x_0, y_0) = \frac{\partial}{\partial y} f(\mathbf{x}(t)).$$

Also, since $\mathbf{x}(t)$ is differentiable

$$\lim_{h \rightarrow 0} \frac{x_1 - x_0}{h} = \lim_{h \rightarrow 0} \frac{x(t+h) - x(t)}{h} = x'(t)$$

and

$$\lim_{h \rightarrow 0} \frac{y_1 - y_0}{h} = \lim_{h \rightarrow 0} \frac{y(t+h) - y(t)}{h} = y'(t)$$

Since the limit of a product is the product of the limits,

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{f(\mathbf{x}(t+h)) - f(\mathbf{x}(t))}{h} &= \left[\frac{\partial}{\partial x} f(\mathbf{x}(t)) \right] x'(t) + \left[\frac{\partial}{\partial y} f(\mathbf{x}(t)) \right] y'(t) \\ &= \nabla f(\mathbf{x}(t)) \cdot \mathbf{x}'(t). \end{aligned}$$

This is what we had to show. \square

4.1.3 The geometric meaning of the gradient

The gradient of a function is a vector. As such, it has a *length*, and a *direction*. To understand the gradient in geometric terms, let us try to understand what the length and direction are telling us.

The key to this is the formula

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta . \quad (4.8)$$

Now pick any point \mathbf{x}_0 and any unit vector \mathbf{u} in \mathbb{R}^n . Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ has continuous partial derivatives at \mathbf{x}_0 , and consider the directional derivative of f at \mathbf{x}_0 in the direction \mathbf{u} . By Theorem 1, this is $\mathbf{u} \cdot \nabla f(\mathbf{x}_0)$.

By 4.8 and the fact that \mathbf{u} is a unit vector (i.e., a pure direction vector),

$$\mathbf{u} \cdot \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\| \cos \theta$$

where θ is the angle between $\nabla f(\mathbf{x}_0)$ and \mathbf{u} . (This is defined as long as $\nabla f(\mathbf{x}_0) \neq 0$, in which case the right hand side is zero.)

As \mathbf{u} ranges over the set of unit vectors in \mathbb{R}^n , i.e., the $n - 1$ dimensional unit sphere in \mathbb{R}^n , $\cos \theta$ varies between -1 and 1 , and hence

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

Recall that by Theorem 1, $\mathbf{u} \cdot \nabla f(\mathbf{x}_0)$ is the slope at \mathbf{x}_0 of the slice of the graph $z = f(\mathbf{x})$ that you get when slicing along $\mathbf{x}_0 + t\mathbf{u}$. Hence we can rephrase this as

$$-\|\nabla f(\mathbf{x}_0)\| \leq [\text{slope of a slice at } \mathbf{x}_0] \leq \|\nabla f(\mathbf{x}_0)\|$$

That is,

- The magnitude of the gradient, $\|\nabla f(\mathbf{x}_0)\|$ tells us the minimum and maximum values of the slopes of all slices of $z = f(\mathbf{x})$ through \mathbf{x}_0 .

The slope has the maximal value, $\|\nabla f(\mathbf{x}_0)\|$, exactly when $\theta = 0$; i.e., when \mathbf{u} and $\nabla f(\mathbf{x}_0)$ point in the same direction. In other words:

- The gradient of f at \mathbf{x}_0 points in the direction of steepest increase of f at \mathbf{x}_0

For the same reasons, we get the steepest negative slope by taking \mathbf{u} to point in the direction of $-\nabla f(\mathbf{x}_0)$.

Example 61 (Which way the water runs). Let $f(x) = \frac{xy^2}{1+x^2+y^2}$, $\mathbf{x}_0 = (1, 1)$ and let $\mathbf{x}_0 = (0, 1)$. If $z = f(\mathbf{x})$ denotes the altitude at \mathbf{x} , and you stood at \mathbf{x}_0 , and spilled a glass of water, which way would the water run?

For purposes of this question, let's say that the direction of the positive x axis is due East, and the direction of the positive y axis is due North.

But now, from Example 59, we have that

$$\nabla f(\mathbf{x}) = \frac{1}{(1+x^2+y^2)^2} (y^2(1+y^2-x^2), 2xy(1+x^2)) ,$$

and hence, substituting $x = 0$ and $y = 1$, we have

$$\nabla f(\mathbf{x}_0) = \frac{1}{4}(2, 0) .$$

Thus, the gradient points due East. This is the “straight uphill” direction. The water will run in the “straight downhill” direction, which is opposite. That is the water will run due West.

4.1.4 Critical points

Theorem 41 has important application to *minimization and maximization problems*, which are problems in which we look for minimum and maximum values of f , and the inputs \mathbf{x} that produce them. Indeed, you see that:

- If $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, then there is an “uphill” direction and a “downhill” direction at \mathbf{x}_0 .

If it is possible to “move uphill” from \mathbf{x}_0 , then $f(\mathbf{x}_0)$ cannot possibly be a maximum value of f . Likewise, if it is possible to “move downhill” from \mathbf{x}_0 , then $f(\mathbf{x}_0)$ cannot possibly be a minimum value of f .

- If we are looking for either minimum values of f or maximum values of f in some open set U , and f has continuous partial derivatives everywhere in U , then it suffices to look among only at those points \mathbf{x} at which $\nabla f(\mathbf{x}) = \mathbf{0}$.

Notice that it is vitally important that the set U be open. In an open set, starting from any point, you can move around, at least a little bit, in all directions while staying inside the set. Hence if $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$ at $\mathbf{x}_0 \in U$, you can always move at least a little bit in any uphill or downhill direction away from \mathbf{x}_0 . However, consider a set that is not open, and has boundary points. At a boundary point, the uphill direction, say, might take you out of the set. So the boundary point still might be a maximum. Consider for example the function $f(x, y) = x^2 + y^2$ defined on the closed unit disc. Each of the boundary points $(\cos \theta, \sin \theta)$ maximizes f on this set, even though the gradient is non-zero at each of them: The uphill direction specified by the gradient leads outside the set, and the fact that one can go further uphill upon leaving the set is irrelevant when we seek to maximize the function f in the set.

The discussion so far leads us to the definition of a critical point:

Definition 46 (Critical point). *Suppose that f is defined and has all of its partial derivatives continuous in a neighborhood of some point \mathbf{x}_0 . Then \mathbf{x}_0 is a critical point of f in case $\nabla f(\mathbf{x}_0) = \mathbf{0}$.*

Example 62 (Computing critical points). *Let $f(x, y) = x^4 + y^4 + 4xy$. We readily compute*

$$\nabla f(x, y) = 4(x^3 + y, y^3 + x) .$$

We find that $\nabla f(x, y) = (0, 0)$ if and only if $x^3 + y = 0$ and $y^3 + x = 0$. Thus if $x = 0$, $y = 0$, If $x \neq 0$, $x^8 = 1$, and so $x = \pm 1$, and $y = -x$. Thus, there are the three critical points, namely

$$(0, 0) \quad (-1, 1) \quad \text{and} \quad (1, -1) .$$

The three critical points found in Example 62 are the *only* points at which f can possibly take on either a maximum value or a minimum value. Computing the values of f at these critical points, we find:

$$f(0, 0) = 0 \quad \text{and} \quad f(-1, 1) = f(1, -1) = -2 .$$

One might be tempted to conclude from this calculation that the maximum value of f on \mathbb{R}^2 is 0, and the minimum value of f on \mathbb{R}^2 is -2 . The answer is only half-right, and the reasoning is

wrong: On the basis of what has been worked out so far, we do not know that either a minimum or a maximum exist.

Indeed, as you can easily check,

$$\lim_{n \rightarrow \infty} f(n, n) = \infty .$$

This means that f has no maximum on \mathbb{R}^2 . On the other hand, f does have a minimum value, and it is -2 . It requires a bit more thinking to see this:

Example 63 (Finding a minimum). *Let us show that $f(x, y) = x^4 + y^4 + 4xy$ does have a minimum value, and it is -2 . First observe that when $\|\mathbf{x}\|$ is large, $f(\mathbf{x})$ is also large, so there is no point in looking outside a compact set for the minimum. To make this precise, let us compare f with a function of $\|\mathbf{x}\|$, as we did in connection with the squeeze principle in Chapter Two. Using the inequalities introduced there, we have*

$$x^4 + y^4 \geq \frac{(x^2 + y^2)^2}{2} \quad \text{and} \quad 2|xy| \leq x^2 + y^2 .$$

Thus,

$$f(\mathbf{x}) \geq \frac{\|\mathbf{x}\|^4}{2} - 2\|\mathbf{x}\|^2 = \frac{1}{2}(\|\mathbf{x}\|^2 - 2)^2 - 2 .$$

In particular,

$$\|\mathbf{x}\| \geq 2 \quad \Rightarrow \quad f(\mathbf{x}) \geq 0 . \tag{4.9}$$

The set $C := \{(x, y) : \|\mathbf{x}\| \leq 2\}$ is closed and bounded. Therefore, by one of the key theorems of Chapter 3, f has a minimum and maximum on C . While the maximum may (and in fact, does) lie on the boundary of C , the minimum does not. Indeed, by (4.9), $f(\mathbf{x}) \geq 0$ for any \mathbf{x} on the boundary of C , and since $f(-1, 1) = f(1, -1) = -2$, no such point can be a minimizer of f on C . Hence, the minimizers all lie in the interior of C , and must be critical points of f .

In Example 62, we have computed the critical points of f , and found that they are $(0, 0)$, $(-1, 1)$ and $(1, -1)$, all of which lies in C . At least one of these must be a minimizer, and since $f(0, 0) = 0$, $f(-1, 1) = g(1, -1) = -2$, we see that $(-1, 1)$ and $(1, -1)$ are minimizers of f , and the minimum value of f is -2 .

Later in the course, we shall return to minimization and maximization problems, and study them in considerable detail. As you see from this example, finding critical points is one important part of finding maxima and minima, but only one part.

4.1.5 The gradient and tangent planes

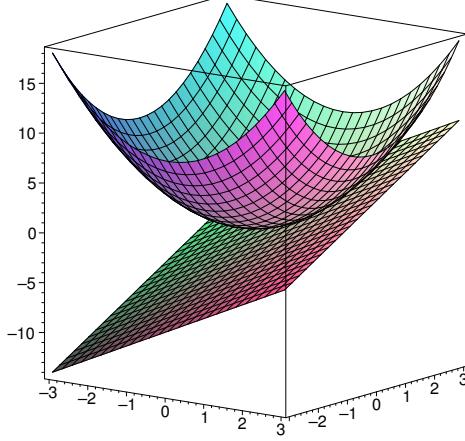
Let g be a differentiable function on \mathbb{R} . Then the graph of $y = g(x)$ is a curve in \mathbb{R}^2 , and

$$y = g(x_0) + g'(x_0)(x - x_0) \tag{4.10}$$

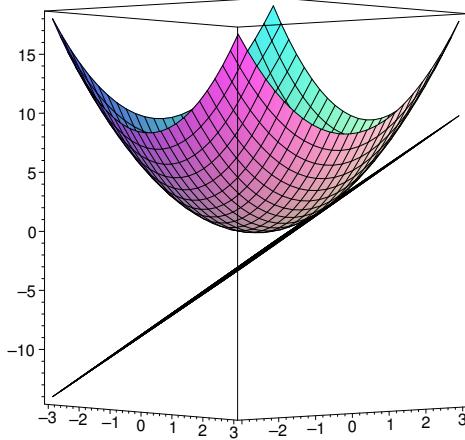
is the equation of the tangent line to this curve at x_0 . This is the line that “best fits” the graph $y = g(x)$ at x_0

Now consider a differentiable function f from \mathbb{R}^2 to \mathbb{R} . The graph of $z = f(x, y)$ is a surface. We now show that when f is differentiable, there is a unique plane, the *tangent plane* that “best fits” the graph $z = f(\mathbf{x})$ at \mathbf{x}_0 . Again, the derivative of f tells us what the tangent plane is.

For example here is the graph of $z = x^2 + y^2$, together with the tangent plane to this graph at the point $(1, 1)$.



Here is another picture of the same thing from a different vantage point, giving a better view of the point of contact:



We now ask:

- How does one compute that equation of the tangent plane to the graph of a differentiable function f from \mathbb{R}^2 to \mathbb{R} at $\mathbf{x}_0 \in \mathbb{R}^2$? For that matter, what is the precise definition of this tangent plane?

Let f be such that all of its partial derivatives are continuous in an open set $U \subset \mathbb{R}^2$. Fix any $\mathbf{x}_0 \in U$. Then for some $r > 0$, U contains every point in the ball of radius r centered on \mathbf{x}_0 . Hence, whenever $\|\mathbf{x} - \mathbf{x}_0\| < r$, $\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0) \in U$ for all $0 \leq t \leq 1$. That is, U contains the line segment connecting \mathbf{x}_0 with \mathbf{x} .

Then by Theorem 41 and the fundamental Theorem of Calculus,

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}_0) &= \int_0^1 \frac{d}{dt} f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) dt \\ &= \int_0^1 \nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) \cdot (\mathbf{x} - \mathbf{x}_0) dt \end{aligned} \tag{4.11}$$

Now when $\mathbf{x} - \mathbf{x}_0$ is small, $\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)$ is close to \mathbf{x}_0 for all $t \in [0, 1]$

Let us suppose that \mathbf{x} is very close to \mathbf{x}_0 , so that $\|\mathbf{x} - \mathbf{x}_0\|$ is very small. Then, if we make the approximation $\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0) \approx \mathbf{x}_0$, then by the continuity of the partial derivatives of f ,

$$\nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) \approx \nabla f(\mathbf{x}_0)$$

for all $t \in [0, 1]$. But the right hand side is independent of t , and so if we use this approximation in (4.11), the integral in t is trivial, and we get

$$f(\mathbf{x}) - f(\mathbf{x}_0) \approx \int_0^1 \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) dt = \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) . \quad (4.12)$$

As we shall see below, this is a very good approximation near \mathbf{x}_0 ; it is the *tangent plane approximation*.

Using the approximation

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) ,$$

the graph of $z = f(\mathbf{x})$ is approximated near \mathbf{x}_0 by the graph of

$$z = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) .$$

To better appreciate the simplicity of this, let us write $\mathbf{x} = (x, y)$, $\nabla f(\mathbf{x}_0) := (a, b)$ and $d := f(\mathbf{x}_0) - \nabla f(\mathbf{x}_0) \cdot \mathbf{x}_0$. Then this becomes $z = ax + by + d$, or equivalently,

$$ax + by - z = d .$$

This is the equation of a non-vertical plane in \mathbb{R}^3 .

Definition 47 (Tangent plane). *Let f be a function on $U \subset \mathbb{R}^2$ such that all of its partial derivatives are continuous on U . Let $\mathbf{x}_0 \in U$. Then the tangent plane to the graph of $z = f(x, y)$ at $\mathbf{x}_0 = (x_0, y_0)$ is given by*

$$z = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) . \quad (4.13)$$

By what we have noted above, if we define

$$\begin{aligned} \mathbf{A} &= \left(\frac{\partial}{\partial x} f(x_0, y_0) , \frac{\partial}{\partial y} f(x_0, y_0) , -1 \right) \\ \mathbf{X}_0 &= (x_0, y_0, f(x_0, y_0)) \\ \mathbf{X} &= (x, y, z) , \end{aligned}$$

the equation (4.13) is equivalent to

$$\mathbf{A} \cdot (\mathbf{X} - \mathbf{X}_0) = 0 .$$

Notice how the gradient of f at \mathbf{x}_0 determines, but is not equal to, the normal vector \mathbf{A} : The gradient is a vector in \mathbb{R}^2 , and the normal vector is a vector in \mathbb{R}^3 .

Example 64 (Tangent planes). *Consider the function $f(x, y) = x^2 + y^2$ with $x_0 = 1$ and $y_0 = 1$. Since f is a polynomial, its partial derivatives are continuous everywhere. Then $f(x_0, y_0) = 2$ and $\nabla f(x_0, y_0) = (2, 2)$. Therefore (4.13) becomes*

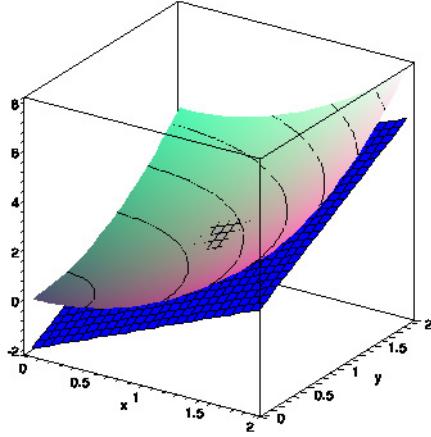
$$z = 2 + (2, 2) \cdot (x - 1, y - 1) = 2x + 2y - 2 .$$

Thus the tangent plane to the graph of f at \mathbf{x}_0 is given by

$$z = 2x + 2y - 2 .$$

Here is a three dimensional graph of f and together with this tangent plane for the region

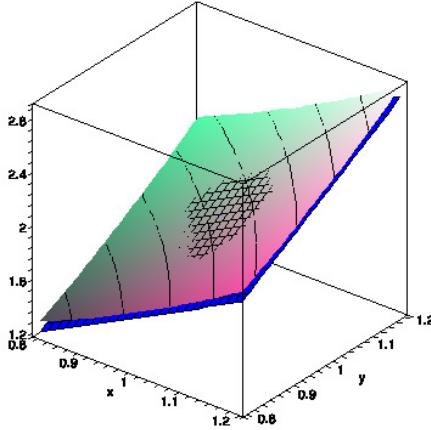
$$|x - 1| \leq 1 \quad \text{and} \quad |y - 1| < 1 : .$$



As you see, the graphs are almost indistinguishable for x and y in the region

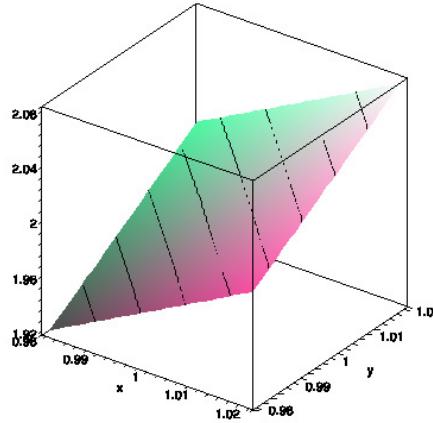
$$|x - 1| \leq 0.2 \quad \text{and} \quad |y - 1| < 0.2 .$$

Let us “zoom in” on this region:



The vertical separation between the graphs is getting to be a pretty small percentage of the displayed distances. The graphs are almost indistinguishable. Let’s zoom in by a factor of 10, and have a look for

$$|x - 1| \leq 0.02 \quad \text{and} \quad |y - 1| < 0.02 .$$



Now, the graphs really are indistinguishable. The geometric meaning of the differentiability of f at \mathbf{x}_0 is exactly this “good fit” between the graph of $z = f(x, y)$ and $z = f(x_0, y_0) + (x - x_0, y - y_0) \cdot \nabla f(x_0, y_0)$.

In our definition of the tangent plane, we have included the requirement that all of the partial derivatives of f be continuous. This is what guaranteed that the tangent plane in the previous example had the very close fit to the graph of the original function f that we saw in the last example. As we see in the next example, without the continuity hypothesis, they may be no plane at all that fits so well, and hence no plane at all deserving to be called a tangent plane.

Example 65 (No tangent plane). Consider the function f defined by

$$f(x, y) = \begin{cases} \frac{2xy}{x^4 + y^4} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases} \quad (4.14)$$

As you can check, both partial derivatives are defined everywhere and

$$f(0, 0) = 0 \quad \frac{\partial f}{\partial x}(0, 0) = 0 \quad \text{and} \quad \frac{\partial f}{\partial y}(0, 0) = 0 .$$

Hence, if we naively applied that tangent plane approximation formula without first checking continuity, we would conclude that

$$z = 0$$

is a good approximation to $z = f(x, y)$ near $(0, 0)$. In fact, it is a terrible approximation. For instance

$$f(t, t) = t^{-2}$$

which gets larger and larger as (t, t) approaches $(0, 0)$. The more you zoom in, the worse the approximation looks.

Moreover, no other plane does any better. If we try the approximation $f(x, y) \approx ax + by + d$ for any a, b and d , we have

$$|f(t, t) - [at + bt + d]| = |t^{-2} - at - bt - d|$$

and this approaches infinity as (t, t) approaches $(0, 0)$. The problem is that the partial derivatives of f are not continuous at $(0, 0)$.

We have just seen that if the partial derivatives of f are not continuous, the tangent plane approximation may not be at all meaningful. We now show that whenever the partial derivatives are continuous, then it is a very good approximation, getting better and better the more one “zooms in” as in Example 64.

First, we *define* differentiability of a function f from \mathbb{R}^n to \mathbb{R} so that, for $n = 2$, it means exactly that that f has a good tangent plan approximation at \mathbf{x}_0 :

Definition 48 (Differentiability of functions from \mathbb{R}^n to \mathbb{R}). *A function f from \mathbb{R}^n to \mathbb{R} is differentiable at \mathbf{x}_0 in case there is a vector $\mathbf{a} \in \mathbb{R}^n$ such that for all $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that*

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta_\epsilon \Rightarrow |f(\mathbf{x}) - [f(\mathbf{x}_0) + \mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0)]| < \epsilon \|\mathbf{x} - \mathbf{x}_0\|. \quad (4.15)$$

An equivalent way to express (4.15), which “hides” the ϵ and δ in the definition of a limit, is

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|f(\mathbf{x}) - [f(\mathbf{x}_0) + \mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0)]|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0. \quad (4.16)$$

Before going further, let’s carefully consider what the definition says when $n = 2$. Note that the graph of the function $h(\mathbf{x}) := f(\mathbf{x}_0) + \mathbf{a}(\mathbf{x} - \mathbf{x}_0)$ is a plane in \mathbb{R}^3 passing through $(x_0, y_0, f(x_0, y_0))$, as does the graph of f itself. The definition says that if you “zoom in” enough – plot the graphs for $\|\mathbf{x} - \mathbf{x}_0\| < \delta$ for some sufficiently small δ – then $|f(\mathbf{x}) - h(\mathbf{x})|$, the vertical separation of the graphs of f and h is an arbitrarily small percentage of $\|\mathbf{x} - \mathbf{x}_0\|$, and of course the graph of h is plane. Thus if you “zoom in” enough, you will not be able to see any noticeable difference in the graphs: The graph of f will appear planar on this scale. Refer back to the pictures in Example 64, and make sure you see how as $\|\mathbf{x} - \mathbf{x}_0\|$ gets small, not only does the vertical also separation get small – it becomes *small as a percentage* of the already small quantity $\|\mathbf{x} - \mathbf{x}_0\|$.

Having explained the relation between the definition of differentiability and the tangent plane approximation in $n = 2$, we return to the general case of arbitrary n .

Observe that if f is differentiable at \mathbf{x}_0 , so that (4.16) is satisfied for *some* \mathbf{a} , there is exactly one \mathbf{a} for which (4.16) is satisfied, namely $\mathbf{a} = \nabla f(\mathbf{x}_0)$. To see this, consider $t \neq 0$ and $\mathbf{x} = \mathbf{x}_0 + t\mathbf{e}_j$ for some $j \in \{1, \dots, n\}$. Then

$$\begin{aligned} \frac{|f(\mathbf{x}) - [f(\mathbf{x}_0) + \mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0)]|}{\|\mathbf{x} - \mathbf{x}_0\|} &= \frac{|f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0) - t\mathbf{a} \cdot \mathbf{e}_j|}{|t|} \\ &= \left| \frac{f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0)}{t} - \mathbf{a} \cdot \mathbf{e}_j \right|. \end{aligned}$$

Then (4.16) is satisfied if and only if

$$\lim_{t \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0)}{t} = \mathbf{a} \cdot \mathbf{e}_j.$$

The left hand side is $\frac{\partial}{\partial x_j} f(\mathbf{x}_0)$ and hence, if f is differentiable at \mathbf{x}_0 , for each $j = 1, \dots, n$, $\frac{\partial}{\partial x_j} f(\mathbf{x}_0)$ exists, and the unique vector \mathbf{a} for which (4.16) is satisfied is $\mathbf{a} = \nabla f(\mathbf{x}_0)$.

Therefore, when f is differentiable at \mathbf{x}_0 , it makes sense to refer to $\nabla f(\mathbf{x}_0)$ as *the derivative of f at \mathbf{x}_0* , and we shall do so. However, as Example 65 shows, mere existence of the partial derivatives necessary to write down $\nabla f(\mathbf{x}_0)$ is not a sufficient condition for f to be differentiable at \mathbf{x}_0 . Fortunately, only a little more does suffice:

Theorem 42. Let f be a function on \mathbb{R}^n , and suppose that the partial derivatives of f , $\frac{\partial}{\partial x_j} f(\mathbf{x})$, $j = 1, \dots, n$, all exist and are continuous on some open set $U \subset \mathbb{R}^n$. Then for any $\mathbf{x}_0 \in U$, f is differentiable at \mathbf{x}_0 .

Proof. By the Fundamental Theorem of Calculus and the Chain Rule, exactly as in (4.11),

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \int_0^1 \nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) \cdot (\mathbf{x} - \mathbf{x}_0) dt ,$$

Now subtracting $f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$ from both sides, we have

$$|f(\mathbf{x}) - [f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)]| = \left| \int_0^1 [\nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) - \nabla f(\mathbf{x}_0)] \cdot (\mathbf{x} - \mathbf{x}_0) dt \right| . \quad (4.17)$$

By the Cauchy-Schwarz inequality,

$$|\nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) - \nabla f(\mathbf{x}_0)| \leq \|\nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) - \nabla f(\mathbf{x}_0)\| \|\mathbf{x} - \mathbf{x}_0\| . \quad (4.18)$$

Consider the function $g(\mathbf{y})$ defined by

$$g(\mathbf{y}) := \|\nabla f(\mathbf{x}_0 + \mathbf{y}) - \nabla f(\mathbf{x}_0)\| . \quad (4.19)$$

Then we may rewrite (4.18) as

$$\frac{|\nabla f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) - \nabla f(\mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|} \leq \int_0^1 g(t(\mathbf{x} - \mathbf{x}_0)) dt . \quad (4.20)$$

Since the partial derivatives of f are continuous, g is continuous and $g(\mathbf{0}) = 0$. Thus for all $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$\|\mathbf{y}\| < \delta_\epsilon \Rightarrow g(\mathbf{y}) < \epsilon . \quad (4.21)$$

Then for \mathbf{x} such that $\|\mathbf{x} - \mathbf{x}_0\| < \delta_\epsilon$, $g(t(\mathbf{x} - \mathbf{x}_0)) < \epsilon$ for all $t \in [0, 1]$, and then $\int_0^1 g(t(\mathbf{x} - \mathbf{x}_0)) dt < \epsilon$. Using this in (4.20), we have

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta_\epsilon \Rightarrow |f(\mathbf{x}) - [f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)]| < \epsilon \|\mathbf{x} - \mathbf{x}_0\| ,$$

which is (4.15) with $\mathbf{a} = \nabla f(\mathbf{x}_0)$. □

What makes the tangent plane approximation so useful is that it provides “the best linear approximation” to the possibly complicated function $f(\mathbf{x}) - f(\mathbf{x}_0)$ by the linear function $ax + by$ where $(a, b) = \nabla f(\mathbf{x}_0)$. It is the “best” such approximation in that the function $h(\mathbf{x}) = \mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0)$ is such that

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|(f(\mathbf{x}) - f(\mathbf{x}_0)) - h(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0$$

only for for $\mathbf{a} = \nabla f(\mathbf{x}_0)$: Any other linear approximation fails to fit this well: No matter how much one “zooms in” on the graphs near (x_0, y_0) , they will stay distinguishable.

- They key idea of the differential calculus is to approximate non-linear functions by their “best linear approximation” wherever possible.

For functions f from \mathbb{R}^2 to \mathbb{R} that have continuous partial derivatives, it is the tangent plane that provides the best linear approximation. But what about functions \mathbf{f} from \mathbb{R}^n to \mathbb{R}^m ? To move forward, we need to do three things:

- (1) We need to explain what a linear function from \mathbb{R}^n to \mathbb{R}^m is.
- (2) We need to explain what “best linear approximation” means in precise terms.
- (3) We need to explain why it is so useful to approximate non-linear functions by linear functions, and how methods of exact calculation can be based on a sequence of successive linear approximations.

In the next section we deal with the first task.

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

Definition 49 (Linear functions). *Let \mathbf{f} be a function defined on \mathbb{R}^n with values in \mathbb{R}^m . Then \mathbf{f} is a linear function in case for all $s, t \in \mathbb{R}$ and all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

$$\mathbf{f}(s\mathbf{x} + t\mathbf{y}) = s\mathbf{f}(\mathbf{x}) + t\mathbf{f}(\mathbf{y}) . \quad (4.22)$$

Specializing to $\mathbf{y} = \mathbf{0}$, we see that for all $s \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{f}(s\mathbf{x}) = s\mathbf{f}(\mathbf{x}) .$$

A function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with this property is said to be *homogeneous*. Thus, linear functions are always homogeneous.

Further specializing to $s = 0$, since $0\mathbf{x} = \mathbf{0}$ for all \mathbf{x} , (4.22) becomes $\mathbf{f}(\mathbf{0}) = \mathbf{0}$: A linear function *always* has the output value $\mathbf{0}$ at the input value $\mathbf{0}$. The following theorem establishes another key property of linear functions

Theorem 43 (Linear functions respect linear combination). *Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear. Then for any linear combination $\sum_{j=1}^r x_j \mathbf{x}_j$ in \mathbb{R}^n ,*

$$\mathbf{f}\left(\sum_{j=1}^r x_j \mathbf{x}_j\right) = \sum_{j=1}^r x_j \mathbf{f}(\mathbf{x}_j) . \quad (4.23)$$

Proof. If $r = 2$, this follows from the definition of linearity, or the remark made on homogeneity following this definition. The general case is proved by induction:

$$\mathbf{f}\left(\sum_{j=1}^r x_j \mathbf{x}_j\right) = \mathbf{f}\left(x_1 \mathbf{x}_1 + 1 \left(\sum_{j=2}^r x_j \mathbf{x}_j\right)\right) = x_1 \mathbf{f}(\mathbf{x}_1) + \mathbf{f}\left(\sum_{j=2}^r x_j \mathbf{x}_j\right) ,$$

where in the last equality we have used the definition of linearity. Now making the inductive hypothesis that the theorem is true for all linear combinations of $r - 1$ or fewer vectors,

$$\mathbf{f}\left(\sum_{j=2}^r x_j \mathbf{x}_j\right) = \sum_{j=2}^r x_j \mathbf{f}(\mathbf{x}_j) .$$

Combining results, we have proved (4.23). □

4.2.1 The matrix representation of linear functions

Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear. Theorem 43 tells us that for any $\mathbf{x} = \sum_{j=1}^n x_j \mathbf{e}_j \in \mathbb{R}^n$,

$$\mathbf{f}(\mathbf{x}) = \mathbf{f} \left(\sum_{j=1}^n x_j \mathbf{e}_j \right) = \sum_{j=1}^n x_j \mathbf{f}(\mathbf{e}_j) . \quad (4.24)$$

The right hand side of (4.24) is a linear combination of the n vectors

$$f(\mathbf{e}_j) \quad j = 1, \dots, n .$$

- All of the data needed to compute $\mathbf{f}(\mathbf{x})$ for any $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ is the list of the n vectors $f(\mathbf{e}_j)$ for $j = 1, \dots, n$.

Example 66 (Evaluating a linear function \mathbf{f} given $\mathbf{f}(\mathbf{e}_j)$, $j = 1, \dots, n$). Consider a linear function $\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ such that

$$\mathbf{f}(\mathbf{e}_1) = (-2, 1, 2) \quad \mathbf{f}(\mathbf{e}_2) = (1, -2, 2) \quad \text{and} \quad \mathbf{f}(\mathbf{e}_3) = (1, 1, 1) . \quad (4.25)$$

Then,

$$\begin{aligned} \mathbf{f}(2, 3, 4) &= \mathbf{f}(2\mathbf{e}_1 + 3\mathbf{e}_2 + 4\mathbf{e}_3) \\ &= 2\mathbf{f}(\mathbf{e}_1) + 3\mathbf{f}(\mathbf{e}_2) + 4\mathbf{f}(\mathbf{e}_3) \\ &= 2(-2, 1, 2) + 3(1, -2, 2) + 4(1, 1, 1) \\ &= (3, 0, 14) . \end{aligned}$$

- This feature of linear functions is the essence of their simplicity. Though there are infinitely many possible inputs \mathbf{x} at which a linear function \mathbf{f} might be evaluated, once you know the values of $\mathbf{f}(\mathbf{e}_j)$ for each $j \in \{1, \dots, n\}$, you know how to evaluate $\mathbf{f}(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^n$.

Definition 50 (Matrix of a linear transformation). Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear. The matrix $A_{\mathbf{f}}$ corresponding to \mathbf{f} is the list of the n vectors $\{\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)\}$ in \mathbb{R}^m , which is written as an $m \times n$ array with $\mathbf{f}(\mathbf{e}_j)$ in the j th column of the array. We express this by writing

$$A_{\mathbf{f}} = [\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)] . \quad (4.26)$$

An $m \times n$ matrix A is any such $m \times n$ array of numbers. The i, j th entry in the matrix A is denoted $A_{i,j}$.

Example 67. Let \mathbf{f} be the function from \mathbb{R}^3 to \mathbb{R}^3 given by (4.25). Then placing the three vectors $\mathbf{f}(\mathbf{e}_1)$, $\mathbf{f}(\mathbf{e}_2)$ and $\mathbf{f}(\mathbf{e}_3)$, respectively, as the columns in a 3×3 array, we have

$$A_{\mathbf{f}} := \begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 2 & 2 & 1 \end{bmatrix} .$$

With regard to most problems we shall encounter here, it is more helpful to think of matrices as lists of vectors rather than rectangular arrays of numbers. That is, we will find it useful to think of the matrix A_f in Example 67 as a list of three vectors:

$$A_f = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] ,$$

and *sometimes* to write the vectors themselves as *vertical lists* of numbers, departing from our practice up to now of writing vectors as horizontal lists of numbers. In this vertical list notation, we would write, still referring to Example 67,

$$\mathbf{v}_1 = \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix} \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} .$$

Example 68 (The matrix of a Householder reflection in \mathbb{R}^3). Let $\mathbf{u} = \frac{1}{\sqrt{3}}(-1, 1, -1)$, and consider the Householder reflection \mathbf{h}_u . We have seen in Chapter One that $\mathbf{x} \mapsto \mathbf{h}_u(\mathbf{x})$ is a linear transformation from \mathbb{R}^3 to \mathbb{R}^3 . What is the 3×3 matrix that represents this linear transformation?

To answer this question, we need only compute $\mathbf{h}_u(\mathbf{e}_1)$, $\mathbf{h}_u(\mathbf{e}_2)$ and $\mathbf{h}_u(\mathbf{e}_3)$, and place these vectors in the first, second and third columns of a 3×3 matrix. From the formula $\mathbf{h}_u(\mathbf{x}) = \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{u})\mathbf{u}$, it is easy to compute $\mathbf{h}_u(\mathbf{e}_1)$, $\mathbf{h}_u(\mathbf{e}_2)$ and $\mathbf{h}_u(\mathbf{e}_3)$, and hence the corresponding matrix. One finds:

$$A_{\mathbf{h}_u} = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix} .$$

This is the 3×3 matrix representing the linear transformation \mathbf{h}_u . Make sure you do the computations yourself, in full detail.

Example 69 (The matrix of a cross product transformation). Let $\mathbf{b} \in \mathbb{R}^3$ be given by

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

Consider the function $\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{b} \times \mathbf{x} .$$

We have seen in Chapter One that for any $s, t \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$,

$$\mathbf{b} \times (s\mathbf{x} + t\mathbf{y}) = s(\mathbf{b} \times \mathbf{x}) + t(\mathbf{b} \times \mathbf{y}) .$$

Thus, \mathbf{f} is linear. What is the 3×3 matrix that represents this linear transformation?

To answer this question, we need only compute $\mathbf{f}(\mathbf{e}_1)$, $\mathbf{f}(\mathbf{e}_2)$ and $\mathbf{f}(\mathbf{e}_3)$, and place these vectors in the first, second and third columns of a 3×3 matrix. The result is:

$$A_f = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix} .$$

Make sure you do the computations yourself, in full detail.

Example 70 (The identity matrix). *The next example is really simple, but also really important. The identity transformation \mathbf{Id} on \mathbb{R}^n is given by*

$$\mathbf{Id}(\mathbf{x}) = \mathbf{x} .$$

What could be more simple?

Since, by definition,

$$\mathbf{Id}(s\mathbf{x} + t\mathbf{y}) = s\mathbf{x} + t\mathbf{y} = s\mathbf{Id}(\mathbf{x}) + t\mathbf{Id}(\mathbf{y}) ,$$

the identity transformation is linear. What is the corresponding $n \times n$ matrix?

Since $\mathbf{Id}(\mathbf{e}_j) = \mathbf{e}_j$, it is the $n \times n$ matrix with \mathbf{e}_j , written as column vector, in the j th column. For example, the 4×4 identity matrix, representing the identity transformation on \mathbb{R}^4 , is

$$I_{4 \times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

So far, we have seen that a list of the vectors $[\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)]$ is all that one needs to compute $\mathbf{f}(\mathbf{x})$ given $\mathbf{x} = (x_1, \dots, x_n)$. How can we automate this computational process?

- *The multiplication of matrices and vectors is defined so that if A_f is the matrix corresponding to the linear function \mathbf{f} , the matrix product $A_f \mathbf{x}$, yields $\mathbf{f}(\mathbf{x})$, the value of the function \mathbf{f} at \mathbf{x} .*
- *Before proceeding, let us be clear on the notation we shall use: Some texts make a distinction between row vectors and column vectors. We shall not. Instead, we shall simply write vectors in \mathbb{R}^n in either row or column form, using whichever notation seems convenient at the time. In particular, the dot product of two vectors \mathbf{r} and \mathbf{x} means exactly what it meant in Chapter 1, even if now we write one vector as a row and the other as a column.*

Definition 51 (Matrix-vector multiplication). *Let $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ be an $m \times n$ matrix, where the j th column of A is the vector $\mathbf{v}_j \in \mathbb{R}^m$. For every vector $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, the product of the matrix A and the vector \mathbf{x} is the vector*

$$A\mathbf{x} = [\mathbf{v}_1, \dots, \mathbf{v}_n](x_1, \dots, x_n) = \sum_{j=1}^n x_j \mathbf{v}_j . \quad (4.27)$$

Note that in (4.27) we could have written $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ in place of (x_1, \dots, x_n) , and it would not matter (apart from taking more space). What matters is the equality between the first and third terms in (4.27).

In particular, for a linear function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and $A_f = [\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)]$, we have

$$A_f \mathbf{x} = [\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)](x_1, \dots, x_n) = \sum_{j=1}^n x_j \mathbf{f}(\mathbf{e}_j) = \mathbf{f} \left(\sum_{j=1}^n x_j \mathbf{e}_j \right) = \mathbf{f}(\mathbf{x}) .$$

Thus, the definition of matrix-vector multiplication has been arranged so that

$$\mathbf{f}(\mathbf{x}) = A_{\mathbf{f}} \mathbf{x} . \quad (4.28)$$

By Definition 51, for any $m \times n$ matrix A , $A\mathbf{e}_j$ is the j th column of A . Since $A_{i,j}$ is by definition the i th entry of the j th column of A , and since the i th entry of any vector $\mathbf{y} \in \mathbb{R}^m$ is given by $y_i = \mathbf{e}_i \cdot \mathbf{y}$, it follows that

$$A_{i,j} = (A\mathbf{e}_j)_i = \mathbf{e}_i \cdot A\mathbf{e}_j . \quad (4.29)$$

So far, we have considered $m \times n$ matrices as lists of n vectors in \mathbb{R}^m , through the relation $A = [A\mathbf{e}_1, \dots, A\mathbf{e}_n]$.

It is also profitable to consider an $m \times n$ matrix as a list of m vectors in \mathbb{R}^n , namely the n rows in A . For example, if

$$A = \begin{bmatrix} -2 & 1 & 1 & 5 \\ 1 & -2 & 1 & 0 \\ 2 & 2 & 1 & -1 \end{bmatrix} .$$

the three rows of A , \mathbf{a}_1 , \mathbf{a}_2 and \mathbf{a}_3 are

$$\begin{aligned} \mathbf{a}_1 &= (-2, 1, 1, 5) \\ \mathbf{a}_2 &= (1, -2, 1, 0) \\ \mathbf{a}_3 &= (2, 2, 1, -1) . \end{aligned} \quad (4.30)$$

Thinking of A as a list of its rows allows us to think of matrix multiplication in terms of the dot product, and brings geometry into the picture. The key observation is that the i th entry of $A\mathbf{x}$ is the dot product of the i th row of A with \mathbf{x} . Indeed, the i th entry of $A\mathbf{x}$ is given by

$$(A\mathbf{x})_i = \mathbf{e}_i \cdot A\mathbf{x} = \mathbf{e}_i \cdot \left(\sum_{j=1}^n (A\mathbf{e}_j)x_j \right) = \sum_{j=1}^n (\mathbf{e}_i \cdot A\mathbf{e}_j)x_j = \sum_{j=1}^n A_{i,j}x_j \quad (4.31)$$

Defining the vector

$$\mathbf{a}_i = (A_{i,1}, \dots, A_{i,n}), \quad (4.32)$$

which is the i th row of A , we can rewrite the conclusion of (4.31) as $(A\mathbf{x})_i = \mathbf{a}_i \cdot \mathbf{x}$. This gives us a formula for matrix-vector multiplication in terms of the rows of A :

$$A\mathbf{x} = (\mathbf{a}_1 \cdot \mathbf{x}, \dots, \mathbf{a}_n \cdot \mathbf{x}) \quad \text{for} \quad A = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix} . \quad (4.33)$$

Along the way to (4.33) we encountered another formula, with which you may be familiar:

$$(A\mathbf{x})_i = \sum_{j=1}^n A_{i,j}x_j . \quad (4.34)$$

The formulae (4.33) and (4.34) are two ways of expressing the same thing. If one primarily thinks of m by n matrices as m by n rectangular arrays of numbers, then (4.34) is natural. And it certainly has its uses.

However, the geometric interpretation of the dot product together with (4.33) will allow us to use geometric methods to solve equations involving linear transformations. This turns out to be far more useful than one might expect. Let us summarize:

Theorem 44 (Matrix vector multiplication in terms of matrix rows). *Let $A = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix}$ be an $m \times n$ matrix expressed as a list of its m rows. Then for any $\mathbf{x} \in \mathbb{R}^n$,*

$$A\mathbf{x} = (\mathbf{a}_1 \cdot \mathbf{x}, \dots, \mathbf{a}_m \cdot \mathbf{x}).$$

The “row representation” is closely connected with our practice of writing any sort of function from \mathbb{R}^n to \mathbb{R}^m in the form

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x})),$$

that is, in terms of a list of m functions from \mathbb{R}^n to R . As you see, each $f_i(\mathbf{x})$ is given by

$$f_i(\mathbf{x}) = \mathbf{e}_i \cdot \mathbf{f}(\mathbf{x}). \quad (4.35)$$

Now suppose that \mathbf{f} happens to be the linear transformation from \mathbb{R}^n to \mathbb{R}^m given by an $m \times n$ matrix A by $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$. Let us write A in terms of its rows: $A = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix}$. Then by (4.31), (4.32) and (4.33), (4.36) becomes

$$f_i(\mathbf{x}) = \mathbf{e}_i \cdot A\mathbf{x} = \mathbf{a}_i \cdot \mathbf{x}. \quad (4.36)$$

The following theorem gives us an alternative way to think about linear functions from \mathbb{R}^n to \mathbb{R}^m :

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 , and all $\mathbf{x} \in \mathbb{R}^m$,*

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x}.$$

Proof. We have seen that if \mathbf{f} is linear, then with $A = [\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)]$, $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ for all \mathbf{x} . Conversely, suppose A is any $m \times n$ matrix, let $\mathbf{a}_1, \dots, \mathbf{a}_m$ be its rows so that

$$A\mathbf{x} = (\mathbf{a}_1 \cdot \mathbf{x}, \dots, \mathbf{a}_m \cdot \mathbf{x}). \quad (4.37)$$

Note that fact that for all \mathbf{a}, \mathbf{y} and $\mathbf{z} \in \mathbb{R}^n$ and all s and t ,

$$\mathbf{a} \cdot (s\mathbf{y} + t\mathbf{z}) = s\mathbf{a} \cdot \mathbf{y} + t\mathbf{a} \cdot \mathbf{z}. \quad (4.38)$$

Taking $\mathbf{x} = s\mathbf{y} + t\mathbf{z}$ in (4.37) and using (4.38) in each entry, $A(s\mathbf{y} + t\mathbf{z}) = sA\mathbf{y} + tA\mathbf{z}$, so the function defined by sending \mathbf{x} to $A\mathbf{x}$ is linear for all $n \times n$ matrices A . \square

As we shall see, there are many convenient “matrix manipulation” methods for solving equations such as $A\mathbf{x} = \mathbf{b}$. In fact, *linear algebra* provides a complete set of methods for answering almost any question concerning linear transformations. There is no such complete theory for non-linear transformations, which is why linear approximation is so important.

Thus, linear algebra is an essential part of the theory of multivariable calculus, and we shall introduce many aspects of linear algebra as we develop the theory of multivariable calculus. Of course in single variable calculus, $m = n = 1$, and there is not much to say about 1×1 matrices: Linear algebra in one variable is trivial, and it does not get mentioned by name in single variable calculus. But already with two variables, it plays an essential role.

In closing this subsection, we reiterate a point about notation: A vector \mathbf{x} in \mathbb{R}^n is an ordered list of n real numbers, and how we record the order – top down in a column, left to right in a row, or some other choice does not really matter. An $m \times n$ matrix A can be thought of as a list of vectors in two distinct ways, and this distinction matters: A can be regarded as the list of its m rows in \mathbb{R}^n , or as the list of its n columns in \mathbb{R}^m , and it is natural to write the row lists vertically, and the column lists horizontally, and we always do this.

4.2.2 Composition of linear functions and matrix multiplication

Let \mathbf{f} be a linear function from \mathbb{R}^n to \mathbb{R}^m , and let \mathbf{g} be a linear function from \mathbb{R}^m to \mathbb{R}^p . Since the range of \mathbf{f} lies in the domain of \mathbf{g} , the composition $\mathbf{g} \circ \mathbf{f}$ is well defined by

$$\mathbf{g} \circ \mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x})) .$$

Let $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ be the $m \times n$ matrix representing \mathbf{f} , so that $\mathbf{f}(\mathbf{x}) = \sum_{j=1}^n x_j \mathbf{v}_j$. Then since \mathbf{g} is linear, there is an $p \times m$ matrix B representing \mathbf{g} . We then have, using first the linearity of \mathbf{g} and then the matrix representation,

$$\mathbf{g} \circ \mathbf{f}(\mathbf{x}) = \mathbf{g} \left(\sum_{j=1}^n x_j \mathbf{v}_j \right) = \sum_{j=1}^n x_j \mathbf{g}(\mathbf{v}_j) = \sum_{j=1}^n x_j B \mathbf{v}_j .$$

In particular, if we define the $p \times n$ matrix C by $C = [B\mathbf{v}_1, \dots, B\mathbf{v}_n]$. then $\mathbf{g} \circ \mathbf{f}(\mathbf{x}) = C\mathbf{x}$.

Since matrix multiplication is a linear operation – see Theorem 45 – the composition of linear functions \mathbf{g} and \mathbf{f} , $\mathbf{g} \circ \mathbf{f}$, is linear, and is therefore represented by a matrix. *We define matrix multiplication so that the matrix representing $\mathbf{g} \circ \mathbf{f}$ is the matrix product of the matrices representing \mathbf{g} and \mathbf{f} .*

Definition 52 (Matrix multiplication). *Let A be an $m \times n$ matrix with columns $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^m$ so that*

$$A = [\mathbf{v}_1, \dots, \mathbf{v}_n] .$$

Let B be an $p \times m$ matrix Then the matrix product of B and A is the $p \times n$ matrix BA where

$$BA = [B\mathbf{v}_1, \dots, B\mathbf{v}_n] .$$

Example 71. Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 1 \end{bmatrix}$. Let $B = \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & -1 \end{bmatrix}$. Then $\mathbf{v}_1 = (1, 2, 1)$ and $\mathbf{v}_2 = (2, 1, 1)$ be the columns of A . We compute $B\mathbf{v}_1 = (3, 3)$ and $B\mathbf{v}_2 = (2, 4)$. Writing these vectors in as the columns of BA we get

$$BA = \begin{bmatrix} 3 & 2 \\ 3 & 4 \end{bmatrix} .$$

Next, let $\mathbf{w}_1 = (1, 2)$, $\mathbf{w}_3 = (2, 1)$ and $\mathbf{w}_3 = (-2, -1)$ be the columns of B . We then compute

$$A\mathbf{w}_1 = (5, 4, 3) , \quad A\mathbf{w}_2 = (4, 5, 3) , \quad \text{and} \quad A\mathbf{w}_3 = (-4, -5, -3) .$$

Writing these vectors in as the columns of AB we get

$$AB = \begin{bmatrix} 5 & 4 & -4 \\ 4 & 5 & -5 \\ 3 & 3 & -3 \end{bmatrix} .$$

Note that $AB \neq BA$. That is matrix multiplication is not commutative. In this example AB and BA are even matrices of different sizes. Even worse, let C be any 2×2 matrix. Then, since A is 3×2 , AC is defined, but CA is not even defined.

Theorem 46 (Matrix multiplication and composition of linear functions). *Let \mathbf{f} be a linear function from \mathbb{R}^n to \mathbb{R}^m , and let \mathbf{g} be a linear function from \mathbb{R}^m to \mathbb{R}^p . Let A be the $m \times n$ matrix representing \mathbf{f} , and let B be the $p \times m$ matrix representing \mathbf{g} . Then $\mathbf{g} \circ \mathbf{f}$ is a linear function from \mathbb{R}^n to \mathbb{R}^p , and BA is matrix representative.*

Proof. By definition, the matrix representing $\mathbf{g} \circ \mathbf{f}$ is $C := [\mathbf{g} \circ \mathbf{f}(\mathbf{e}_1), \dots, \mathbf{g} \circ \mathbf{f}(\mathbf{e}_n)]$ by definition of B , $\mathbf{g} \circ \mathbf{f}(\mathbf{e}_j) = B\mathbf{f}(\mathbf{e}_j)$. Thus,

$$C = [B\mathbf{f}(\mathbf{e}_1), \dots, B\mathbf{f}(\mathbf{e}_n)] = B[\mathbf{f}(\mathbf{e}_1), \dots, \mathbf{f}(\mathbf{e}_n)] = BA .$$

□

Corollary 5 (Associativity of matrix multiplication). *Let A be an $n \times m$, let B be a $m \times p$ matrix, and let C be a $p \times q$ matrix. Then*

$$A(BC) = (AB)C .$$

In other words, matrix multiplication is associative.

Proof. Let \mathbf{f} , \mathbf{g} and \mathbf{h} be the linear transformations corresponding to A , B and C respectively. Then $A(BC)$ is the matrix representative of $\mathbf{f} \circ (\mathbf{g} \circ \mathbf{h})$ and $(AB)C$ is the matrix representative of $(\mathbf{f} \circ \mathbf{g}) \circ \mathbf{h}$. But by definition, for all \mathbf{x} ,

$$[\mathbf{f} \circ (\mathbf{g} \circ \mathbf{h})](\mathbf{x}) = \mathbf{f}(\mathbf{g} \circ \mathbf{h})(\mathbf{x}) = \mathbf{f}(\mathbf{g}(\mathbf{h}(\mathbf{x}))) = (\mathbf{f} \circ \mathbf{g})(\mathbf{h}(\mathbf{x})) = [(\mathbf{f} \circ \mathbf{g}) \circ \mathbf{h}](\mathbf{x}) .$$

In other words, since the composition of functions is associative – linear or not – matrix multiplication is associative because it directly represents composition of linear functions. □

4.2.3 Solving the equation $Ax = b$

Let f be a linear function from \mathbb{R}^n to \mathbb{R}^m . Let $A := A_f$ be its $m \times n$ matrix. Let b be a given vector in \mathbb{R}^m . A basic problem of linear algebra is to find all vectors $x \in \mathbb{R}^n$, if any, that satisfy the equation

$$Ax = b , \quad (4.39)$$

or what is the same thing, to find all vectors $x \in \mathbb{R}^n$, if any, that satisfy $f(x) = b$. There are two basic questions: *When do solutions exist? Supposing a solution does exist, is it unique?* These are usually referred to as the existence and uniqueness questions.

Concerning the existence question, let $A = [v_1, \dots, v_n]$ be an $m \times n$ matrix written as a list of its columns. If $x = (x_1, \dots, x_n)$, then $Ax = \sum_{j=1}^n x_j v_j$. Note that the right hand is the general element of $\text{Span}(\{v_1, \dots, v_n\})$. Hence there exists a vector $x \in \mathbb{R}^n$ such that $Ax = b$ if and only if $b \in \text{Span}(\{v_1, \dots, v_n\})$. This proves:

Theorem 47. *Let $A = [v_1, \dots, v_n]$ be an $m \times n$ matrix, and let $b \in \mathbb{R}^m$. Then $Ax = b$ has at least one solution if and only if $b \in \text{Span}(\{v_1, \dots, v_n\})$.*

Definition 53 (Column space of a Matrix). *$A = [v_1, \dots, v_n]$ be an $m \times n$ matrix. Then the subspace of \mathbb{R}^m – which may be all of \mathbb{R}^m – spanned by the columns of A is called the column space of A , or the range of A , and is denoted $\text{Ran}(A)$. That is, $\text{Ran}(A) = \text{Span}(\{v_1, \dots, v_n\})$*

We now turn to the uniqueness question, which we shall see is intimately related to the corresponding question for the *homogeneous equation*

$$Ax = 0 . \quad (4.40)$$

While $Ax = b$ may have *no solutions*, the equation $Ax = 0$ always has at least one solution, namely $x = 0$, which is called the *trivial solution*.

Theorem 48. *Let $A = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix}$ be an $m \times n$ matrix. The solution set of the equation $Ax = 0$ is the set of vectors that is orthogonal to each row of A ; that is, the subspace of \mathbb{R}^n given by $\{\mathbf{a}_1, \dots, \mathbf{a}_m\}^\perp$.*

Proof. This is an immediate consequence of (4.33), the formula for matrix-vector multiplication in terms of rows. \square

The orthogonal complement of any set of vectors in \mathbb{R}^n is always a subspace of \mathbb{R}^n ; see Theorem 19. and moreover, again by Theorem 19, it is also the orthogonal complement of the span of this set of vectors. That is, $\text{Null}(A)$ is the orthogonal complement of the span of the rows of A .

Definition 54 (The null space and the row space of a matrix). *The null space of an $m \times n$ matrix A , denoted $\text{Null}(A)$, is the orthogonal complement of the set of rows of A , or equivalently by Theorem 48, the set of solutions of $Ax = 0$. The row space of A is the subspace of \mathbb{R}^n spanned by the rows of A .*

Theorem 49. *Let A be an $m \times n$ matrix. Then solutions of $Ax = b$ are unique whenever they exist if and only if $\text{Null}(A) = \{0\}$; i.e., if and only if the only solution of $Ax = 0$ is the trivial solution.*

Proof. If $A\mathbf{z} = \mathbf{0}$ for some $\mathbf{z} \neq \mathbf{0}$, and if $A\mathbf{x}_0 = \mathbf{b}$, then for any $t \in \mathbb{R}$, $A(\mathbf{x}_0 + t\mathbf{z}) = A\mathbf{x}_0 + tA\mathbf{z} = \mathbf{b}$, so that $\mathbf{x} + t\mathbf{z}$ is a line of infinitely many solutions of $A\mathbf{x} = \mathbf{b}$. Hence $\text{Null}(A) = \{\mathbf{0}\}$ is necessary for solutions to be unique when they exist.

To see that it is sufficient, suppose that $\text{Null}(A) = \{\mathbf{0}\}$, and suppose that $A\mathbf{x}_1 = \mathbf{b}$ and $A\mathbf{x}_2 = \mathbf{b}$. Then $A(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{b} - \mathbf{b} = \mathbf{0}$, so that $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$. That is, the two solutions \mathbf{x}_1 and \mathbf{x}_2 are the same. \square

We are now ready to draw some important conclusions. Let A be an $m \times n$ matrix. According to Theorem 47, $A\mathbf{x} = \mathbf{b}$ has at least one solution for every $\mathbf{b} \in \mathbb{R}^m$ if and only if $\text{Ran}(B) = \mathbb{R}^m$. According to Theorem 49, solutions of $A\mathbf{x} = \mathbf{b}$ are unique when they exist if and only if $\text{Null}(A) = \{0\}$. But since $\text{Null}(A)$ is the orthogonal complement of the row space of A . $\text{Null}(A) = \{0\}$ if and only if the row space of A is all of \mathbb{R}^n .

Thus, an $m \times n$ matrix A defines a linear transformation from \mathbb{R}^n to \mathbb{R}^m that is *onto* \mathbb{R}^n ; i.e., for every $\mathbf{b} \in \mathbb{R}^m$ there is at least one $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$, if and only if the column space of A is all of \mathbb{R}^m . Moreover, it is *one-to-one*; i.e., if $A\mathbf{x}_1 = \mathbf{b}$ and $A\mathbf{x}_2 = \mathbf{b}$, then $\mathbf{x}_1 = \mathbf{x}_2$, if and only if the row space of A is all of \mathbb{R}^n .

Since any function is invertible if and only if it is both onto and one-to-one, the linear transformation from \mathbb{R}^n to \mathbb{R}^m defined by an $m \times n$ matrix A is invertible if and only if the column space of A is all of \mathbb{R}^m , and the row space of A is all of \mathbb{R}^n . The next theorem says, among other things, that this can only happen when $m = n$.

Theorem 50 (Fundamental Theorem of Linear Algebra). *Let A be an $m \times n$ matrix. Then the dimension of the columns space of A equals the dimension of the row space of A .*

Theorem 50 is proved at the end of the next subsection; the proof is based on the Gram-Schmidt algorithm. First, we discuss what the theorem says.

Definition 55 (Rank of a matrix). *The rank of an $m \times n$, $\text{rank}(A)$, matrix A is the dimension of its column space, or, equivalently the dimension of its row space.*

By Theorem 50, the rank of an $m \times n$ matrix A can be no larger than $\min\{m, n\}$, and hence when $m \neq n$, it is impossible to have both that the row space is \mathbb{R}^n and the column space is \mathbb{R}^m . Thus, only square matrices; i.e., those with $m = n$, can ever represent invertible transformations. This is the negative message coming from Theorem 50.

The positive message is what makes the theorem fundamentally important: Let A be an $n \times n$ matrix. Suppose we know that $A\mathbf{x} = \mathbf{b}$ has a solution for every \mathbf{b} in \mathbb{R}^n . Then the column space of A is all of \mathbb{R}^n , and so the rank of A is n . But then by Theorem 50, the row space of A is all of \mathbb{R}^n , and hence $\text{Null}(A) = \mathbf{0}$. In short, if $A\mathbf{x} = \mathbf{b}$ has a solution for each \mathbf{b} , it automatically follows that this solutions is unique. In other words, if the linear transformation from \mathbb{R}^n to \mathbb{R}^n represented by A is onto, it is also one-to-one and hence invertible.

The same reasoning shows that if A is an $n \times n$ matrix, and the the linear transformation from \mathbb{R}^n to \mathbb{R}^n represented by A is one-to-one, then it is also onto, and hence invertible.

- To prove invertibility of a linear transformation \mathbf{f} from \mathbb{R}^n to \mathbb{R}^n , one need only show that it is either onto or one-to-one; the rest follows automatically. Moreover, this can be done by computing the rank of the matrix A that represents \mathbf{f} , and showing that the rank is n .

One way to compute the rank of a matrix A is to apply the Gram-Schmidt Algorithm to the columns of A to produce an orthonormal basis for the column space of A . The rank of A is the number r of vectors in this orthonormal basis. As we shall see in the next subsection, computing such an orthonormal basis tells us much more than the rank of A : It is the basis of a powerful method of solving $A\mathbf{x} = \mathbf{b}$ in general, and it is the basis of a simple proof of the Fundamental Theorem of Linear Algebra.

4.2.4 QR factorization

For special types of matrices A , it is very easy to solve $A\mathbf{x} = \mathbf{b}$. Here is one such case. Suppose $m = n$, and let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be any orthonormal basis of \mathbb{R}^n . Let Q be the $n \times n$ matrix $Q = [\mathbf{u}_1, \dots, \mathbf{u}_n]$. To explicitly solve $Q\mathbf{x} = \mathbf{b}$, all we have to do is to find numbers x_1, \dots, x_n such that $\mathbf{b} = \sum_{j=1}^n x_j \mathbf{u}_j$. Since $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is an orthonormal basis of \mathbb{R}^n , $\mathbf{b} = \sum_{j=1}^m (\mathbf{b} \cdot \mathbf{u}_j) \mathbf{u}_j$. Therefore, the unique solution to $Q\mathbf{x} = \mathbf{b}$ is the vector \mathbf{x} defined by

$$\mathbf{x} = (\mathbf{b} \cdot \mathbf{u}_1, \dots, \mathbf{b} \cdot \mathbf{u}_m) = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} \mathbf{b} . \quad (4.41)$$

Definition 56 (Transpose of a matrix). Let $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ be an $n \times n$ matrix. The transpose of A is the $n \times m$ matrix A^T given by $\begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_n \end{bmatrix}$.

Using this definition, we may write (4.41) as $\mathbf{x} = Q^T \mathbf{b}$. Summarizing, if Q is an $n \times n$ matrix with orthonormal columns, for all $\mathbf{b} \in \mathbb{R}^n$, $Q\mathbf{x} = \mathbf{b}$ has a unique solution which is $\mathbf{x} = Q^T \mathbf{b}$.

Example 72. Let $Q = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & -2 & -1 \\ 2 & 1 & 2 \end{bmatrix}$ and $\mathbf{b} = (1, 2, 3)$. As one readily checks, the columns of Q are orthonormal.

The unique solution of $Q\mathbf{x} = \mathbf{b}$ is then the vector

$$\mathbf{x} = Q^T \mathbf{b} = (\mathbf{b} \cdot \mathbf{u}_1, \mathbf{b} \cdot \mathbf{u}_2, \mathbf{b} \cdot \mathbf{u}_3) = \frac{1}{3}(11, 1, 2) .$$

Of course there is nothing special about the vector $(1, 2, 3)$.

More generally consider $m \times r$ matrix $Q = [\mathbf{u}_1, \dots, \mathbf{u}_r]$ with orthonormal columns. Since the columns belong to \mathbb{R}^m , $r \leq m$. If $r = m$ it is essentially the case we just considered, so suppose that $r < m$.

By definition, $Q^T Q = [Q^T \mathbf{u}_1, \dots, Q^T \mathbf{u}_r]$, and for each j , $Q^T \mathbf{u}_j = (\mathbf{u}_1 \cdot \mathbf{u}_j, \dots, \mathbf{u}_n \cdot \mathbf{u}_j) = \mathbf{e}_j$. That is,

$$Q^T Q = I_{r \times r}, \quad (4.42)$$

where $I_{r \times r}$ is the $r \times r$ identity matrix.

The identity (4.42) implies that $\text{Null}(Q) = \{\mathbf{0}\}$: If $Q\mathbf{x} = \mathbf{0}$, then $\mathbf{x} = Q^T Q\mathbf{x} = Q^T \mathbf{0} = \mathbf{0}$. In fact, more is true: If $\mathbf{x} = (x_1, \dots, x_r)$, $Q\mathbf{x} = \sum_{j=1}^r x_j \mathbf{u}_j$, and by the Pythagorean Theorem,

$$\|Q\mathbf{x}\|^2 = \left\| \sum_{j=1}^r x_j \mathbf{u}_j \right\|^2 = \sum_{j=1}^r x_j^2 = \|\mathbf{x}\|^2.$$

We have proved:

Lemma 12. *Let Q be an $m \times r$ matrix whose columns are orthonormal. Then (4.42) is valid, and for all $\mathbf{x} \in \mathbb{R}^r$, $\|Q\mathbf{x}\| = \|\mathbf{x}\|$. In particular, $\text{Null}(Q) = \{\mathbf{0}\}$.*

Matrices with orthonormal columns are therefore very special, and square ($n \times n$) matrices with this property are more special still. Therefore, the fact that we can easily solve $Q\mathbf{x} = \mathbf{b}$ when Q is an $n \times n$ matrix with orthonormal columns may appear to be an exceptional curiosity. It is not.

Let $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ be any $m \times n$ matrix. By Theorem 47, the set of vectors \mathbf{b} for which $A\mathbf{x} = \mathbf{b}$ has a solution is precisely the span of the columns of A , $\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_n\})$.

Applying the Gram-Schmidt Algorithm to $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ produces an orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ of r vectors, and we have seen that $r \leq \min\{m, n\}$. By Theorem 16,

$$\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_n\}) = \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\}). \quad (4.43)$$

Therefore, for each $1 \leq j \leq n$, $\mathbf{v}_j \in \text{Span}(\{\mathbf{u}_1, \dots, \mathbf{u}_r\})$ and hence there are numbers $R_{i,j}$, $1 \leq i \leq r$, $1 \leq j \leq n$, such that

$$\mathbf{v}_j = \sum_{i=1}^r R_{i,j} \mathbf{u}_i. \quad (4.44)$$

Theorem 6 gives us an explicit formula for $R_{k,j}$, namely

$$R_{i,j} = \mathbf{u}_i \cdot \mathbf{v}_j.$$

Define R to be the $r \times n$ matrix whose i, j entry is $R_{i,j} = \mathbf{u}_i \cdot \mathbf{v}_j$. Define Q to be the $m \times r$ matrix whose k th column is \mathbf{u}_k . Then since $(R_{1,j}, \dots, R_{r,j})$ is the j th column of R , (4.44) says that \mathbf{v}_j is the j th column of QR . That is, $A = QR$. This is the *QR factorization of A* .

It is extremely important because A is a general $m \times n$ matrix, while Q and R are very special: The columns of Q are orthonormal, while R has a “staircase structure”, which makes R another particularly simple kind of matrix to deal with.

Example 73 (QR factorization). Consider the 3×3 matrix $A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 2 & -3 \\ 2 & 5 & 6 \end{bmatrix}$. Write this in the

form $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$. Apply the Gram-Schmidt algorithm to $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ to produce $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$,

and define $Q = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$. Doing the computations, one finds $Q = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & -2 & -1 \\ 2 & 1 & 2 \end{bmatrix}$.

It is now a simple matter to compute

$$R = \begin{bmatrix} \mathbf{u}_1 \cdot \mathbf{v}_1 & \mathbf{u}_1 \cdot \mathbf{v}_2 & \mathbf{u}_1 \cdot \mathbf{v}_3 \\ \mathbf{u}_2 \cdot \mathbf{v}_1 & \mathbf{u}_2 \cdot \mathbf{v}_2 & \mathbf{u}_2 \cdot \mathbf{v}_3 \\ \mathbf{u}_3 \cdot \mathbf{v}_1 & \mathbf{u}_3 \cdot \mathbf{v}_2 & \mathbf{u}_3 \cdot \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 3 \\ 0 & 3 & 6 \\ 0 & 0 & 3 \end{bmatrix}.$$

You can now easily check that $A = QR$, though we have already proved this to be true.

Notice that the matrix R is “upper triangular”, meaning that all of the entries below the main diagonal are zero.

This is not an accident. By the very nature of the Gram-Schmidt Algorithm, \mathbf{v}_1 is a linear combination of \mathbf{u}_1 alone, \mathbf{v}_2 is a linear combinations of only \mathbf{u}_1 and \mathbf{u}_2 . This gives the coefficient matrix R its triangular structure,

The upper triangular matrix with positive diagonal entries that we encountered in the previous example belongs to a general class of simple matrices that we now introduce.

Definition 57 (Echelon form). An $m \times n$ is in echelon form in case the first non-zero entry in each row lies strictly to the right of the first non zero entry in the row just above it for all rows after the first row. The first non-zero entries in each row are called the pivotal entries.

Here is a schematic picture of a 4×7 matrix in echelon form. In the schematic, a \bullet denotes an entry that is definitely non-zero, and an $*$ denotes an entry that may be zero or non zero.

$$\begin{bmatrix} 0 & \bullet & * & * & * & * & * \\ 0 & 0 & \bullet & * & * & * & * \\ 0 & 0 & 0 & 0 & \bullet & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (4.45)$$

Notice that the requirement that the first non-zero entry in each row lies strictly to the right of the first non zero entry in the row just above it gives such matrices a “staircase” structure, with steps occurring at each pivotal entry.

Lemma 13. Let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be any set of n vectors in \mathbb{R}^m . Suppose that the Gram-Schmidt Algorithm applied to $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ yields an orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ of r vectors. Define the $r \times n$ matrix R by $R_{i,j} = \mathbf{u}_i \cdot \mathbf{v}_j$ for all $i \leq i \leq r$ and all $1 \leq j \leq n$. Then R is in echelon form with no zero rows and all pivotal entries are positive.

Proof. For each $1 \leq i \leq r$, define $j(i)$ to be the least value of j such that applying the Gram-Schmidt Algorithm to $\{\mathbf{v}_1, \dots, \mathbf{v}_j\}$ produces a set of i vectors. Then, by definition, the vector $\mathbf{v}_{j(i)}$ is not discarded, and

$$\mathbf{0} \neq \mathbf{w}_i = \mathbf{v}_{j(i)} - \sum_{k=1}^{i-1} (\mathbf{v}_{j(i)} \cdot \mathbf{u}_k) \mathbf{u}_k.$$

Since $\mathbf{w}_i = \|\mathbf{w}_i\| \mathbf{u}_i$, taking the dot product of both sides with \mathbf{u}_i yields $\|\mathbf{w}_i\| = \mathbf{u}_i \cdot \mathbf{v}_{j(i)}$. That is $R_{i,j(i)} = \mathbf{u}_i \cdot \mathbf{v}_{j(i)} > 0$. Next, for $j < j(i)$, by Theorem 16, \mathbf{v}_j is a linear combination of $\{\mathbf{u}_1, \mathbf{u}_{j(i)-1}\}$, and hence is orthogonal to \mathbf{u}_i . That is $R_{i,j} = 0$ for all $j < j(i)$.

We have proven so far that the first non-zero entry in the i th row of R occurs in the $j(i)$ th column. It remains to observe that, by its very definition, $j(i)$ is a strictly increasing function of i . Therefore, the first non-zero entry in each row after the first lies strictly to the right of the first non zero entry in the row just above. \square

Theorem 51 (*QR factorization*). *Let A be an $m \times n$ matrix, and let r be dimension of the column space of A . Then there exist an $m \times r$ matrix Q whose columns are orthonormal, and an $r \times n$ matrix in echelon form such that $A = QR$.*

Proof. The proof is simply a recapitulation of the discussion above, and it tells us how to compute Q and R . Let $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$. Apply the Gram-Schmidt Algorithm to the columns of A to produce $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$. Define $Q = [\mathbf{u}_1, \dots, \mathbf{u}_r]$. Next, define R by $R_{i,j} = \mathbf{u}_i \cdot \mathbf{v}_j$. Then for each $j = 1, \dots, n$,

$$\mathbf{v}_j = \sum_{i=1}^r (\mathbf{u}_i \cdot \mathbf{v}_j) \mathbf{u}_i = \sum_{i=1}^r R_{i,j} \mathbf{u}_i .$$

The right hand side is simply the matrix-vector product of Q and $(R_{1,j}, \dots, R_{r,j})$, which is the j th column of R . Hence, with \mathbf{r}_j denoting the j th column of R , we have $[\mathbf{v}_1, \dots, \mathbf{v}_n] = [Q\mathbf{r}_1, \dots, Q\mathbf{r}_n]$ which means that $A = QR$. Finally, by Lemma 13, R is in echelon form. \square

Example 74 (Computing and using a *QR* factorization). Consider the 3×4 matrix $A = \begin{bmatrix} 1 & 4 & 3 & 3 \\ 2 & 2 & 0 & -3 \\ 2 & 5 & 3 & 6 \end{bmatrix}$.

Apply the Gram-Schmidt algorithm to the set of columns $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$. The result is an orthonormal set of 3 vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, and defining $Q = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$ we have $Q = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & -2 & -1 \\ 2 & 1 & 2 \end{bmatrix}$.

It is now a simple matter to compute

$$R = \begin{bmatrix} \mathbf{u}_1 \cdot \mathbf{v}_1 & \mathbf{u}_1 \cdot \mathbf{v}_2 & \mathbf{u}_1 \cdot \mathbf{v}_3 & \mathbf{u}_1 \cdot \mathbf{v}_4 \\ \mathbf{u}_2 \cdot \mathbf{v}_1 & \mathbf{u}_2 \cdot \mathbf{v}_2 & \mathbf{u}_2 \cdot \mathbf{v}_3 & \mathbf{u}_2 \cdot \mathbf{v}_4 \\ \mathbf{u}_3 \cdot \mathbf{v}_1 & \mathbf{u}_3 \cdot \mathbf{v}_2 & \mathbf{u}_3 \cdot \mathbf{v}_3 & \mathbf{u}_3 \cdot \mathbf{v}_4 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 3 & 3 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 0 & 3 \end{bmatrix} .$$

You can now easily check that $A = QR$, though we have already proved this to be true.

The computation shows that $\text{rank}(A) = 3$, and hence $A\mathbf{x} = \mathbf{b}$ has a solution for every \mathbf{b} in \mathbb{R}^3 , and in fact infinitely many of them since $\text{Null}(A)$ is a one dimensional subspace of \mathbb{R}^4 .

We can easily solve for these solutions using a 2 step procedure. Since $Q^T Q = I_{3 \times 3}$, if $A\mathbf{x} = \mathbf{b}$, then $Q^T A\mathbf{x} = Q^T \mathbf{b}$, but $Q^T A\mathbf{x} = Q^T Q R\mathbf{x} = R\mathbf{x}$, and hence $R\mathbf{x} = Q^T \mathbf{b}$. Because of the echelon form of R , we can easily find all solutions by back substitution. To carry this out, compute

$$\mathbf{y} := Q^T \mathbf{b} = (\mathbf{u}_1 \cdot \mathbf{b}, \mathbf{u}_2 \cdot \mathbf{b}, \mathbf{u}_3 \cdot \mathbf{b}) ,$$

For the specific choice $\mathbf{b} = (1, 2, 3)$, we find that $\mathbf{y} = \frac{1}{3}(11, 1, 2)$.

Now let us find all solutions $\mathbf{x} = (x, y, z, w)$ of $R\mathbf{x} = \mathbf{y}$. This is equivalent to the system

$$3x + 6y + 3z + 3w = 11/3$$

$$3z + 6w = 1/3$$

$$3w = 2/3$$

Because of the structure of this system – a direct consequence of the fact that R is in echelon form – we immediately see from the last equation that $w = 2/9$, and then from the second equation that $z = 1$. Using these values in the first equation, it reduces to $3x + 6y = 7/3$. Solving for x , we find $x = 7/9 - 2y$. Hence the general solution of $R(x, y, z, w) = \frac{1}{3}(11, 1, 2)$ is

$$\mathbf{x}(y) = (7/9 - 2y, y, 1, 1/9).$$

We have a one parameter family of solutions, parametrized by the “free” or undetermined variable y . We could have also solved for y in terms of x , and kept x as the parameter, but the choice we made, y taking as the free variable the variable corresponding to a column with no pivot, has some advantages as we shall see.

Example 75 (QR factorization and $A\mathbf{x} = \mathbf{0}$). Again consider the 3×4 matrix $A = \begin{bmatrix} 1 & 4 & 3 & 3 \\ 2 & 2 & 0 & -3 \\ 2 & 5 & 3 & 6 \end{bmatrix}$. As we saw in the previous example, the QR factorization of A is given by $Q = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & -2 & -1 \\ 2 & 1 & 2 \end{bmatrix}$

$$\text{and } R = \begin{bmatrix} 3 & 6 & 3 & 3 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

Suppose that \mathbf{x} is any solution of $A\mathbf{x} = \mathbf{0}$. Then $Q(R\mathbf{x}) = \mathbf{0}$. Since the columns of Q are orthonormal, and hence linearly independent, this is possible only in case $R\mathbf{x} = \mathbf{0}$. Hence it suffices to find all solutions $\mathbf{x} = (x, y, z, w)$ of this latter equation. This is equivalent to the system

$$\begin{aligned} 3x + 6y + 3z + 3w &= 0 \\ 3z + 6w &= 0 \\ 3w &= 0 \end{aligned}$$

The last two equations tell us that $w = z = 0$, and then the first reduces to $x + 2y = 0$, which means that $x = -2y$. Thus, the general solution is

$$\mathbf{x}(y) = (-2y, y, 0, 0) = y(-2, 1, 0, 0).$$

That is, the set of all solutions of $A\mathbf{x} = \mathbf{0}$ is precisely the set of all multiples of the vector \mathbf{x}_0 given by $\mathbf{x}_0 = (-2, 1, 0, 0)$. Notice that from the row vector formula for matrix-vector multiplication, $A\mathbf{x}_0 = \mathbf{0}$ is equivalent to the statement that \mathbf{x}_0 is orthogonal to each of the rows of A .

We close this subsection by proving the Fundamental Theorem of Linear Algebra.

Proof of Theorem 50. Let A be an $m \times n$ matrix, and let $A = QR$ be its QR factorization. By construction, the columns of Q are an orthonormal basis for $\text{Ran}(A)$, so if Q is an $m \times r$ matrix, the column rank of A is r .

Next we claim that $\text{Null}(A) = \text{Null}(R)$. To see this, suppose that $A\mathbf{z} = \mathbf{0}$, and define $\mathbf{y} := R\mathbf{z}$. Then $Q\mathbf{y} = A\mathbf{z} = \mathbf{0}$. By Lemma 12, $\|Q\mathbf{y}\| = \|R\mathbf{z}\|$, and hence $\mathbf{y} = \mathbf{0}$. Evidently if $R\mathbf{z} = \mathbf{0}$, $A\mathbf{z} = QR\mathbf{z} = A\mathbf{0} = \mathbf{0}$. This proves the claim.

By Theorem 19 and Theorem 48 together, the row space of a matrix is the orthogonal complement of its null space. Therefore, since A and R have the same null space, they have the same row space. That is, the span of the rows of R is equal to the span of the rows of A . But from the echelon structure of R , it is evident that applying the Gram-Schmidt algorithm from the bottom up that each of the r rows yields a new unit vector, and hence the algorithm yields a set of r vectors in \mathbb{R}^n . Hence the row rank of A is also r . \square

Definition 58 (Linear independence). *A set $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of n vectors in \mathbb{R}^m is linearly independent in case $\sum_{j=1}^n x_j \mathbf{v}_j = \mathbf{0}$ if and only if $x_j = 0$ for all j .*

It follows immediately from the definition that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is linearly independent in \mathbb{R}^m if and only if the matrix $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ satisfies $\text{Null}(A) = \{\mathbf{0}\}$, and hence the rows of A span all of \mathbb{R}^n , so that, by the Fundamental Theorem of Linear Algebra, the columns of A also span \mathbb{R}^n .

Conversely, by Theorem 47, $\text{Span}(\{\mathbf{v}_1, \dots, \mathbf{v}_n\}) = \mathbb{R}^n$ if and only if the column rank of A is n , and then by the Fundamental Theorem of Linear Algebra, the row rank of A is also n , and hence $\text{Null}(A) = \{\mathbf{0}\}$, which is the same as saying that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ spans \mathbb{R}^n . This proves:

Theorem 52. *A set $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of n vectors in \mathbb{R}^n is linearly independent if and only if it spans \mathbb{R}^n .*

Remark 4. *The QR factorization method is important because it is the basis of numerically stable methods of computation for large matrices, and because it leads to an easy proof of the Fundamental Theorem of Linear Algebra. However, using it with pencil and paper is often laborious since complicated square roots typically arise when doing Gram-Schmidt on the columns of a matrix – our examples were carefully prepared to avoid this. Other methods that work on the rows of a matrix are developed in the exercises. In the next section we will deduce formulas for matrix inverses, using the Fundamental Theorem of Linear Algebra, and with these formulas we will be able to easily solve many of the equations of the form $A\mathbf{x} = \mathbf{b}$ that arise in the next chapters.*

4.2.5 Matrix inverses

Theorem 53. *Let \mathbf{f} be an invertible linear transformation from \mathbb{R}^n to \mathbb{R}^n . Then its inverse transformation \mathbf{f}^{-1} is also linear.*

Proof. Suppose \mathbf{f} is an invertible linear transformation. For any numbers a, b and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{f}^{-1}(a\mathbf{x} + b\mathbf{y})$ is the unique vector \mathbf{z} such that

$$\mathbf{f}(\mathbf{z}) = a\mathbf{x} + b\mathbf{y}. \quad (4.46)$$

But since \mathbf{f} is linear, $\mathbf{f}(a\mathbf{f}^{-1}(\mathbf{x}) + b\mathbf{f}^{-1}(\mathbf{y})) = a\mathbf{f}(\mathbf{f}^{-1}(\mathbf{x})) + b\mathbf{f}(\mathbf{f}^{-1}(\mathbf{y})) = a\mathbf{x} + b\mathbf{y}$. Thus, $\mathbf{z} = a\mathbf{f}^{-1}(\mathbf{x}) + b\mathbf{f}^{-1}(\mathbf{y})$ solves (4.46), and hence $\mathbf{f}^{-1}(a\mathbf{x} + b\mathbf{y}) = a\mathbf{f}^{-1}(\mathbf{x}) + b\mathbf{f}^{-1}(\mathbf{y})$. This proves that \mathbf{f}^{-1} is linear. \square

Now let $A := [\mathbf{v}_1, \dots, \mathbf{v}_n]$ be the matrix of an invertible linear transformation \mathbf{f} from \mathbb{R}^n to \mathbb{R}^n . Let B be the $n \times n$ matrix representation of its inverse. (This makes sense since the inverse is linear,

so it has a matrix). Write B in its row form: $B = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{bmatrix}$.

Since for each $j = 1, \dots, n$, $\mathbf{v}_j = A\mathbf{e}_j = \mathbf{f}(\mathbf{e}_j)$, it follows that $\mathbf{f}^{-1}(\mathbf{v}_j) = B\mathbf{v}_j = \mathbf{e}_j$. By Theorem 44, it follows that $B\mathbf{v}_j = (\mathbf{w}_1 \cdot \mathbf{v}_j, \dots, \mathbf{w}_n \cdot \mathbf{v}_j) = \mathbf{e}_j$. That is, we must have:

$$\mathbf{w}_i \cdot \mathbf{v}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \quad (4.47)$$

This motivates the following theorem:

Theorem 54. *Let $A := [\mathbf{v}_1, \dots, \mathbf{v}_n]$ be the matrix of a linear transformation \mathbf{f} from \mathbb{R}^n to \mathbb{R}^n . Then \mathbf{f} is invertible if and only if there exists a set $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ of n vectors in \mathbb{R}^n such that (4.47) is satisfied. In this case, the matrix of \mathbf{f}^{-1} is given by $\begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{bmatrix}$, and the set $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ is unique.*

Proof. By what we have explained above, when \mathbf{f} is invertible, there does exist such a set $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$. Since $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ gives a formula for the inverse of \mathbf{f} , and since there is only one inverse of \mathbf{f} , there is exactly one such set $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$.

Conversely, suppose that such a set $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ exists. Define a function \mathbf{g} from \mathbb{R}^n to \mathbb{R}^n by

$$\mathbf{g}(\mathbf{z}) := (\mathbf{w}_1 \cdot \mathbf{z}, \dots, \mathbf{w}_n \cdot \mathbf{z}).$$

Since by definition, $\mathbf{f}(\mathbf{x}) = \sum_{j=1}^n x_j \mathbf{v}_j$,

$$\begin{aligned} \mathbf{g}(\mathbf{f}(\mathbf{x})) &= (\mathbf{w}_1 \cdot \sum_{j=1}^n x_j \mathbf{v}_j, \dots, \mathbf{w}_n \cdot \sum_{j=1}^n x_j \mathbf{v}_j) \\ &= (\sum_{j=1}^n x_j (\mathbf{w}_1 \cdot \mathbf{v}_j), \dots, \sum_{j=1}^n x_j (\mathbf{w}_n \cdot \mathbf{v}_j)) \\ &= (x_1, \dots, x_n) = \mathbf{x}, \end{aligned} \quad (4.48)$$

where in the last line we have used (4.47).

This shows that \mathbf{f} is one to one, since if $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{y})$, then $\mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{g}(\mathbf{f}(\mathbf{y}))$, and then by (4.48), $\mathbf{x} = \mathbf{y}$. By Theorem 50, \mathbf{f} is invertible. Now that we know that \mathbf{f} is invertible, (4.48) shows that \mathbf{g} is the inverse of \mathbf{f} , and that the matrix representation of \mathbf{g} is what it was claimed to be. \square

Before proceeding to the applications, we highlight an important point. It may seem that once we have derived (4.48), this alone proves that \mathbf{g} is the inverse of \mathbf{f} , and have no need to invoke Theorem 50. *This is not the case.*

Indeed, consider the two functions f and g from the set \mathbb{N} of the natural numbers into itself that are given by

$$f(n) = n + 1 \quad \text{and} \quad g(n) = \begin{cases} 1 & n = 1 \\ n - 1 & n > 1 \end{cases}. \quad (4.49)$$

Then $g(f(n)) = n$ for all $n \in \mathbb{N}$, but neither f nor g is invertible. Indeed, there is no $n \in \mathbb{N}$ with $f(n) = 1$, so that f does not transform \mathbb{N} onto \mathbb{N} . Also, $g(1) = g(2) = 1$, so that g is not a one-to-one transformation of \mathbb{N} into \mathbb{N} , though it does transform \mathbb{N} onto \mathbb{N} .

The pathology seen in this example does not occur for transformations of finite sets into themselves. Let X denote the finite set $X := \{1, \dots, N\}$ where N is some positive integer. Let f be a function defined on X with values in X ; i.e., a transformation from X to X . It is not hard to show that if f is one-to-one, then f is necessarily onto, and that if f is onto, then f is necessarily one-to-one. This is left as an exercise.

Thus, for a transformation f on a finite set, to check for invertibility, it suffices to check *either* whether f is one-to-one, or whether f is onto. For transformations on infinite sets, such as \mathbb{R}^3 , this is not necessarily the case, as the example (4.49) shows. However, linear transformations are very special, and for them we have Theorem 50.

Now let us apply Theorem 54. We begin with $n = 2$. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{so that} \quad A = [\mathbf{v}_1, \mathbf{v}_2],$$

where the columns of A are $\mathbf{v}_1 = (a, c)$ and $\mathbf{v}_2 = (b, d)$.

First observe that if $\mathbf{v}_1 = \mathbf{0}$, then $A\mathbf{e}_1 = A\mathbf{0} = \mathbf{0}$, so that A is not one-to-one, and therefore not invertible. Similar considerations apply to \mathbf{v}_2 , and so if A is invertible, neither column is the zero vector. Therefore, in trying to find an inversion formula for A , we may suppose neither column of A is the zero vector.

In order to have $\mathbf{w}_1 \cdot \mathbf{v}_2 = 0$ and $\mathbf{w}_2 \cdot \mathbf{v}_1 = 0$, we must have that \mathbf{w}_1 and \mathbf{w}_2 are multiples of \mathbf{v}_2^\perp and \mathbf{v}_1^\perp , respectively:

$$\mathbf{w}_1 = \alpha(-d, b) \quad \text{and} \quad \mathbf{w}_2 = \beta(-c, a).$$

We then compute

$$\mathbf{w}_1 \cdot \mathbf{v}_1 = \alpha(bc - ad) \quad \text{and} \quad \mathbf{w}_2 \cdot \mathbf{v}_2 = \beta(ad - bc).$$

Thus, it is possible to achieve (4.47) if and only if $ad - bc \neq 0$, and in this case we have

$$\mathbf{w}_1 = \frac{1}{ad - bc}(d, -b) \quad \text{and} \quad \mathbf{w}_2 = \frac{1}{ad - bc}(-c, a). \tag{4.50}$$

Now applying Theorem 54, we conclude:

Theorem 55. *Let \mathbf{f} be the linear transformation from \mathbb{R}^2 to \mathbb{R}^2 whose matrix is*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Then \mathbf{f} is invertible if and only if $ad - bc \neq 0$, and in this case the inverse transformation has the matrix

$$A^{-1} := \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

We now turn to $n = 3$. Let $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$. As before, if A is invertible no column is the zero vector. But even more is true: Suppose \mathbf{v}_2 is a multiple of \mathbf{v}_1 ; say $\mathbf{v}_2 = t\mathbf{v}_1$ for some $t \in \mathbb{R}$. But then

$$A(\mathbf{e}_2 - t\mathbf{e}_1) = A\mathbf{e}_2 - tA\mathbf{e}_1 = \mathbf{v}_2 - t\mathbf{v}_1 = \mathbf{0} = A\mathbf{0}$$

and since $\mathbf{e}_2 - t\mathbf{e}_1 = (-t, 1, 0) \neq (0, 0, 0)$, the transformation sending \mathbf{x} to $A\mathbf{x}$ is not one to one, and hence is not invertible. It follows that if A is invertible, then no column of A can be a multiple of any other column of A , or, what is the same thing,

$$\mathbf{v}_i \times \mathbf{v}_j \neq \mathbf{0} \quad \text{for each } 1 \leq i < j \leq 3. \quad (4.51)$$

Therefore, let us assume (4.51), and try to find an inversion formula for A .

Theorem 54 tells us what to look for: We first seek a vector \mathbf{w}_1 that is orthogonal to \mathbf{v}_2 and \mathbf{v}_3 , and also such that $\mathbf{w}_1 \cdot \mathbf{v}_1 = 1$. Because of (4.51), to achieve the orthogonality, we *must* choose \mathbf{w}_1 to be a multiple of $\mathbf{v}_2 \times \mathbf{v}_3$. We can then achieve $\mathbf{w}_1 \cdot \mathbf{v}_1 = 1$ if and only if $\mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3 \neq 0$, in which case we must choose

$$\mathbf{w}_1 = \frac{1}{\mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3} \mathbf{v}_2 \times \mathbf{v}_3.$$

This determines \mathbf{w}_1 . The same sort of reasoning determines \mathbf{w}_2 and \mathbf{w}_3 . At first it might appear that one gets different normalization factor each time, but by the properties of the triple product,

$$\mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3 = \mathbf{v}_2 \cdot \mathbf{v}_3 \times \mathbf{v}_1 = \mathbf{v}_3 \cdot \mathbf{v}_1 \times \mathbf{v}_2,$$

and hence one divides by the same quantity in each case. We arrive at:

Theorem 56. *Let \mathbf{f} be the linear transformation from \mathbb{R}^3 to \mathbb{R}^3 whose matrix is $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$. Then \mathbf{f} is invertible if and only if $\mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3 \neq 0$, and in this case the inverse transformation has the matrix*

$$A^{-1} := \frac{1}{\mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3} \begin{bmatrix} \mathbf{v}_2 \times \mathbf{v}_3 \\ \mathbf{v}_3 \times \mathbf{v}_1 \\ \mathbf{v}_1 \times \mathbf{v}_2 \end{bmatrix}.$$

The matrix A^{-1} corresponding to the inverse of the transformation $\mathbf{x} \mapsto A\mathbf{x}$ is called the *matrix inverse of A* .

The two theorems we have just proved bring us to the following definition:

Definition 59 (Determinants of 2×2 and 3×3 matrices). *Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then the determinant of A , $\det(A)$, is defined by $\det(A) := ad - bc$. Let $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ be a 3×3 matrix, specified as a list of its three column vectors in \mathbb{R}^3 . Then the determinant of A , $\det(A)$, is defined by $\det(A) := \mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3$.*

The nomenclature is justified by the fact that if A is a 2×2 or a 3×3 matrix, then whether the corresponding linear transformation $\mathbf{f}(\mathbf{x}) := A\mathbf{x}$ is invertible or not is determined by whether $\det(A)$ differs from zero or not. We shall later define the determinant function on $n \times n$ matrices for all n so that the corresponding statement is true.

We close with one final inverse formula for an important special case. Let $Q = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ be any $n \times n$ matrix with orthonormal columns. Then the i, j entry of $Q^T A$ is

$$[Q^T Q]_{i,j} = (\text{row } i \text{ of } Q^T) \cdot (\text{column } j \text{ of } Q) = \mathbf{u}_i \cdot \mathbf{u}_j = [I_{n \times n}]_{i,j}.$$

That is, $Q^T Q = I_{n \times n}$. Since the column space of Q is \mathbb{R}^n , and since the column space of Q is the row space of Q^T , it follows that the rank of Q^T is n , and hence Q^T is invertible. Then the formula we have just proved shows that the unique solution of $Q^T \mathbf{x} = \mathbf{b}$ is $\mathbf{x} = Q^T \mathbf{b}$ since

$$Q^T(Q\mathbf{b}) = (Q^T Q)\mathbf{b} = I_{n \times n}\mathbf{b} = \mathbf{b}.$$

In other words, Q is the inverse of Q^T – and then Q^T is the inverse of Q . We conclude that if Q is any $n \times n$ matrix with orthonormal columns, then Q^T and Q represent linear transformation that are inverse to one another.

Definition 60 (Orthogonal matrix). *An $n \times n$ matrix Q is an orthogonal matrix in case its columns are orthonormal.*

Theorem 57. *Q is an orthogonal matrix if and only if $Q^{-1} = Q^T$.*

Proof. We have seen above that if Q is orthogonal, then Q is invertible and $Q^{-1} = Q^T$. Now let $Q = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ and suppose that $Q^T Q = I_{n \times n}$. Then the i, j entry of $Q^T Q$ is $\mathbf{u}_i \cdot \mathbf{u}_j$ and the i, j entry of $I_{n \times n}$ is 1 if $i = j$ and 0 otherwise. Thus, $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is orthonormal. \square

Another simple formula concerning the transpose, that we will often use later on, can be combined with Theorem 57 to produce a remarkable conclusion.

Theorem 58. *Let A be an $m \times n$ matrix. Then for all $\mathbf{x} \in \mathbb{R}^n$ and all $\mathbf{y} \in \mathbb{R}^m$,*

$$\mathbf{y} \cdot A\mathbf{x} = (A^T \mathbf{y}) \cdot \mathbf{x}. \quad (4.52)$$

Proof. Let $A = [\mathbf{a}_1, \dots, \mathbf{a}_n]$. Then $A^T \mathbf{y} = (\mathbf{a}_1 \cdot \mathbf{y}, \dots, \mathbf{a}_n \cdot \mathbf{y})$, so that

$$(A^T \mathbf{y}) \cdot \mathbf{x} = \sum_{j=1}^n (\mathbf{a}_j \cdot \mathbf{y}) x_j = \mathbf{y} \cdot \left(\sum_{j=1}^n x_j \mathbf{a}_j \right) = \mathbf{y} \cdot A\mathbf{x}$$

where we have used both the row and column formulas for matrix-vector multiplication. \square

Corollary 6. *The columns of an $n \times n$ matrix A are orthonormal if and only if the rows of A are orthonormal.*

Proof. Suppose the columns of A are orthonormal. Then A is an orthogonal matrix, and by Theorem 57, A is invertible, and $A^{-1} = A^T$, and hence $(A^T)^{-1} = A$. Let $A^T = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ so that for $j = 1, \dots, n$, $\mathbf{v}_j = A^T \mathbf{e}_j$. Then for $1 \leq i, j \leq n$, since $AA^T = I_{n \times n}$,

$$\mathbf{v}_i \cdot \mathbf{v}_j = A^T \mathbf{e}_i \cdot A^T \mathbf{e}_j = \mathbf{e}_i \cdot AA^T \mathbf{e}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Hence the columns of A^T , which are the rows of A are orthonormal.

Now suppose that the rows of A are orthonormal. Then A^T is orthogonal, and then by the first part, the columns of $(A^T)^T$ are orthonormal. But $(A^T)^T = A$. \square

4.2.6 Continuity of matrix inverses

Approximation and convergence are basic notions in analysis. In this subsection, we introduce a simple but useful notion of the distance between two $m \times n$ matrices, and hence a notion of continuity of matrix valued functions. We then discuss continuity of the matrix inverse function.

Definition 61 (Frobenius norm of a matrix and Frobenius distance). *Let A be an $m \times n$ matrix.*

The non-negative number $\|A\|_F$ defined by

$$\|A\|_F = \left(\sum_{j=1}^n \sum_{i=1}^m |A_{i,j}|^2 \right)^{1/2} \quad (4.53)$$

is called the Frobenius norm of the matrix A .

Notice that if $A = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_m \end{bmatrix} = [\mathbf{a}_1, \dots, \mathbf{a}_n]$, then

$$\|A\|_F^2 = \sum_{i=1}^m \|\mathbf{r}_i\|^2 = \sum_{j=1}^n \|\mathbf{a}_j\|^2 = \sum_{j=1}^n \|A\mathbf{e}_j\|^2. \quad (4.54)$$

Theorem 59. *Let A be an $m \times n$ matrix. Then for all $\mathbf{x} \in \mathbb{R}^n$,*

$$\|A\mathbf{x}\| \leq \|A\|_F \|\mathbf{x}\|. \quad (4.55)$$

Proof. Let $\mathbf{r}_1, \dots, \mathbf{r}_m$ denote the m rows of A . Then, $A\mathbf{x} = (\mathbf{r}_1 \cdot \mathbf{x}, \dots, \mathbf{r}_m \cdot \mathbf{x})$. Therefore,

$$\|A\mathbf{x}\|^2 = \sum_{i=1}^m (\mathbf{r}_i \cdot \mathbf{x})^2 \leq \sum_{i=1}^m (\|\mathbf{r}_i\| \|\mathbf{x}\|)^2 = \left(\sum_{i=1}^m \|\mathbf{r}_i\|^2 \right) \|\mathbf{x}\|^2,$$

from which (4.55) follows by (4.54). \square

Next, $A\mathbf{x} - A\mathbf{y} = A(\mathbf{x} - \mathbf{y})$ so that (4.55) implies $\|A\mathbf{x} - A\mathbf{y}\| \leq \|A\|_F \|\mathbf{x} - \mathbf{y}\|$. In particular, if \mathbf{f} is a linear transformation from \mathbb{R}^n to \mathbb{R}^m , and $A_{\mathbf{f}}$ is its corresponding matrix, then for any $\epsilon > 0$,

$$\|\mathbf{x} - \mathbf{y}\| < \frac{\epsilon}{\|A_{\mathbf{f}}\|_F} \Rightarrow \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < \epsilon.$$

While finding explicit values of δ to go with given values of ϵ can be a chore for nonlinear functions, it is simple in the linear case: simply compute $\|A_{\mathbf{f}}\|_F$.

The Frobenius distance between two $m \times n$ matrices A and B is the quantity $d_F(A, B)$ given by

$$d_F(A, B) = \|A - B\|_F.$$

Note that if we may identify an $m \times n$ matrix with a vector in \mathbb{R}^{mn} by the simple device of writing out the rows, one after another in a long vector \mathbf{v}_A . For example, for $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, we would have

$$\mathbf{v}_A = (1, 2, 3, 4, 5, 6).$$

Notice that the definition is set up so that $\|A\|_F = \|\mathbf{v}_A\|$. That is, the Frobenius norm of A is nothing other than the length of the vector one gets by “stretching A out” as a vector $\mathbf{v}_A \in \mathbb{R}^{nm}$

Notice also that because of the way matrix addition (and subtraction) is defined, for any two $m \times n$ matrices A and B , $\mathbf{v}_{A-B} = \mathbf{v}_A - \mathbf{v}_B$.

- It follows from this that the Frobenius distance d_F satisfies the triangle inequality

$$d_F(A, C) \leq d_F(A, B) + d_F(B, C)$$

for all $m \times n$ matrices A , B and C

Since it is clear that $d_F(A, B) = 0$ if and only if $A = B$, and that $d_F(A, B) = d_F(B, A)$, this shows that $d_F(A, B)$ is a metric on the set of $m \times n$ matrices. It is therefore correct to refer to it as a distance.

- The set of $m \times n$ matrices equipped with the Frobenius distance d_F therefore provides us with an example of a metric space – albeit one that is essentially just the Euclidean metric space \mathbb{R}^{mn} . Now we can not only do algebra with matrices: We can do analysis, and use limiting processes to solve problems.

Theorem 60 (The Frobenius norm and matrix multiplication). *Let A be an $m \times n$ matrix, and let B be an $n \times p$ matrix so that the matrix product AB is defined. Then*

$$\|AB\|_F \leq \|A\|_F \|B\|_F . \quad (4.56)$$

Proof. By (4.54) and Theorem 59,

$$\|AB\|_F^2 = \sum_{j=1}^n \|AB(\mathbf{e}_j)\|^2 \leq \|A\|_F^2 \left(\sum_{j=1}^n \|B(\mathbf{e}_j)\|^2 \right) = \|A\|_F^2 \|B\|_F^2 .$$

□

Theorem 61. *Let A be an $n \times n$ matrix such that $\|A - I_{n \times n}\|_F = r < 1$. Then A is invertible, and $\|A^{-1}\|_F \leq \sqrt{n}(1 - r)^{-1}$.*

More generally, let B be an invertible $n \times n$ matrix, and A is any $n \times n$ matrix such that $\|A - B\|_F = r\|B^{-1}\|_F^{-1}$, with $r < 1$, then A is invertible, and $\|A^{-1}\|_F \leq \sqrt{n}(1 - r)^{-1}\|B^{-1}\|_F$.

Proof. For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} = (\mathbf{x} - A\mathbf{x}) + A\mathbf{x}$, and then by the triangle inequality and Theorem 59,

$$\|\mathbf{x}\| \leq \|(A - I_{n \times n})\mathbf{x}\| + \|A\mathbf{x}\| \leq \|A - I_{n \times n}\|_F \|\mathbf{x}\| + \|A\mathbf{x}\| ,$$

so that

$$\|A\mathbf{x}\| \geq \|\mathbf{x}\| - r\|\mathbf{x}\| = (1 - r)\|\mathbf{x}\| . \quad (4.57)$$

In particular, $A\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. That is, $\text{Null}(A) = \{\mathbf{0}\}$, and then by the Fundamental Theorem of Linear Algebra, A is invertible. Now applying (4.57) with $\mathbf{x} = A^{-1}\mathbf{e}_j$, $j = 1, \dots, n$, we find

$$1 = \|\mathbf{e}_j\| = \|A(A^{-1}\mathbf{e}_j)\| \geq (1 - r)\|A^{-1}\mathbf{e}_j\| .$$

Therefore, $\|A^{-1}\|_F^2 = \sum_{j=1}^n \|A^{-1}\mathbf{e}_j\|^2 \leq \frac{n}{1-r}$. This proves the first part.

For the second, define $C := B^{-1}A$, and write $A - B = B(C - I_{n \times n})$, so that $C - I_{n \times n} = B^{-1}(A - B)$. By Theorem 60,

$$\|C - I_{n \times n}\|_F \leq \|B^{-1}\|_F \|A - B\|_F .$$

It follows that since $\|A - B\|_F = r\|B^{-1}\|_F^{-1}$ with $r < 1$, then $\|C - I_{n \times n}\|_F \leq r < 1$. By the first part C is then invertible. But since $A = BC$, A is the product of invertible matrices, and is itself invertible: $A^{-1} = C^{-1}B^{-1}$. By Theorem 60 again, $\|A^{-1}\|_F \leq \|C^{-1}\|_F \|B^{-1}\|_F$, and the bound on $\|C^{-1}\|_F$ from the first part completes the proof. \square

In what follows we will often be working with matrix valued functions: The *derivative* at \mathbf{x} of a differentiable function $\mathbf{f} = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$ from \mathbb{R}^n to \mathbb{R}^m is, as we discuss next, given by an $m \times n$ matrix $A(\mathbf{x})$ whose entries are partial derivatives of the functions $f_j(\mathbf{x})$.

Definition 62. Let $A(\mathbf{x})$ be an $m \times n$ matrix valued function on \mathbb{R}^p . Then A is continuous at $\mathbf{x}_0 \in \mathbb{R}^p$ in case for all $\epsilon > 0$, there is a $\delta > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow \|A(\mathbf{x}) - A(\mathbf{x}_0)\|_F < \epsilon .$$

There is nothing really new here; “stretching A out” to identify it with the vector $\mathbf{v}_A \in \mathbb{R}^{mn}$ as above, this is just the notion of continuity of vector valued functions in different notation. However, combining this definition with Theorem 61 does give us something new and important:

Theorem 62. Let A be a continuous $n \times n$ matrix valued function on \mathbb{R}^p . Suppose that $A(\mathbf{x}_0)$ is invertible. Then there is a $\delta > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow A(\mathbf{x}) \text{ is invertible} .$$

In particular, the set U on which $A(\mathbf{x})$ is invertible is open. Moreover, $A^{-1}(\mathbf{x})$ is a continuous function on U .

Proof. By the continuity of A , there is a $\delta > 0$ so that for $\|\mathbf{x} - \mathbf{x}_0\| < \delta$,

$$\|A(\mathbf{x}) - A(\mathbf{x}_0)\|_F < \frac{1}{2}\|A^{-1}(\mathbf{x}_0)\|_F^{-1} ,$$

and then by Theorem 61, for $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, $A(\mathbf{x})$ is invertible. Thus, the set U contains an open ball of some positive radius about each point in it, which means that U is open.

Moreover, Theorem 61 gives us the bound $\|A^{-1}(\mathbf{x})\|_F \leq 2\sqrt{n}\|A^{-1}(\mathbf{x}_0)\|_F$ for all \mathbf{x} such that $\|\mathbf{x} - \mathbf{x}_0\| < \delta$. We use this as follows: For any such \mathbf{x} , we have identity

$$A^{-1}(\mathbf{x}) - A^{-1}(\mathbf{x}_0) = A^{-1}(\mathbf{x}_0)[A(\mathbf{x}_0) - A(\mathbf{x})]A^{-1}(\mathbf{x}) .$$

By Theorem 60 and the bound $\|A^{-1}(\mathbf{x})\|_F \leq 2\sqrt{n}\|A^{-1}(\mathbf{x}_0)\|_F$,

$$\begin{aligned} \|A^{-1}(\mathbf{x}) - A^{-1}(\mathbf{x}_0)\|_F &\leq \|A^{-1}(\mathbf{x}_0)\|_F \|A(\mathbf{x}_0) - A(\mathbf{x})\|_F \|A^{-1}(\mathbf{x})\|_F \\ &\leq (2\sqrt{n}\|A^{-1}(\mathbf{x}_0)\|_F) \|A(\mathbf{x}_0) - A(\mathbf{x})\|_F . \end{aligned}$$

Since A is continuous at \mathbf{x}_0 , $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \|A(\mathbf{x}_0) - A(\mathbf{x})\|_F = 0$, and then $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \|A^{-1}(\mathbf{x}_0) - A^{-1}(\mathbf{x})\|_F = 0$, which means that A^{-1} is continuous at \mathbf{x}_0 . \square

4.3 Differentiability of functions from \mathbb{R}^n to \mathbb{R}^m

4.3.1 Differentiability and best linear approximation in several variables

Definition 63 (Differentiable functions from \mathbb{R}^n to \mathbb{R}^m). A function \mathbf{f} from \mathbb{R}^n to \mathbb{R}^m is differentiable at \mathbf{x}_0 in case there is some $m \times n$ matrix A from \mathbb{R}^n to \mathbb{R}^m such that for each $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta(\epsilon) \Rightarrow \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\| < \epsilon \|\mathbf{x} - \mathbf{x}_0\|. \quad (4.58)$$

An equivalent way to express (4.58), which “hides” the ϵ and δ in the definition of a limit, is that

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0. \quad (4.59)$$

In other words, \mathbf{f} is differentiable at \mathbf{x}_0 when for all \mathbf{x} “sufficiently close” to \mathbf{x}_0 , the deviation of \mathbf{f} from its value at \mathbf{x}_0 is given by $A(\mathbf{x} - \mathbf{x}_0)$ for some $m \times n$ matrix A , up to errors that are negligibly small compared to $\|\mathbf{x} - \mathbf{x}_0\|$, where we have chosen ϵ to be “negligible”. Thus, we have the *linear approximation*

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) \approx A(\mathbf{x} - \mathbf{x}_0). \quad (4.60)$$

This is a very important definition. We now carefully examine it to become completely familiar with it.

First of all, suppose A and B are two $m \times n$ matrices and for all $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ so that whenever $\|\mathbf{x} - \mathbf{x}_0\| < \delta(\epsilon)$, both

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\| < \epsilon \|\mathbf{x} - \mathbf{x}_0\| \quad \text{and} \quad \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - B(\mathbf{x} - \mathbf{x}_0)\| < \epsilon \|\mathbf{x} - \mathbf{x}_0\|$$

are true. Then, for such \mathbf{x} , we have from the triangle inequality,

$$\begin{aligned} \|(A - B)(\mathbf{x} - \mathbf{x}_0)\| &= \|[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - B(\mathbf{x} - \mathbf{x}_0)] - [\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)]\| \\ &\leq \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - B(\mathbf{x} - \mathbf{x}_0)\| + \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\| \\ &\leq 2\epsilon \|\mathbf{x} - \mathbf{x}_0\|. \end{aligned}$$

Next, for any j , if $0 < t < \delta(\epsilon)$, then $\mathbf{x} := \mathbf{x}_0 + t\mathbf{e}_j$ satisfies $\|\mathbf{x} - \mathbf{x}_0\| = t < \delta(\epsilon)$. Evaluating both sides above for this choice of \mathbf{x} , and canceling a factor of t , we obtain $\|(A - B)\mathbf{e}_j\| \leq \epsilon$. Since ϵ can be chosen arbitrarily small, we conclude $\|(A - B)\mathbf{e}_j\| = 0$ for each j . But this means that $A\mathbf{e}_j = B\mathbf{e}_j$ for each j , and hence $A = B$. In summary:

- There can be at most one $n \times m$ matrix A for which (4.58) is true for all $\epsilon > 0$.

The uniqueness of A that we have just proved justifies referring to the approximation in (4.58) as the *best linear approximation to \mathbf{f} at \mathbf{x}_0* : It is the only one for which (4.58) is true. Any other linear approximation entails errors that are not negligible at the first order in $\|\mathbf{x} - \mathbf{x}_0\|$. The uniqueness also allows us to make the following definition.

Definition 64 (Derivative and continuity of derivatives). Let \mathbf{f} be a function from \mathbb{R}^n to \mathbb{R}^m that is differentiable at some $\mathbf{x}_0 \in \mathbb{R}^n$. Then the unique $m \times n$ matrix A such that 4.58 is true for all $\epsilon > 0$

is called the derivative of \mathbf{f} at \mathbf{x}_0 . It is denoted by $[D\mathbf{f}(\mathbf{x}_0)]$, and is sometimes called the Jacobian matrix of \mathbf{f} at \mathbf{x}_0 . The corresponding linear transformation from \mathbb{R}^n to \mathbb{R}^m is simply denoted by $D\mathbf{f}(\mathbf{x}_0)$.

Furthermore, if \mathbf{f} is differentiable in an open set U , then we say that \mathbf{f} is continuously differentiable in U in case the function $\mathbf{x} \mapsto [D\mathbf{f}(\mathbf{x})]$ is continuous from U to the space of $m \times n$ matrices.

Remark 5. Recall that a vector valued function is continuous if and only if all of its entry functions are continuous. Since the Frobenius distance on the space of $m \times n$ matrices is essentially the Euclidean distance in \mathbb{R}^{mn} , a matrix valued function is continuous if and only if each of its entry functions is continuous. That is, $\mathbf{x} \mapsto [D\mathbf{f}(\mathbf{x})]$ is continuous on U if and only if $\mathbf{x} \mapsto [D\mathbf{f}(\mathbf{x})]_{i,j}$ is continuous from U to \mathbb{R} for each i, j .

So far we know that the derivative $[D\mathbf{f}(\mathbf{x}_0)]$ is unique when it exists. But how do we decide when it exists, and how do we compute it? Also, how is this notion of derivative related to the notions of directional derivatives and partial derivatives? To answer these questions, let us look again at the definition.

Let us pick some $t \neq 0$, some $1 \leq j \leq n$, and set $\mathbf{x} := \mathbf{x}_0 + t\mathbf{e}_j$. With this choice of \mathbf{x} ,

$$\begin{aligned} \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} &= \frac{\|\mathbf{f}(\mathbf{x}_0 + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x}_0) - tA\mathbf{e}_j\|}{|t|} \\ &= \left\| \frac{\mathbf{f}(\mathbf{x}_0 + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x}_0)}{t} - A\mathbf{e}_j \right\|. \end{aligned}$$

By the definition of differentiability, this must approach 0 as t approaches 0, and hence we must have

$$\lim_{t \rightarrow 0} \frac{\mathbf{f}(\mathbf{x}_0 + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x}_0)}{t} = A\mathbf{e}_j. \quad (4.61)$$

Therefore, if \mathbf{f} is differentiable at \mathbf{x}_0 , (4.61) gives us a formula for the j th column of $A = [D\mathbf{f}(\mathbf{x}_0)]$. Moreover, this tells us how to compute $A = [D\mathbf{f}(\mathbf{x}_0)]$ by partial differentiation. To do this, write $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$ so that $f_i(\mathbf{x}) = \mathbf{e}_i \cdot \mathbf{f}(\mathbf{x})$ for each $i = 1, \dots, n$. Taking the dot product of both sides of (4.61) by \mathbf{e}_i we obtain

$$A_{i,j} = \mathbf{e}_i \cdot A\mathbf{e}_j = \lim_{t \rightarrow 0} \frac{f_i(\mathbf{x}_0 + t\mathbf{e}_j) - f_i(\mathbf{x}_0)}{t} = \frac{\partial}{\partial x_j} f_i(\mathbf{x}_0).$$

We conclude that if \mathbf{f} is differentiable at \mathbf{x}_0 , then each of the partial derivatives

$$\frac{\partial}{\partial x_j} f_i(\mathbf{x}_0) \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$$

exist, and for each such i and j ,

$$[D\mathbf{f}(\mathbf{x}_0)]_{i,j} = \frac{\partial}{\partial x_j} f_i(\mathbf{x}_0).$$

There is a nice way to write this: Notice that the i th row of $[D\mathbf{f}(\mathbf{x}_0)]$ is $\nabla f_i(\mathbf{x}_0)$, so that

$$[D\mathbf{f}(\mathbf{x}_0)] = \begin{bmatrix} \nabla f_1(\mathbf{x}_0) \\ \vdots \\ \nabla f_m(\mathbf{x}_0) \end{bmatrix}. \quad (4.62)$$

One might hope that if all of the partial derivatives of \mathbf{f} exist at \mathbf{x}_0 , then \mathbf{f} would be differentiable at \mathbf{x}_0 . However, this is not the case. Fortunately, it is the case if one assumes only a little more:

Theorem 63 (Differentiability of \mathbf{f} and partial derivatives). *Let $\mathbf{f} = (f_1, \dots, f_m)$ be a function from \mathbb{R}^n to \mathbb{R}^m . Suppose that on some open set U including \mathbf{x}_0 , all of the partial derivatives of each component f_i of \mathbf{f} exist, and are continuous at \mathbf{x}_0 . Then \mathbf{f} is differentiable at \mathbf{x}_0 .*

Proof. Divide and conquer:
$$\frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - A(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} \leq \sum_{i=1}^m \frac{|f_i(\mathbf{x}) - f_i(\mathbf{x}_0) - \nabla f_i(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|}.$$

Therefore, it suffices to show that for each i ,

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|f_i(\mathbf{x}) - f_i(\mathbf{x}_0) - \nabla f_i(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0$$

under the assumption that the partial derivatives of each f_i are continuous. But this has already been done in Theorem 42. \square

Example 76 (Differentiability of rational functions). *Let $p(\mathbf{x})$ and $q(\mathbf{x})$ be polynomials in the n variables x_1, \dots, x_n . Let U be an open set in \mathbb{R}^n such that $q(\mathbf{x}) \neq 0$ for any $\mathbf{x} \in U$. Then the rational function*

$$f(\mathbf{x}) := \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

is well-defined on U .

By the quotient rule of single variable calculus, for each $j = 1, \dots, n$,

$$\frac{\partial}{\partial x_j} f(\mathbf{x}) = \left(\frac{\partial}{\partial x_j} q(\mathbf{x}) \right)^{-2} \left[\left(\frac{\partial}{\partial x_j} p(\mathbf{x}) \right) q(\mathbf{x}) - \left(\frac{\partial}{\partial x_j} q(\mathbf{x}) \right) p(\mathbf{x}) \right].$$

Since $\frac{\partial}{\partial x_j} p(\mathbf{x})$ and $\frac{\partial}{\partial x_j} q(\mathbf{x})$ are both polynomials in the n variables x_1, \dots, x_n , so is $\frac{\partial}{\partial x_j} f(\mathbf{x})$, and the denominator is not zero for any $\mathbf{x} \in U$. Therefore, each of the partial derivatives of f is continuous in U , and hence f is differentiable at each point \mathbf{x} in U .

As a special case, $f(x, y) = x^2 + y^2$ is differentiable at each (x_0, y_0) in \mathbb{R}^2 .

4.3.2 The general chain rule

Let \mathbf{f} be a function from \mathbb{R}^n to \mathbb{R}^m that is differentiable at $\mathbf{x}_0 \in \mathbb{R}^n$. Let \mathbf{g} be a differentiable function from \mathbb{R}^m to \mathbb{R}^ℓ that is differentiable at $\mathbf{f}(\mathbf{x}_0)$. Then for \mathbf{x} near \mathbf{x}_0 ,

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}_0) + [\mathbf{D}\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0),$$

and for \mathbf{y} near $\mathbf{f}(\mathbf{x}_0)$,

$$\mathbf{g}(\mathbf{y}) \approx \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) + [\mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{x}_0))](\mathbf{y} - \mathbf{f}(\mathbf{x}_0)).$$

If we take $\mathbf{y} := \mathbf{f}(\mathbf{x})$, then for \mathbf{x} sufficiently close to \mathbf{x}_0 , \mathbf{y} will be close to $\mathbf{f}(\mathbf{x}_0)$, and so we will have

$$\begin{aligned} \mathbf{g}(\mathbf{f}(\mathbf{x})) &\approx \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) + [\mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{x}_0))](\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)) \\ &\approx \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) + [\mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{x}_0))][\mathbf{D}\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0). \end{aligned}$$

Thus, the composite function $\mathbf{g} \circ \mathbf{f}$ will have the linear approximation

$$\mathbf{g}(\mathbf{f}(\mathbf{x})) \approx \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) + [\mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{x}_0))][\mathbf{D}\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)$$

at \mathbf{x}_0 , which suggests that $\mathbf{g} \circ \mathbf{f}$ is differentiable, and its derivative is the matrix product $[\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))][\mathbf{Df}(\mathbf{x}_0)]$. Remembering that matrix multiplication represents composition of linear functions, this formula would be very natural. In fact, it is valid without any extra hypotheses.

Theorem 64. *Let \mathbf{f} be a function from \mathbb{R}^n to \mathbb{R}^m that is differentiable at $\mathbf{x}_0 \in \mathbb{R}^n$. Let \mathbf{g} be a differentiable function from \mathbb{R}^m to \mathbb{R}^ℓ that is differentiable at $\mathbf{f}(\mathbf{x}_0)$. Then the composite function $\mathbf{g} \circ \mathbf{f}$ is differentiable at \mathbf{x}_0 and*

$$[\mathbf{Dg} \circ \mathbf{f}(\mathbf{x}_0)] = [\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))][\mathbf{Df}(\mathbf{x}_0)] ,$$

where the right hand side is the product of the Jacobian matrices of \mathbf{f} and \mathbf{g} at the indicated points.

Proof. Define $\mathbf{y} := \mathbf{f}(\mathbf{x})$ and $\mathbf{y}_0 := \mathbf{f}(\mathbf{x}_0)$ and then define

$$\mathbf{w} := \mathbf{f}(\mathbf{x}) - (\mathbf{f}(\mathbf{x}_0) - [\mathbf{Df}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)) \quad \text{and} \quad \mathbf{z} := \mathbf{g}(\mathbf{y}) - (\mathbf{g}(\mathbf{y}_0) + [\mathbf{Dg}(\mathbf{y}_0)](\mathbf{y} - \mathbf{y}_0)) .$$

With the definitions,

$$\begin{aligned} \mathbf{g}(\mathbf{f}(\mathbf{x})) - \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) &= \mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{y}_0) \\ &= [\mathbf{Dg}(\mathbf{y}_0)](\mathbf{y} - \mathbf{y}_0) + \mathbf{z} \\ &= [\mathbf{Dg}(\mathbf{y}_0)](\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)) + \mathbf{z} \\ &= [\mathbf{Dg}(\mathbf{y}_0)][[\mathbf{Df}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) + \mathbf{w}] + \mathbf{z} \\ &= [\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))][\mathbf{Df}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) + [\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w} + \mathbf{z} . \end{aligned}$$

Thus by the triangle inequality, and then Theorem 60,

$$\begin{aligned} \frac{\|\mathbf{g}(\mathbf{f}(\mathbf{x})) - \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) - [\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))][\mathbf{Df}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} &\leq \frac{\|[\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}\|}{\|\mathbf{x} - \mathbf{x}_0\|} + \frac{\|\mathbf{z}\|}{\|\mathbf{x} - \mathbf{x}_0\|} \\ &\leq \|[\mathbf{Dg}(\mathbf{f}(\mathbf{x}_0))]\|_{\mathbb{F}} \frac{\|\mathbf{w}\|}{\|\mathbf{x} - \mathbf{x}_0\|} + \frac{\|\mathbf{z}\|}{\|\mathbf{x} - \mathbf{x}_0\|} . \end{aligned}$$

Thus, it suffices to prove that $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{\|\mathbf{w}\|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0$ and $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{\|\mathbf{z}\|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0$. The first of these equations is satisfied since \mathbf{w} is differentiable. Next, recalling the definitions of \mathbf{y} and \mathbf{z} ,

$$\frac{\|\mathbf{z}\|}{\|\mathbf{x} - \mathbf{x}_0\|} = \frac{\|\mathbf{g}(\mathbf{y}) - (\mathbf{g}(\mathbf{y}_0) + [\mathbf{Dg}(\mathbf{y}_0)](\mathbf{y} - \mathbf{y}_0))\|}{\|\mathbf{y} - \mathbf{y}_0\|} \frac{\|\mathbf{y} - \mathbf{y}_0\|}{\|\mathbf{x} - \mathbf{x}_0\|}$$

We now claim that as $\mathbf{x} \rightarrow \mathbf{x}_0$, not only do we have that $\mathbf{y} \rightarrow \mathbf{y}_0$, but also that the ratio $\frac{\|\mathbf{y} - \mathbf{y}_0\|}{\|\mathbf{x} - \mathbf{x}_0\|}$ is bounded by some constant M in some open set about \mathbf{x}_0 . Suppose this is true. Then, by the differentiability of \mathbf{g} ,

$$0 \leq \lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{\|\mathbf{z}\|}{\|\mathbf{x} - \mathbf{x}_0\|} \leq M \lim_{\mathbf{y} \rightarrow \mathbf{y}_0} \frac{\|\mathbf{g}(\mathbf{y}) - (\mathbf{g}(\mathbf{y}_0) + [\mathbf{Dg}(\mathbf{y}_0)](\mathbf{y} - \mathbf{y}_0))\|}{\|\mathbf{y} - \mathbf{y}_0\|} = 0 ,$$

which is what we need.

To conclude, use the definitions of \mathbf{y} and \mathbf{y}_0 , add and subtract $\mathbf{Df}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$, and use the triangle inequality and then Theorem 59 to obtain

$$\begin{aligned} \frac{\|\mathbf{y} - \mathbf{y}_0\|}{\|\mathbf{x} - \mathbf{x}_0\|} &= \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} \leq \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - \mathbf{Df}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} + \frac{\|\mathbf{Df}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} \\ &\leq \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - \mathbf{Df}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} + \|\mathbf{Df}(\mathbf{x}_0)\|_{\mathbb{F}} , \end{aligned}$$

Again by differentiability of \mathbf{f} at \mathbf{x}_0 , the first term on the right goes to 0 as $\mathbf{x} \rightarrow \mathbf{x}_0$, and hence taking $M := 1 + \|D\mathbf{f}(\mathbf{x}_0)\|_F$, there is an $r > 0$ so that $\|\mathbf{y} - \mathbf{y}_0\| \leq M\|\mathbf{x} - \mathbf{x}_0\|$ whenever $\|\mathbf{x} - \mathbf{x}_0\| < r$. \square

4.4 Newton's Method

4.4.1 Linear approximation and Newton's iterative scheme

Let \mathbf{f} be a differentiable function from \mathbb{R}^n to \mathbb{R}^n . Suppose we want to solve the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. For example, if we are given a differentiable function f from \mathbb{R}^n to \mathbb{R} , and we want to find its critical points, than we must solve the equation $\nabla f(\mathbf{x}) = \mathbf{0}$. Defining $\mathbf{f}(\mathbf{x}) := \nabla f(\mathbf{x})$, we have arrived at the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$.

The equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ arises in many other ways as well. Let \mathbf{g} and \mathbf{h} be functions from \mathbb{R}^n to \mathbb{R}^n , and suppose that we want to solve $\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$. Defining \mathbf{f} by $\mathbf{f}(\mathbf{x}) := \mathbf{g}(\mathbf{x}) - \mathbf{h}(\mathbf{x})$, we see that

$$\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x}) \iff \mathbf{f}(\mathbf{x}) = \mathbf{0}.$$

This example is worth bearing in mind: *Many equations can be written in the form $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ by the simple device of “moving everything to the left of the equal sign”.*

Newton's method for solving $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ is to successively improve an approximate solution using the linear approximation to \mathbf{f} that is provided by differentiation. The multivariable version works in exactly the same way as the single variable version.

Recall that the *linear approximation* to \mathbf{f} at \mathbf{x}_0 is given by

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0). \quad (4.63)$$

Now replace \mathbf{f} in the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ by its linear approximation to obtain an equation that approximates (4.63):

$$\mathbf{f}(\mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) = \mathbf{0}, \quad (4.64)$$

which is the same as $[D\mathbf{f}(\mathbf{x}_0)]\mathbf{x} = [D\mathbf{f}(\mathbf{x}_0)]\mathbf{x}_0 - \mathbf{f}(\mathbf{x}_0)$. This is nothing other than the matrix equation

$$A\mathbf{x} = \mathbf{b} \quad \text{where} \quad A := [D\mathbf{f}(\mathbf{x}_0)] \quad \text{and} \quad \mathbf{b} := [D\mathbf{f}(\mathbf{x}_0)]\mathbf{x}_0 - \mathbf{f}(\mathbf{x}_0).$$

As long as the matrix $A := [D\mathbf{f}(\mathbf{x}_0)]$ is invertible, there is a unique solution which is $\mathbf{x} = A^{-1}\mathbf{b}$, or, more explicitly,

$$\mathbf{x} = \mathbf{x}_0 - [D\mathbf{f}(\mathbf{x}_0)]^{-1}\mathbf{f}(\mathbf{x}_0).$$

Now, in so far as \mathbf{x}_0 is an approximate solution to $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, so that we may expect there to be an exact solution nearby, and in so far as (4.63) is a good approximation, we can expect the solution of (4.64) to be a more accurate approximate solution. That is, defining

$$\mathbf{x}_1 := \mathbf{x}_0 - [D\mathbf{f}(\mathbf{x}_0)]^{-1}\mathbf{f}(\mathbf{x}_0),$$

we can hope that \mathbf{x}_1 is a better approximate solution than \mathbf{x}_0 .

Now simply iterate this “improvement” procedure: Given the starting point \mathbf{x}_0 , we define the infinite sequence $\{\mathbf{x}_n\}$ by

$$\mathbf{x}_{n+1} := \mathbf{x}_n - [D\mathbf{f}(\mathbf{x}_n)]^{-1}\mathbf{f}(\mathbf{x}_n). \quad (4.65)$$

Of course, the construction of the sequence is only meaningful if $[D\mathbf{f}(\mathbf{x}_n)]$ is invertible for each n .

Newton's method is a “successive approximations method”. It takes a starting guess for the solution \mathbf{x}_0 , and iteratively improves the guess. The iteration scheme produces an *infinite sequence* of approximate solutions $\{\mathbf{x}_n\}$. Under favorable circumstances, this sequence will converge *very rapidly* toward an exact solution. In fact, the number of correct digits in each entry of \mathbf{x}_n will more or less double at each step, once you get reasonably close. If you have one digit right at the outset, you may expect about a million correct digits after 20 iterations – more than you are ever likely to want to keep!

To explain the use of Newton's method, we have to cover three points:

- (i) How one picks the starting guess \mathbf{x}_0 .
- (ii) How the iterative loop runs; i.e., the rule for determining \mathbf{x}_{n+1} given \mathbf{x}_n , which we have already explained.
- (iii) How to break out of the iterative loop – we need a “stopping rule” which assures us that there is an actual solution within some prescribed distance ϵ of \mathbf{x}_n , and hence we may stop iterating at the n th step.

We begin by explaining (ii), the nature of the loop. Once we are familiar with the loop, we can better understand what we have to do to start it and stop it. Let us run through an example.

Consider the following system of non linear equations:

$$\begin{aligned} x^2 + 2yx &= 4 \\ xy &= 1 . \end{aligned} \tag{4.66}$$

As noted above, any system of two equations in two variables can be written in the form

$$\begin{aligned} f(x, y) &= 0 \\ g(x, y) &= 0 . \end{aligned}$$

In this case we define $f(x, y) = x^2 + 2xy - 4$ and $g(x, y) = xy - 1$.

Next, introducing $\mathbf{f}(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x}))$, we can write (4.66) as a single vector equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. In the case of (4.66), we have

$$\mathbf{f}(x, y) = (x^2 + 2xy - 4, xy - 1) . \tag{4.67}$$

In this example, we can solve $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ by algebra alone: Using the second equation in (4.66) to eliminate y , the first equation becomes $x^2 = 2$. Hence $x = \pm\sqrt{2}$. The second equation says that $y = 1/x$ and so we have two solutions $(\sqrt{2}, 1/\sqrt{2})$ and $(-\sqrt{2}, -1/\sqrt{2})$. In other examples, it may be quite hard to eliminate either variable, and algebra alone cannot deliver solutions. Now let's see how Newton's Method works with this example:

Example 77 (Using Newton's iteration). *Consider the system of equations $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ where \mathbf{f} is given by (4.67). We will choose a starting point so that at least one of the equations in the system is satisfied, and the other is not too far off. This seems reasonable enough. Notice that with $x = y = 1$, $xy - 1 = 0$,*

while $x^2 - 2xy - 4 = -1$. With $\mathbf{x}_0 = (1, 1)$ we have $\mathbf{f}(\mathbf{x}_0) = (-1, 0)$. Computing the derivative of \mathbf{f} , we find $[\mathbf{D}\mathbf{f}(\mathbf{x})] = \begin{bmatrix} 2x + 2y & 2x \\ y & x \end{bmatrix}$ and hence $[\mathbf{D}\mathbf{f}(\mathbf{x}_0)] = \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}$. Then $\mathbf{x}_1 := \mathbf{x}_0 - [\mathbf{D}\mathbf{f}(\mathbf{x}_0)]^{-1} \mathbf{f}(\mathbf{x}_0)$ is

$$\mathbf{x}_1 = (1, 1) - \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}^{-1} (-1, 0).$$

Since $\begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -2 \\ -1 & 4 \end{bmatrix}$, we find $\mathbf{x}_1 = (3/2, 1/2)$. Notice that \mathbf{x}_1 is indeed considerably closer to the exact solution $(\sqrt{2}, 1/\sqrt{2})$ than \mathbf{x}_0 . Moreover, $\mathbf{f}(\mathbf{x}_1) = -\frac{1}{4}(1, 1)$. This is a better approximate solution; it is much closer to the actual solution. If you now iterate this further, you will find a sequence of approximate solutions converging to the exact solution $(\sqrt{2}, 1/\sqrt{2})$. You should compute \mathbf{x}_2 and \mathbf{x}_3 and observe the speed of convergence.

4.4.2 The Geometry of Newton's Method

To understand when and how Newton's method works, it is useful to relate the affine approximation of \mathbf{f} to the tangent plane approximation of each of the entry functions f_j in $\mathbf{f} = (f_1, \dots, f_n)$.

To be able to draw graphs, and to keep the discussion as geometrically clear as possible, let us take $n = 2$, and write $\mathbf{f} = (f, g)$. Then $\mathbf{f} = \mathbf{0}$ is equivalent to the system of equations

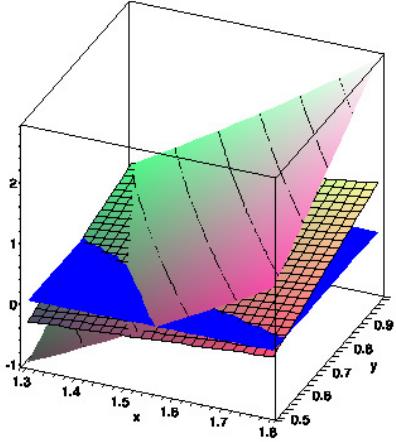
$$\begin{aligned} f(x, y) &= 0 \\ g(x, y) &= 0. \end{aligned} \tag{4.68}$$

Replace this by the equivalent system

$$\begin{aligned} z &= f(x, y) \\ z &= g(x, y) \\ z &= 0. \end{aligned} \tag{4.69}$$

From an algebraic standpoint, we have taken a step backwards – we have gone from two equations in two variables to three equations in three variables. However, (4.69) has an interesting geometric meaning: The graph of $z = f(x, y)$ is a surface in \mathbb{R}^3 , as is the graph of $z = g(x, y)$. The graph of $z = 0$ is just the x, y plane – a third surface. Hence the solution set of (4.69) is given by the intersection of 3 surfaces.

For example, in the plot below you see the three surfaces in (4.69) when $f(x, y) = x^2 + 2xy - 4$ and $g(x, y) = xy - 1$, as in Example 77. Here, we have plotted $1.3 \leq x \leq 1.8$ and $0.5 \leq y \leq 1$, which includes one exact solution of the system (4.68) in this case. The plane $z = 0$ is the surface in solid color, $z = f(x, y)$ shows the contour lines, and $z = g(x, y)$ is the surface showing a grid. You see where all three surfaces intersect, and that is the where the solution lies.



You also see in this graph that the tangent plane approximation is pretty good in this region, so replacing the surfaces by their tangent planes will not wreak havoc on the graph. So here is what we do: Take any point \$(x_0, y_0)\$ so that the three surface intersect *near* \$(x_0, y_0, 0)\$. Then replace the surfaces \$z = f(x, y)\$ and \$z = g(x, y)\$ by their tangent planes at \$(x_0, y_0)\$, and compute the intersection of the tangent planes with the plane \$z = 0\$. This is a simple linear algebra problem that we know how to solve. Replacing \$z = f(x, y)\$ and \$z = g(x, y)\$ by the equations of their tangent planes at \$(x_0, y_0)\$ amounts to the replacement

$$z = f(x, y) \quad \rightarrow \quad z = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

and

$$z = g(x, y) \quad \rightarrow \quad z = g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

where \$\mathbf{x}_0 = (x_0, y_0)\$. This transforms (4.69) into

$$\begin{aligned} z &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) \\ z &= g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) \\ z &= 0. \end{aligned} \tag{4.70}$$

Now we can eliminate \$z\$, and pass to the simplified system

$$\begin{aligned} f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) &= 0 \\ g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) &= 0. \end{aligned} \tag{4.71}$$

Since \$[\mathbf{D}f(\mathbf{x}_0)] = \begin{bmatrix} \nabla f(\mathbf{x}_0) \\ \nabla g(\mathbf{x}_0) \end{bmatrix}\$, this is equivalent to (4.64) by the rules for matrix multiplication.

We see from this analysis that how close we come to an exact solution in one step of Newton's method depends on how good the tangent plane approximation is at the current approximate solution. To really understand how well Newton's method works, we need to understand something about how the surfaces given by \$z = f(x, y)\$ and \$z = g(x, y)\$ "curve away" from their tangent planes at \$\mathbf{x}_0\$. That is, we will need to know something about the curvature of surfaces. As you may guess from our investigation of curvature of curves \$\mathbf{x}(t)\$, this will involve second derivatives of functions from \$\mathbb{R}^n\$ to \$\mathbb{R}\$. We take up the topic of higher derivatives of functions from \$\mathbb{R}^n\$ to \$\mathbb{R}\$ in the next chapter. However, there is much that can be accomplished considering only first derivatives.

4.4.3 Starting and stopping the iteration

It is not always a simple matter to determine how many solutions $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ may have, or to find approximate locations of these solutions. Ideally, before starting the iterative procedure, one would like to know how many solutions $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ has, at least in some region of interest, and then one needs to find a good starting point close to each of these. Let us begin with a simple example where this is readily accomplished.

Example 78 (Graphical location of starting points). *Consider the system*

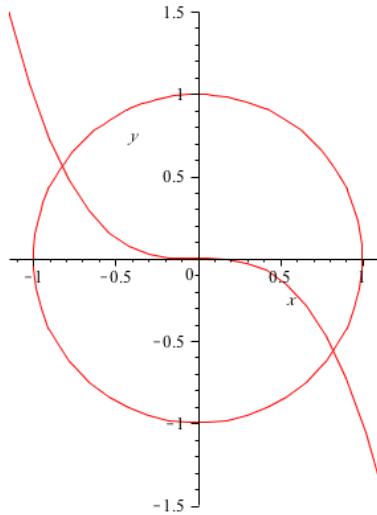
$$(\mathbf{f}(x, y), g(x, y)) = \mathbf{0}$$

where $f(x, y) = x^4 + xy$, and $g(x, y) = 1 - y^2 - x^2$.

The equation $g(x, y) = 0$ is equivalent to the equation $x^2 + y^2 = 1$, i.e., the equation of the unit circle. Therefore, all solutions of $g(x, y) = 0$ lie on the unit circle, and hence all solutions of $(\mathbf{f}(x, y), g(x, y)) = \mathbf{0}$ lie on the unit circle.

Now, what about the equation $f(x, y) = 0$? Can we find a similar explicit description of its solution set? Yes, since $f(x, y) = x(x^3 + y)$, $f(x, y) = 0$ if and only if $x = 0$, which is the equation of the y -axis, or if $y = -x^3$, which is the equation of an easily plotted cubic curve.

Hence the solutions of our system of equations are exactly the intersection of the unit circle with the y -axis, and the intersection of the unit circle with the cubic curve $y = -x^3$. Here is a plot showing the unit circle, the y -axis and the cubic curve $y = -x^3$.



As you see, the system has four solutions. Two of them are readily given in closed form: The y -axis intersects the unit circle at the points $(0, 1)$ and $(0, -1)$.

It is actually possible to find an exact formula for the other two solutions by algebraic means. Substituting $y = -x^3$ into $x^2 + y^2 = 1$, we obtain $x^2 + x^6 = 1$. Introducing the variable $t = x^2$, this becomes the cubic equation $t + t^3 = 1$. There is an algebraic formula, known as Cardano's formula for the roots of any cubic.

In many applications, what one would want anyhow would be some such decimal approximation, and not the exact solution obtained by using Cardano's formula (which is considerably more compli-

cated than the quadratic formula). Newton's method provides a more direct route. From the plot you can see that

$$(-0.80, 0.55) \text{ and } (0.80, -0.55)$$

are decent approximate solutions. Using either one as the starting point for the Newton iteration, one readily determines 10 (or more) accurate digits. Moreover, since the second solution is exactly minus the first, one only need to run the iteration for one of these solutions.

If you have more than two variables, plots become harder to use. In such a case, it is often easier to eliminate from consideration a set of points that cannot possibly contain a solutions, and then to choose starting points from among the points that are left over. The following lemma tells us *where not to look for solutions*.

The basic idea is very simple. Suppose at some point \mathbf{x}_0 , the i th component of \mathbf{f} is not zero; i.e.,

$$f_i(\mathbf{x}_0) \neq 0.$$

Suppose for the sake of argument that $f_i(\mathbf{x}_0) > 0$, so that your “altitude” is too high at \mathbf{x}_0 . To get to a solution, you have to get downhill by a height $h = f_i(\mathbf{x}_0)$. But if the slope is bounded in magnitude by $C > 0$, you cannot lose altitude h without traveling a horizontal distance of at least h/C . Since any solution \mathbf{x} of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ must satisfy $f_i(\mathbf{x}) = 0$, there can be no solution within a radius h/C of \mathbf{x}_0 . Here is the precise version:

Lemma 14 (Elimination lemma). *Let B be a bounded closed subset of \mathbb{R}^n . Suppose also that B is convex; i.e., B contains every point on the line segment connecting any two points in B . Let \mathbf{f} be a function from \mathbb{R}^n to \mathbb{R}^n that is continuously differentiable on some open set containing B , and suppose that for some constant C , $\|\mathbf{D}\mathbf{f}(\mathbf{x}_0)\| \leq C$ for all $\mathbf{x} \in B$. Then*

$$\|\mathbf{x} - \mathbf{x}_0\| < \frac{\|\mathbf{f}(\mathbf{x}_0)\|}{C} \Rightarrow \|\mathbf{f}(\mathbf{x})\| > 0.$$

Proof. By the Fundamental Theorem of Calculus and the Chain Rule,

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) = \int_0^1 [\mathbf{D}\mathbf{f}(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0))](\mathbf{x} - \mathbf{x}_0) dt.$$

By the triangle inequality for integrals and Theorem 59, and our hypothesis in $\|\mathbf{D}\mathbf{f}\|$,

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| \leq \left(\int_0^1 \|\mathbf{D}\mathbf{f}(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0))\| dt \right) \|\mathbf{x} - \mathbf{x}_0\| \leq C \|\mathbf{x} - \mathbf{x}_0\|.$$

Then by the triangle inequality,

$$\|\mathbf{f}(\mathbf{x}_0)\| = \|\mathbf{f}(\mathbf{x}_0) - \mathbf{f}(\mathbf{x}) + \mathbf{f}(\mathbf{x})\| \leq C \|\mathbf{x} - \mathbf{x}_0\| + \|\mathbf{f}(\mathbf{x})\|.$$

Hence $\|\mathbf{f}(\mathbf{x})\| \geq \|\mathbf{f}(\mathbf{x}_0)\| - C \|\mathbf{x} - \mathbf{x}_0\|$. □

One way to use this is to compute an upper bound C , the maximum of $\|\mathbf{D}\mathbf{f}(\mathbf{x})\|$ for $\mathbf{x} \in B$, and then to cover B with a square grid of boxes of side length ϵ . Let \mathbf{x}_0 be the center of one of the boxes. For \mathbf{x} in the box, $\|\mathbf{x} - \mathbf{x}_0\| \leq \sqrt{n}\epsilon/2$. Hence if $\|\mathbf{f}(\mathbf{x}_0)\| > C\sqrt{n}\epsilon/2$, there cannot be any \mathbf{x}

with $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ in this box. So to find the solutions, discard all such boxes, and run Newton's method from the center of each box that is not eliminated. note that the smaller you make ϵ , the easier it is to eliminate individual boxes, however the small ϵ is, the more boxes there will be to check.

Now, suppose you run Newton's method from the center of a box, and find a solution in the box. Is it the *only* solution in this box? The answer to that has to do with curvature, and we return to the question later.

Finally there is the matter of stopping Newton's method. Again, we will return to this later, but you will observe the following in the exercises: Once you get one or two digits after the decimal point correct, the number of correct digits approximately doubles with each iteration. Since $2^{20} > 10^6$, it follows that you will have about 10^6 correct digits in 20 iterations provided you picked a starting point, perhaps using one of the methods described above, that has one or two digits to the right of the decimal point correct.

4.5 Exercises

4.1 Let $\mathbf{v}_1 = (1, 1)$ and $\mathbf{v}_2 = (0, 1)$. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be differentiable, and suppose that for some $\mathbf{x}_0 \in \mathbb{R}^2$, $\mathbf{v}_1 \cdot \nabla f(\mathbf{x}_0) = 2$ and $\mathbf{v}_2 \cdot \nabla f(\mathbf{x}_0) = -2$. For $\mathbf{v} = (2, -3)$, compute $\mathbf{v} \cdot \nabla f(\mathbf{x}_0)$.

4.2 Let $\mathbf{v}_1 = (1, 1)$ and $\mathbf{v}_2 = (3, 1)$. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be differentiable, and suppose that for some $\mathbf{x}_0 \in \mathbb{R}^2$, $\mathbf{v}_1 \cdot \nabla f(\mathbf{x}_0) = -2$ and $\mathbf{v}_2 \cdot \nabla f(\mathbf{x}_0) = 3$.

For $\mathbf{v} = (1, -1)$, compute $\mathbf{v} \cdot \nabla f(\mathbf{x}_0)$.

4.3 Let $\mathbf{v}_1 = (1, 1, 1)$, $\mathbf{v}_2 = (0, 1, 1)$ and $\mathbf{v}_3 = (0, 0, 1)$. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be differentiable, and suppose that for some $\mathbf{x}_0 \in \mathbb{R}^3$,

$$\mathbf{v}_1 \cdot \nabla f(\mathbf{x}_0) = 5 \quad \mathbf{v}_2 \cdot \nabla f(\mathbf{x}_0) = 3 \quad \text{and} \quad \mathbf{v}_3 \cdot \nabla f(\mathbf{x}_0) = 2 .$$

For $\mathbf{v} = (1, 2, 3)$, compute $\mathbf{v} \cdot \nabla f(\mathbf{x}_0)$.

4.4 Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^2y + yx - xy^2$. Let $\mathbf{x}(t)$ b given by $\mathbf{x}(t) = (t, t^2)$ Compute $\frac{d}{dt} f(\mathbf{x}(t)) \Big|_{t=1}$.

4.5 Let $f(x, y) = \frac{xy}{(1 + x^2 + y^2)^2}$.

(a) Find all of the critical points of f , and find the value of f at each of the critical points.

(b) Does f have a maximum value? Explain why or why not. If it does, find all points at which the value of f is maximal; i.e, find all maximizers.

(c) Does f have a minimum value? Explain why or why not. If it does, find all points at which the value of f is minimal; i.e, find all minimizers.

4.6 Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^2y + yx - xy^2$.

(a) Compute the gradient of f , and find all critical points of f .

(b) Find the equation of the tangent plane to the graph f at the point $(1, 1)$.

4.7 Let $f(x, y) = 3xy - x^3 - y^3$. Find all points (x, y) at which the tangent plane to the graph of f is orthogonal to the line parameterized by $t(3, 3, 1)$.

4.8 Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^3 + y^3 + 3xy$.

(a) Compute the gradient of f , and find all points (x, y) at which the tangent plane to the graph of f is horizontal.

(b) Find the equation of the tangent plane to the graph of f at the point $(1, 2)$.

(c) If you were standing at the point $(1, 1)$ and wanted to climb uphill as directly as possible, in which compass direction would you head? For purposes of this question, take the direction \mathbf{e}_1 to be East, and the direction \mathbf{e}_2 to be North.

4.9 Let $\mathbf{v}_1 = (1, 1, 1)$, $\mathbf{v}_2 = (0, 1, 1)$ and $\mathbf{v}_3 = (0, 0, 1)$. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be differentiable, and suppose that for some $\mathbf{x}_0 \in \mathbb{R}^3$,

$$\mathbf{v}_1 \cdot \nabla f(\mathbf{x}_0) = 5 \quad \mathbf{v}_2 \cdot \nabla f(\mathbf{x}_0) = 3 \quad \text{and} \quad \mathbf{v}_3 \cdot \nabla f(\mathbf{x}_0) = 2 .$$

Find a right-handed orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of \mathbb{R}^3 such that \mathbf{u}_1 is parallel to $\nabla f(\mathbf{x}_0)$.

4.10 Let $A := \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}$.

(a) Compute $\det(A)$ and the matrix inverse of A .

(b) Find all solutions of the equation $A\mathbf{x} = (1, 2)$.

4.11 Let $A := \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$.

(a) Compute $\det(A)$ and the matrix inverse of A .

(b) Find all solutions of the equation $A\mathbf{x} = (3, 1)$.

4.12 Let $A := \begin{bmatrix} 2 & 0 & 3 \\ -1 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}$.

(a) Compute $\det(A)$ and the matrix inverse of A .

(b) Find all solutions of the equation $A\mathbf{x} = (1, 2, 3)$.

4.13 Let $A := \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$.

(a) Compute $\det(A)$ and the matrix inverse of A .

(b) Find all solutions of the equation $A\mathbf{x} = (4, 0, 8)$.

4.14 Let $f(x, y) = x^2y + (x - 1)(y - 1)^2$.

(a) Find the equations for the tangent planes to the graph of $z = f(x, y)$ at $\mathbf{x}_1 = (2, -1)$ and at $\mathbf{x}_2 = (-1, 1)$.

(b) Parameterize the line that is the intersection of the two planes found in (a).

(c) Let $\mathbf{x}_0 = (-3, 0, 5)$. Compute the distance from \mathbf{x}_0 to the line found in (b), and from \mathbf{x}_0 to the second tangent plane found in (a). The distance to the plane should be smaller than the distance to the line. Explain why.

4.15: Let $f(x, y)$ and $g(x, y)$ be given by

$$f(x, y) = x^2y - 4y + x^3 \quad \text{and} \quad g(x, y) = 4x^2 + 4y^2 - 1 .$$

Define the function \mathbf{f} from \mathbb{R}^2 to \mathbb{R}^2 by

$$\mathbf{f}(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x})) .$$

(a) Let $\mathbf{x}_0 = (-1/2, 0)$. Compute $\mathbf{f}(\mathbf{x}_0)$ and the Jacobian of \mathbf{f} at \mathbf{x}_0 ; i.e., $[D\mathbf{f}(\mathbf{x}_0)]$.

(b) Use \mathbf{x}_0 as a starting point for Newton's method for solving $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, and find \mathbf{x}_1 , the next step.

(c) Evaluate $\mathbf{f}(\mathbf{x}_1)$. Comment on your result - how many solutions are there, and how close is \mathbf{x}_1 to one of them?

4.16 Let $\mathbf{f}(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x}))$ where $f(x, y) = x^3 + xy$, and $g(x, y) = 1 - 4y^2 - x^2$. Let $\mathbf{x}_0 = \mathbf{e}_1$.

(a) Compute $[D\mathbf{f}(\mathbf{x})]$ and $[D\mathbf{f}(\mathbf{x}_0)]$.

(b) Use \mathbf{x}_0 as a starting point for Newton's method, and compute the next approximate solution \mathbf{x}_1 .

(c) Evaluate $\mathbf{f}(\mathbf{x}_1)$, and compare this with $\mathbf{f}(\mathbf{x}_0)$.

4.17 Let $\mathbf{f}(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x}))$ where $f(x, y) = \sqrt{x} + \sqrt{y} - 3$, and $g(x, y) = x^2 + y^2 - 18$. Compute $\mathbf{f}(\mathbf{x}_0)$ for $\mathbf{x}_0 = (3, 3)$. Does this look like a reasonable starting point? Compute $[D\mathbf{f}(\mathbf{x}_0)]$. What happens if you try to use \mathbf{x}_0 as your starting point for Newton's method?

4.18: Let $f(x, y)$ and $g(x, y)$ be given by

$$f(x, y) = (x + 1)^2 + (y + 1)^2 - 4 \quad \text{and} \quad g(x, y) = 4(x - 1)^2 + (y - 1)^2 - 5 .$$

(a) The equation $f(x, y) = 0$ describes a circle. The equation $g(x, y) = 0$ describes an ellipse. Sketch a plot of the circle and the ellipse, and determine how many solutions there are of the system

$$\mathbf{f}(x, y) := (f(x, y), g(x, y)) = \mathbf{0} .$$

Note: It is possible to exactly solve this system by algebraic means, but that is not what is being asked for here.

(b) Your sketch should show one solution of the system in the lower right hand quadrant not too far from $(1, -1)$. Use $\mathbf{x}_0 = (1, -1)$ as a starting point, and apply one step of Newton's method to find \mathbf{x}_1 , a better approximate solution, and compute $f(\mathbf{x}_1)$ and $g(\mathbf{x}_1)$.

4.19 Let $\mathbf{f}(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x}))$ where $f(x, y) = \sin(xy) - x$, and $g(x, y) = x^2y - 1$. Let $\mathbf{x}_0 = (1, 1)$.

(a) Compute $[D\mathbf{f}(\mathbf{x})]$ and $[D\mathbf{f}(\mathbf{x}_0)]$.

(b) Use \mathbf{x}_0 as a starting point for Newton's method, and compute the next approximate solution \mathbf{x}_1 .

(c) Evaluate $\mathbf{f}(\mathbf{x}_1)$, and compare this with $\mathbf{f}(\mathbf{x}_0)$.

(d) How many solutions of this system are there in the region $-2 \leq x \leq 2$ and $0 \leq y \leq 10$? Compute each of them to 10 decimal places of accuracy – using a computer, of course.

4.20 Let A be a 3×3 matrix whose three rows are \mathbf{r}_1 , \mathbf{r}_2 and \mathbf{r}_3 , and whose three columns are \mathbf{v}_1 , \mathbf{v}_2 and \mathbf{v}_3 , so that

$$A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix}.$$

Show that

$$\det(A) = \mathbf{r}_1 \cdot \mathbf{r}_2 \times \mathbf{r}_3.$$

That is, show that the triple product of the rows of a 3×3 matrix is always equal to the triple product of its columns, which, by definition, is the determinant.

4.21 Show that for all \mathbf{a} , \mathbf{b} and \mathbf{c} in \mathbb{R}^3 ,

$$(\mathbf{b} \times \mathbf{c}) \cdot [(\mathbf{c} \times \mathbf{a}) \times (\mathbf{a} \times \mathbf{b})] = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|^2.$$

4.22 Let $\mathbf{f}(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x}))$ where $f(x, y) = xy - x^3 - 1/4$, and $g(x, y) = 1 - 4y^2 - x^2$.

(a) How many solutions to the system $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ are there? Draw a plot showing their approximate location.

(b) In the previous part, you should have found that there is one solution not too far from

$$\mathbf{x}_0 = (-1, 1/2).$$

Compute $[D\mathbf{f}(\mathbf{x})]$, and then use \mathbf{x}_0 as a starting point for Newton's method, and compute the next approximate solution \mathbf{x}_1 .

(c) Evaluate $\mathbf{f}(\mathbf{x}_1)$, and compare this with $\mathbf{f}(\mathbf{x}_0)$.

Chapter 5

THE IMPLICIT FUNCTION THEOREM AND ITS CONSEQUENCES

5.1 Horizontal slices and contour curves

In previous chapters, we have considered *vertical* slices of the graph of $z = f(x, y)$. Considering *horizontal* slices provides a new perspective.

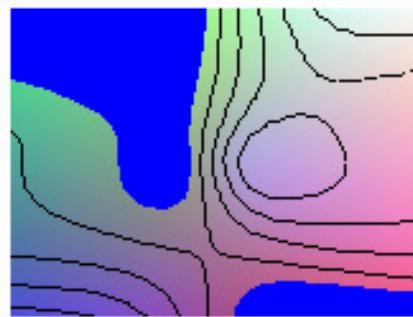
You are probably familiar with topographic maps that show curves along which the altitude is constant. Let horizontal coordinates x and y be given – for example, latitude and longitude in some patch of the Earth’s surface which is small enough that they can be treated as Cartesian coordinates in a plane. Let $f(x, y)$ denote the altitude at the point with latitude x and longitude y . The points (x, y) at which the altitude is a are given by the solution set of the equations $f(x, y) = a$.

Consider function $f(x, y)$ is given by

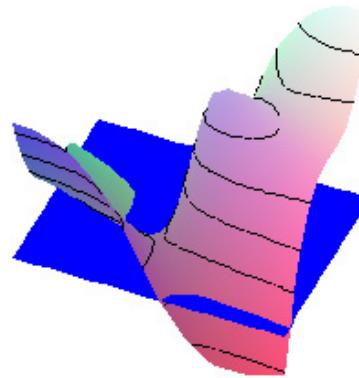
$$f(x, y) = \frac{3(1+x)^2 + xy^3 + y^2}{1+x^2+y^2} \quad (5.1)$$

which is chosen because the plots we will make show a nice “landscape”.

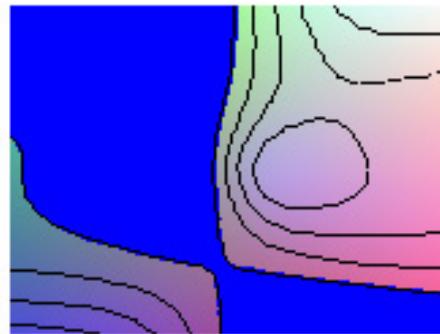
Suppose that a dam is built, and this landscape is flooded, up to an altitude 0.5 in the vertical distance units. This produces a lake that is shown below, in a top view; i.e., an aerial image:



The other lines on the land are the lines at other constant altitudes, specifically $x = 1.5$, $z = 2.5$, $z = 3.5$ and so on. On a topographic map, these curves are called *contour curves*. Here is a sort of side view showing the lake as a horizontal “slice” through the graph $z = f(x, y)$ at height $z = 0.5$:



If the water level is raised further, say to the altitude $z = 1.5$, everything will be flooded up to the next contour curve:

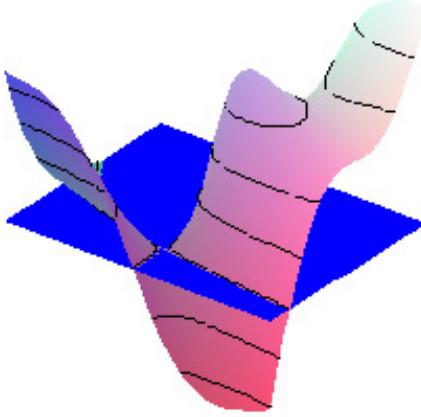


Comparing with the first picture, you clearly see that everything has been flooded up to the $z = 1.5$ contour curve. The isthmus joining the two tall hills is now submerged, and the two regions of the lake in the first graph have merged.

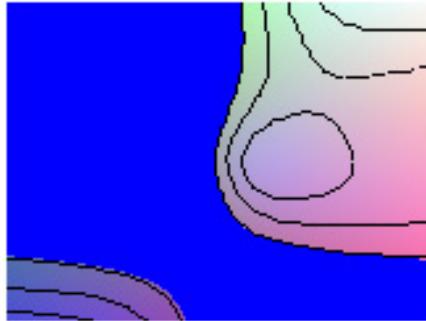
If you walked along the lake shore, your path would trace out the contour curve at $z = 1.5$ in

the first picture.

Here is a side view showing the lake at this level. It shows it as a horizontal “slice” through the graph $z = f(x, y)$ at height $z = 1.5$:



If the water level is raised further, to the height $z = 2.5$, the shore line moves up to the next contour line. Now a walk along the shoreline would trace out the path along the $x = 2.5$ contour line in the first picture. Here is the top view showing the lake at this stage:



The contour curves, which are the results of horizontal slices of the graph of $z = f(x, y)$, tell us a lot about the function $f(x, y)$. This section is an introduction to what they tell us.

Recall a definition from Chapter 3: Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the set of points $\mathbf{x} \in \mathbb{R}^n$ satisfying

$$f(\mathbf{x}) = c$$

is the *level set* of f at c . In other words, the level set of f at c is the solution set of the equation $f(\mathbf{x}) = c$.

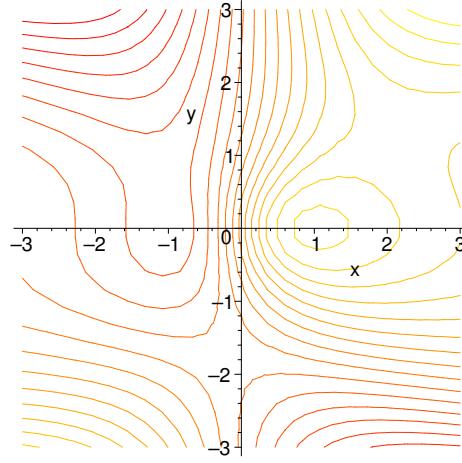
Let us consider the case $n = 2$ more closely. If we think of $f(x, y)$ as representing the altitude at the point with coordinates (x, y) , then the level set of f at height c is the set of all points at which the altitude is c . The level set at height c would be the “shore line” curve if the landscape were flooded up to an altitude c .

Now, here is a very important point, whose validity you can more or less see from the pictures we have displayed:

- Under normal circumstances, the level set of f at c will be a curve in the plane, possibly with several disconnected components.

It is for this reason that it is reasonable to refer to level sets as contour curves. We can plot a number of the level sets on a common graph. A *contour plot* of a function $f(x, y)$ is graph in which level curves of f are plotted at several different “altitudes” c_1, c_2, c_3, \dots . You have probably seen these on maps for hiking.

Here is a contour plot for the function “mountain landscape” function $f(x, y)$ in (5.1):



5.1.1 Implicit and explicit descriptions of planar curves

How could one go about actually drawing the contour curves starting from a formula like (5.1)? That is not so easy in general. You can see a hint of this in the convoluted form of the contour curves plotted here. The difficulty lies here:

- The description of contour curves given by the defining equation $f(x, y) = c$ is an *implicit description*: You have to solve the equation to find the points to plot. A *parametric description* of the curve $(x(t), y(t))$ is *explicit*: There is no equations to solve; simply evaluate $x(t)$ and $y(t)$ to get the points to plot.

To really appreciate this point, one has to understand the distinction between an implicit and an explicit description of a curve. The unit circle is a great example with which to start.

Let $f(x, y) = x^2 + y^2$ and let $c = 1$. Then the level set of f at height c is the set of points (x, y) satisfying

$$x^2 + y^2 = 1 . \quad (5.2)$$

This set, of course, is the unit circle. If we drew a contour plot of f showing the level curves at several altitudes “altitudes” c_1, c_2, c_3, \dots , you would see, several concentric circles.

The equation (5.2) is the *implicit* equation for the unit circle. To get an *explicit description*, just solve the equation (5.2) to find a *parameterization* of the solutions set. In the case of $x^2 + y^2 = 1$, we know how to do this: As t varies between 0 and 2π ,

$$(\cos t, \sin t) \quad (5.3)$$

traces out the unit circle in the counterclockwise direction. This is an explicit description since if you plug in any value of t , you get a point (x, y) on the unit circle, and as you vary t , you “sweep out” all such points. You can easily plot a large number of them and connect the dots to get a good plot of the circle.

Definition 65 (Implicit and explicit descriptions of curves in \mathbb{R}^2). *An equation of the form $f(x, y) = c$ provides an implicit description of a curve. A parameterization $\mathbf{x}(t)$, possibly with $t = x$ and with $y(x)$ given as an explicit function of x , provides an explicit description of a curve.*

Once one has an explicit description, it is easy to generate a plot, just by plugging in values for the parameter, plotting the resulting points, and “connecting the dots”. Passing from an implicit description to an explicit description involves *solving the equation* $f(x, y) = c$ to find an explicit parameterization of the solutions set. Generally, that is easier said than done.

Example 79 (From implicit to explicit by means of algebra). *Consider the function*

$$f(x, y) = 2x^2 - 2xy + y^2 .$$

The level curve at $c = 1$ for this function is given implicitly by the equation

$$2x^2 - 2xy + y^2 = 1 .$$

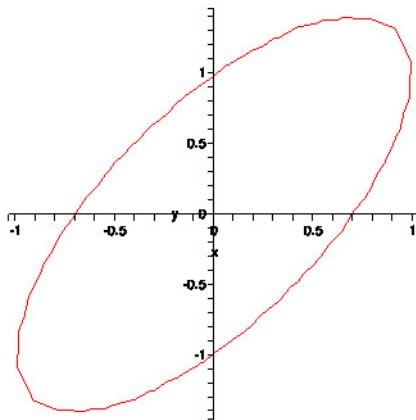
This can be rewritten as $y^2 - 2xy = 1 - 2x^2$. Completing the square in y , we have

$$(y - x)^2 = 1 - x^2 .$$

Therefore, we can solve for y as a function of x , finding

$$y(x) = x \pm \sqrt{1 - x^2} .$$

If we take x as the parameter, evidently y has a real value only for $-1 \leq x \leq 1$. It is now easy to plot the contour curve:



In this example, it was not so difficult passing from an implicit description to an explicit description; i.e., a parameterization, since the equation $f(x, y) = 1$ was quadratic in both x and y : We know how to deal with quadratic equations.

Example 80 (Bernouli Lemiscate). Consider the function

$$f(x, y) = (x^2 + y^2)^2 - 2(x^2 - y^2) ,$$

and consider its level set at 0; i.e., the set of all (x, y) such that $f(x, y) = 0$. It is not so easy to solve this equation for y as a functions of x , but we can easily parameterize the solution set if we use polar coordinates:

For any non-zero point (x, y) in the plane, we let r denote the distance from (x, y) to $(0, 0)$, and let θ be the angle such that if $(1, 0)$ is rotated counterclockwise through the angle θ , then the rotated unit vector points in the same direction as (x, y) .

It follows from this description that

$$r = \sqrt{x^2 + y^2} \quad x = r \cos \theta \quad \text{and} \quad y = r \sin \theta .$$

Making these substitutions, $f(x, y) = 0$ becomes $f(r \cos \theta, r \sin \theta) = 0$, or

$$r^4 - 2r^2(\cos^2 \theta - \sin^2 \theta) = r^4 - r^2 \cos(2\theta) = 0 .$$

We already know that $(0, 0)$, corresponding to $r = 0$, is one solution. For all of the others, $r \neq 0$, and we may divide by r^2 to obtain

$$r^2 = \cos(2\theta) . \tag{5.4}$$

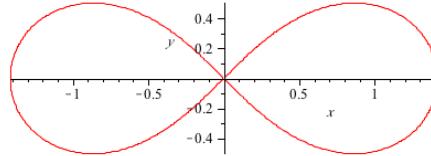
Now notice that the left hand side is always positive, but the right hand side is negative if $\pi/4 < \theta < 3\pi/4$, or if $5\pi/4 < \theta < 7\pi/4$. There are evidently no solutions of (5.4) for such θ . On the other hand, when $-\pi/4 < \theta < \pi/4$ or $3\pi/4 < \theta < 5\pi/4$, the right hand side is positive, and since r is always non-negative by definition (it is a distance), there will be two branches of the solution set, parameterized by

$$\mathbf{x}(\theta) := (r \cos \theta, r \sin \theta) = (\sqrt{\cos(2\theta)} \cos \theta, \sqrt{\cos(2\theta)} \sin \theta) .$$

for

$$-\frac{\pi}{4} < \theta < \frac{\pi}{4} \quad \text{and} \quad \frac{3\pi}{4} < \theta < \frac{5\pi}{4} .$$

One can now plug in values of θ , evaluate, plot the points, and connect the dots. The result is



The moral is that considering a different system of coordinates on the plane can make it much easier to find an explicit description of the curve.

This curve has a name: It is the Bernouli lemniscate, where lemniscate comes form the Latin word for ribbon.

However, changing to polar or other standard coordinate systems is not enough to solve all such problems. In general, we are going to need to extract information on the contour curves *directly* from the implicit description. Fortunately, what we have learned about gradients can help us to do this.

5.1.2 When is the contour curve actually a curve?

When does the equation $f(x, y) = f(x_0, y_0)$ actually define a continuously differentiable curve passing through (x_0, y_0) ? The following theorem gives the answer.

Theorem 65 (Implicit Function Theorem in \mathbb{R}^2). *Let f be a real valued function defined on an open set $U \subset \mathbb{R}^2$, and suppose that f is continuously differentiable in U . Then for any $\mathbf{x}_0 \in U$ such that $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, there is an $a > 0$ and a continuously differentiable curve $\mathbf{x}(t)$, $-a < t < a$, with $\mathbf{x}(0) = \mathbf{x}_0$ such that:*

- (1) *For all $|t| < a$, $f(\mathbf{x}(t)) = f(\mathbf{x}_0)$,*
- (2) *There is an $r > 0$ such that every \mathbf{x} satisfying $f(\mathbf{x}) = f(\mathbf{x}_0)$ and $\|\mathbf{x} - \mathbf{x}_0\| < r$ is of the form*

$$\mathbf{x} = \mathbf{x}(t)$$

for exactly one value of $-a < t < a$.

- (3) *For all $|t| < a$, $\mathbf{x}'(t) \neq \mathbf{0}$.*

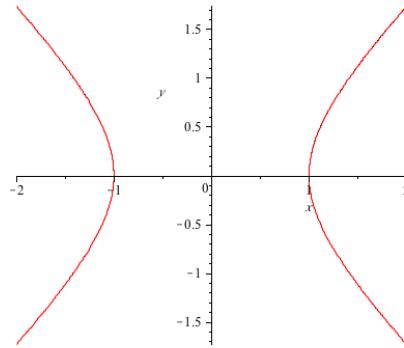
Remark 6. Note that (1) says that for every $t \in (-a, a)$, $f(\mathbf{x}(t)) = f(\mathbf{x}_0)$, while (2) says that the curve gives a one-to-one parameterization of the full solution set in $B_r(\mathbf{x}_0)$. and finally, (3) says that the parameterized curve does not stop, even instantaneously. Altogether, this parameterized curve gives a complete description of the solutions set of $f(\mathbf{x}) = f(\mathbf{x}_0)$ in the disk $B_r(\mathbf{x}_0)$, never stopping as it traces it out. There can be other “branches” of the solution set elsewhere, but not passing through $B_r(\mathbf{x}_0)$.

Example 81 (Branches of implicitly defined curves). Let $f(x, y) = x^2 - y^2$, and note that $\nabla f(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$.

In particular, taking $\mathbf{x}_0 = (1, 0)$, the Implicit Function Theorem assures us that all solutions of $f(\mathbf{x}) = f(\mathbf{x}_0)$ that are sufficiently close to \mathbf{x}_0 lie on a continuously differentiable curve through \mathbf{x}_0 . Indeed, $f(\mathbf{x}_0) = 1$, and the solution of $f(x, y) = 1$ is an hyperbola. This has two branches given by $x = \pm\sqrt{y^2 + 1}$. The branch of the hyperbola passing through $\mathbf{x}_0 = (1, 0)$ may be parameterized by

$$\mathbf{x}(t) = (\sqrt{t^2 + 1}, t).$$

Then, with $r := 2$, all solutions of $f(\mathbf{x}) = f(\mathbf{x}_0)$ with $\|\mathbf{x} - \mathbf{x}_0\| < r$ lie on this branch of the hyperbola; i.e., the curve parameterized by $\mathbf{x}(t)$. However, there are other solutions, on the other branch, further away.



Example 82 (A level set that is a single point). Let $f(x, y) = x^2 + y^2$. Then the level set of f at height 0 is just the single point $(0, 0)$ and not a curve. One could define $\mathbf{x}(t) = \mathbf{0}$ for all t , and thus satisfy (1) and (2) of the Implicit Function Theorem, but not (3).

Example 83 (A level curve that crosses itself). Let $f(x, y) = (x^2 + y^2)^2 - 2(x^2 - y^2)$ as in Example 80. Then the level set of f at height 0, which is the level set of f passing through $(0, 0)$, crosses the origin twice, as you see from the plot there, in two different directions. Here there are two level directions, and however you try to parameterize the level set, you must cover $(0, 0)$ twice. Part (3) of the Implicit Function Theorem cannot be satisfied, but then $\nabla f(0, 0) = \mathbf{0}$, so it does not apply.

There is a general conclusion to be drawn from the last example:

- If a contour curve of a continuously differentiable function f on \mathbb{R}^2 crosses itself at some point, that point must be a critical point of f . Otherwise, the Implicit Function Theorem would preclude the crossing.

We shall state and prove a more general version of the implicit function theorem, for functions from \mathbb{R}^n to \mathbb{R}^m later on. For now, let us focus on what such a theorem is good for, starting with the version stated above.

Let us see why the condition that f be continuously differentiable at \mathbf{x}_0 with $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$ is natural. Notice that since the gradient is continuous, for some $\delta > 0$, $\nabla f(\mathbf{x}) \neq \mathbf{0}$ as long as $\|\mathbf{x} - \mathbf{x}_0\| < \delta$.

Let $\mathbf{x}(s)$ be a smooth curve passing through \mathbf{x}_0 at $s = 0$, and we suppose that s is the arc length, so that $\mathbf{x}'(s) = \mathbf{T}(s)$, the unit tangent vector to the curve. Then by the chain rule,

$$\frac{d}{ds} f(\mathbf{x}(s)) = \nabla f(\mathbf{x}(s)) \cdot \mathbf{T}(s) .$$

Hence f will be constant along the curve if and only if $\mathbf{T}(s)$ is orthogonal to $\nabla f(\mathbf{x}(s))$ for each s . Since $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, there is some $a > 0$ so that $\nabla f(\mathbf{x}(s)) \neq \mathbf{0}$ for all $-a < s < a$. At each such s , we must have

$$\mathbf{T}(s) = \pm \frac{1}{\|\nabla f(\mathbf{x}(s))\|} \nabla f(\mathbf{x}(s))^{\perp} ,$$

and since the curve is smooth, it must be one sign or the other for all $s \in (-a, a)$. Thus, at each $s \in (-a, a)$, the unit tangent vector $\mathbf{T}(s)$ is completely specified, up to a choice of moving backwards or forwards along the curve.

Hence if we take the level curve of f at \mathbf{x}_0 provided by the Implicit Function Theorem, and re-parameterize it by arc-length, and if necessary, change the direction of travel (i.e., replace s by $-s$), we get a curve $\mathbf{x}(s)$ that solves the equation

$$\mathbf{x}'(s) = \frac{1}{\|\nabla f(\mathbf{x}(s))\|} \nabla f(\mathbf{x}(s))^{\perp} . \quad (5.5)$$

for all s near 0, and with $\mathbf{x}(0) = \mathbf{x}_0$. This is an example of an *ordinary differential equation*, and the theory of such equations could be applied to prove that there is a unique solution curve. Intuitively, you can build the solution curve up by starting out at \mathbf{x}_0 , and taking small steps always in the

direction of $\nabla f(\mathbf{x})^\perp$ at whatever point \mathbf{x} you are at. The direction in which you must move is completely specified if and only if $\nabla f(\mathbf{x}) \neq \mathbf{0}$. This reasoning can be made rigorous, but we will take a different approach that have the advantage of also being applicable in higher dimension to deal with, say, implicitly defined surfaces in \mathbb{R}^3 .

Nonetheless, it is worth remembering from this discussion that once one has proved the Implicit Function Theorem in \mathbb{R}^2 , one has proved the existence and uniqueness of a solution for the differential equation (5.5) with $\mathbf{x}(0) = \mathbf{x}_0$. The Implicit function Theorem plays an important role in the theory of ordinary differential equations. We summarize on important conclusion concerning (5.5) for later use:

Theorem 66. *Let f be a real valued function defined on an open set $U \subset \mathbb{R}^2$, and suppose that f is continuously differentiable in U . Then for any $\mathbf{x}_0 \in U$ such that $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, there is an $r > 0$ so that the level set of f at height $f(\mathbf{x}_0)$ in $B_r(\mathbf{x}_0)$ is a differentiable curve whose tangent line is parameterized by*

$$\mathbf{x}_0 + t\nabla f(\mathbf{x}_0)^\perp .$$

5.2 Constrained Optimization in Two variables

An *optimization problem* in two variables is one in which we are given a function $f(x, y)$, and a set D of *admissible points* in \mathbb{R}^2 , and we are asked to find either the maximum or minimum value of $f(x, y)$ as (x, y) ranges over D . In the previous chapter we have considered optimization problems on \mathbb{R}^2 . The new feature here is that we consider the case in which the set D has a boundary that might contain a maximizer, a minimizer, or both. We shall also consider the case in which D is a curve in \mathbb{R}^2 .

Recall that $(x_0, y_0) \in D$ minimizes f in D in case

$$f(x_0, y_0) \leq f(x, y) \quad \text{for all } (x, y) \text{ in } D ,$$

and $(x_1, y_1) \in D$ maximizes f in D in case

$$f(x_1, y_1) \geq f(x, y) \quad \text{for all } (x, y) \text{ in } D .$$

While in general it can be the case there is neither a maximum nor a minimum, we have seen in Chapter 3 that if D is bounded and closed, and if f is a continuous function, then f always has a minimum and a maximum on D .

To solve an optimization problem is to find all maximizers and minimizers, if any, and the corresponding maximum and minimum values. Our goal in this section is to explain a strategy for doing this. As long as D is closed and bounded, and f is continuous, minimizers and maximizers will exist, and our goal now is to compute them.

Recall that if $g(t)$ is a function of the single variable t , and we seek to maximize it on the closed bounded interval $[a, b]$, we proceed in two steps:

(1) We find all values of t in (a, b) at which $g'(t) = 0$. Hopefully there are only finitely many of these, say $\{t_1, t_2, \dots, t_n\}$.

(2) Compute $g(t_1), g(t_2), \dots, g(t_n)$, together with $g(a)$ and $g(b)$. The largest number on this finite list is the maximum value, and the smallest is the minimum value. The maximizers are exactly those numbers from among $\{t_1, t_2, \dots, t_n\}$ together with a and b , at which f takes on the maximum values, and similarly for the minimizers.

The reason this works is that if t belongs to the open interval (a, b) , and $g'(t) \neq 0$, it is possible to move either “uphill” or “downhill” while staying within $[a, b]$ by moving a bit to the right or the left, depending on whether the slope is positive or negative. *Hence no such point can be a maximizer or a minimizer.* This reasoning does not apply to exclude a or b , since at a , taking a step to the left is not allowed, and at b , taking a step to the right is not allowed. Thus, we have a short “list of suspects”, namely the set of solutions of $g'(t) = 0$, together with a and b , and the maximizers and minimizers are there on this list.

In drawing up this “list of suspects”, we are applying the Sherlock Holmes principle:

- *When you have eliminated the impossible, whatever else remains, however unlikely, is the truth.*

When all goes well, the elimination procedure reduces an *infinite* sets of suspects – all of the points in $[a, b]$ – to a *finite* list of suspects: $\{t_1, t_2, \dots, t_n\}$ together with a and b . Finding the minimizers and maximizers among a finite list of points is easy – simply compute the value of f at each point on the list, and see which ones give the largest and smallest values.

We now adapt this to two or more dimensions, focusing first on two. Suppose that D is a closed bounded domain. Let U be the interior of D and let B be the boundary. For example, if D is the closed unit disk

$$D = \{\mathbf{x} : \|\mathbf{x}\| \leq 1\} \quad (5.6)$$

we have

$$U = \{\mathbf{x} : \|\mathbf{x}\| < 1\} \quad \text{and} \quad B = \{\mathbf{x} : \|\mathbf{x}\| = 1\}. \quad (5.7)$$

Notice that in this case, the boundary consists of infinitely many points. *This is the big difference with the one dimensional case.* Hence need a “sieve” to filter the boundary B and eliminate the boundary points that cannot possibly be maximizers or minimizers. Hopefully, one a finite number of “suspects” will remain.

5.2.1 Lagrange’s criterion for optimizers on the boundary

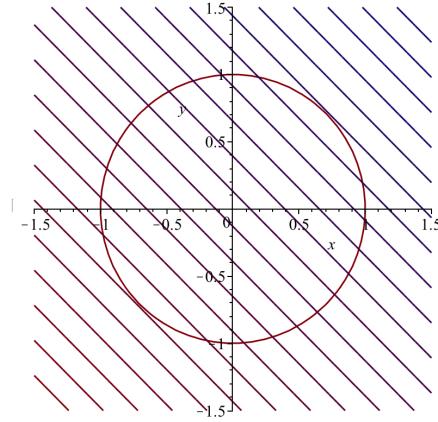
We begin with a problem so simple it can be solved without any calculus at all, but in which the geoemetry of the method to be introduced will be particularly clear. For a really simple choice of the domain D , we take D to be the unit disk (5.6) so that its interior U and boundary B are given by (5.7). For a really simple choice of the the function $f(x, y)$, we take $f(x, y) = x + y$.

Notice that $f(x, y) = (1, 1) \cdot (x, y)$, and then by the Cauchy-Schwarz inequality, for $(x, y) \in D$,

$$f(x, y) \leq \|(1, 1)\| \|(x, y)\| = \sqrt{2} \sqrt{x^2 + y^2} \leq \sqrt{2},$$

and there is equality if and only if $(x, y) = (1/\sqrt{2}, 1/\sqrt{2})$. Hence this is the only maximizer. Similar reasoning shows that $(-1/\sqrt{2}, -1/\sqrt{2})$ is the only minimizer.

Now let's approach this problem from a different point of view to develop a more broadly applicable method. The plot below shows a number of level curves of f – there are lines since f is linear – superimposed on a plot of B .



Notice that except at two points on B , the tangent lines to B “cut across” level curves of f . Hence, by moving backwards or forwards along B at such a point, one can move either “uphill” “downhill”, and no such point can be a maximizer or a minimizer. The only points that can possibly be a minimizer or a maximizer are those at which the tangent line to B is coincides with the tangent line to the level curve of f through that point. *Our next goal is to express this geometric condition in terms of equations that we can solve to find possible minimizers or maximizers.*

Define $g(x, y) := x^2 + y^2 - 1$, and note that $(x, y) \in B$ if and only if $g(x, y) = 0$. Hence B is the 0-level set of g . Since $\nabla g(x, y) = 2(x, y)$, the only critical point of g is $(0, 0)$, and in particular, the conditions of the Implicit function Theorem are met at each point $(x, y) \in B$. (Of course, we know how to explicitly parameterize a circle, but let's try to reason in a way that is generally applicable.) Therefore, by Theorem 66 applied to g at any $\mathbf{x}_0 \in B$, the tangent line to B at \mathbf{x}_0 is parameterized by $\mathbf{x}_0 + t\nabla g(\mathbf{x}_0)^\perp$. Since f has no critical points, Theorem 66 applies to f at \mathbf{x}_0 , and the tangent line to the level curve of f through \mathbf{x}_0 is parameterized by $\mathbf{x}_0 + t\nabla f(\mathbf{x}_0)^\perp$.

These two tangent lines coincide if and only if $\nabla g(\mathbf{x}_0)^\perp$ is a multiple of $\nabla f(\mathbf{x}_0)^\perp$, and evidently this is the case if and only if $\nabla g(\mathbf{x}_0)$ is a multiple of $\nabla f(\mathbf{x}_0)$. That is the tangent lines coincide if and only if for some number λ ,

$$\nabla f(x, y) = \lambda \nabla g(x, y). \quad (5.8)$$

The vector equation (5.8) is a system of two scalar equation. Combining it with the equation for B , namely $g(x, y) = 0$, we have system of 3 equations in the 3 variables x, y, λ . *The only points \mathbf{x}_0 of B that can possibly be a maximizer or a minimizer of f on B are those that satisfy (5.8) for some λ .* (Note that the third equation is satisfied by assumption since we assume that $\mathbf{x}_0 \in B$.)

In our simple example,

$$\nabla f(x, y) = (1, 1) \quad \text{and} \quad \nabla g(x, y) = 2((x, y)).$$

Together with $g(x, y) = 0$, we have the system

$$\begin{aligned} 1 &= 2\lambda x \\ 1 &= 2\lambda y \\ 1 &= x^2 + y^2 \end{aligned}$$

From either of the first two equations, $\lambda \neq 0$, and then $2\lambda x = 2\lambda y$ reduces to $x = y$. The third equation then gives $2x^2 = 1$, so that $x = \pm 1/\sqrt{2}$. Since $y = x$, we have two possible points, namely

$$(1/\sqrt{2}, 1/\sqrt{2}) \quad \text{and} \quad -(1/\sqrt{2}, 1/\sqrt{2}).$$

Evidently, the first is the maximizer on B , and the second is the minimizer on B .

Finally we return to the consideration of f on D . If a maximizer or a minimizer of f in D is not on B , it must be in the interior U , and then it must be a critical point of f . But f has no critical points at all. Hence the maximizer and minimizer that we found on B are also the maximizer and minimizer of f on D , just as we found using the Cauchy-Schwarz inequality.

This second method method we used to solve this problem was invented by Lagrange, and the next thoerem presents it in a general and efficient form in \mathbb{R}^2 , in whch the third variable λ is eliminated right away.

Theorem 67 (Lagrange's Theorem). *Suppose that f and g are two functions on \mathbb{R}^2 with continuous first order partial derivatives. Let B denote the level curve of g given by $g(x, y) = 0$.*

Then if \mathbf{x}_0 is a maximizer or minimizer of f on B ,

$$\det \begin{pmatrix} \nabla f(\mathbf{x}_0) \\ \nabla g(\mathbf{x}_0) \end{pmatrix} = 0 \quad \text{and} \quad g(\mathbf{x}_0) = 0. \quad (5.9)$$

Proof. We must show that if \mathbf{x}_0 does not satisfy both equations in (5.9), then \mathbf{x}_0 cannot be a maximizer or a minimizer of f on B . If $g(\mathbf{x}_0) \neq 0$, then \mathbf{x}_0 is not even on B , and we need only consider points with $g(\mathbf{x}_0) = 0$, but

$$\det \begin{pmatrix} \nabla f(\mathbf{x}_0) \\ \nabla g(\mathbf{x}_0) \end{pmatrix} \neq 0, \quad (5.10)$$

and show that they cannot be maximizers. Clearly, if either $\nabla f(\mathbf{x}_0) = \mathbf{0}$ or $\nabla g(\mathbf{x}_0) = \mathbf{0}$, then the first equation in (5.9) is satisfied, and (5.10) is not. Therefore, whenever (5.10) is satisfied, both $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$ and $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$.

Then, Theorem 66 applies to g , and B is truly given by a parameterized curve through \mathbf{x}_0 , and the tangent line to B at \mathbf{x}_0 is parameterized by $\mathbf{x}_0 + t\nabla g(\mathbf{x}_0)^\perp$. Likewise, since $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, and by Theorem 66, the level curve of f through \mathbf{x}_0 is truly given by a parameterized curve, and the tangnet line to it is paprameterized by $\mathbf{x}_0 + t\nabla f(\mathbf{x}_0)^\perp$.

Moreover, when (5.10) is satisfied, $\nabla f(\mathbf{x}_0)$ is *not* a multiple of $\nabla g(\mathbf{x}_0)$, and therefore, $\nabla f(\mathbf{x}_0)^\perp$ is *not* a multiple of $\nabla g(\mathbf{x}_0)^\perp$. Hence the curve B cuts across level curves of f , and \mathbf{x}_0 cannot possibly be a minimizer or a maximizer. Altogether, the only points in B that can possibly be maximizers or minimizers are those that satsify (5.9). \square

5.2.2 Application of Lagrange's Theorem

Summarizing, our strategy for finding maximizers and minimizers of a continuously differentiable function f in a region D bounded by a curve given by an equation of the form $g(x, y) = 0$ is:

- (1) Find all critical points of f in U , the interior of D .
- (2) Find all points on B , the boundary of D , at which (5.9) holds.
- (3) The combined list of points found in (1) and (2) is a comprehensive list of potential maximizers and minimizers. Hopefully it is a finite list. In this case, evaluate f at each of them, and see which produce the largest and smallest values. Case closed.

Example 84 (Finding minimizers and maximizers). Let $f(x, y) = x^4 + y^4 + 4xy$. Let D be the closed disk of radius 4 centered on the origin. We now find the maximizers and minimizers of f in D .

We can write the equation for the boundary in the form $g(x, y) = 0$ by putting

$$g(x, y) = x^2 + y^2 - 16 .$$

Part (1): First, we look for the critical points of f . We have already examined f in Example 62. There, we found that f has exactly 3 critical points in all of \mathbb{R}^2 , and all of them happen to be in the interior of D . They are

$$(0, 0) \quad (1, -1) \quad \text{and} \quad (-1, 1) . \quad (5.11)$$

Part (2): Next, we look for solutions of (5.9) and $g(x, y) = 0$. Since

$$\nabla f(x, y) = 4(x^3 + y, y^3 + x) \quad \text{and} \quad \nabla g(x, y) = 2(x, y) ,$$

$$\det \begin{pmatrix} \nabla f \\ \nabla g \end{pmatrix} = 8 \det \begin{pmatrix} x^3 + y & y^3 + x \\ x & y \end{pmatrix} = 8(x^3y + y^2 - y^3x - x^2) .$$

Hence (5.9) gives us the equation $x^3y + y^2 - y^3x - x^2 = 0$. Combining this with $g(x, y) = 0$ we have the system of equations

$$\begin{aligned} x^3y + y^2 - y^3x - x^2 &= 0 \\ x^2 + y^2 - 16 &= 0 \end{aligned}$$

The rest is algebra. The key to solving this system of equations is to notice that $x^3y + y^2 - y^3x - x^2$ can be factored:

$$x^3y + y^2 - y^3x - x^2 = (x^2 - y^2)(xy - 1) ,$$

so that the first equation can be written as $(x^2 - y^2)(xy - 1) = 0$. Now it is clear that either $x^2 - y^2 = 0$, or else $xy - 1 = 0$.

Suppose that $x^2 - y^2 = 0$. Then we can eliminate y from the second equation, obtaining $2x^2 = 16$, or $x = \pm 2\sqrt{2}$. If $y^2 = x^2$, then $y = \pm x$, so we get 4 solutions of the system this way:

$$(\pm 2\sqrt{2}, \pm 2\sqrt{2}) . \quad (5.12)$$

On the other hand, if $xy - 1 = 0$, $y = 1/x$, and eliminating y from the second equation gives us

$$x^2 + x^{-2} - 16 = 0 \quad (5.13)$$

Multiplying through by x^2 , and writing $u = x^2$, we get $u^2 - 16u = -1$, so $u = 8 \pm \sqrt{63}$. Since $u = x^2$, there are four values of x that solve (5.13), namely $\pm\sqrt{8 \pm \sqrt{63}}$. The corresponding y values are given by $y = 1/x$. We obtain the final 4 solutions of (5.9):

$$(a, 1/a) \quad \text{with} \quad a = \pm\sqrt{8 \pm \sqrt{63}}. \quad (5.14)$$

Part (3): We now round up and interrogate the suspects. There are 11 of them: Three from (5.11), four from (5.12), and four from (5.14).

Now we interrogate the suspects: At the three critical points we have

$$f(0, 0) = 0, \quad f(1, -1) = f(-1, 1) = -2.$$

From the first group of four boundary points,

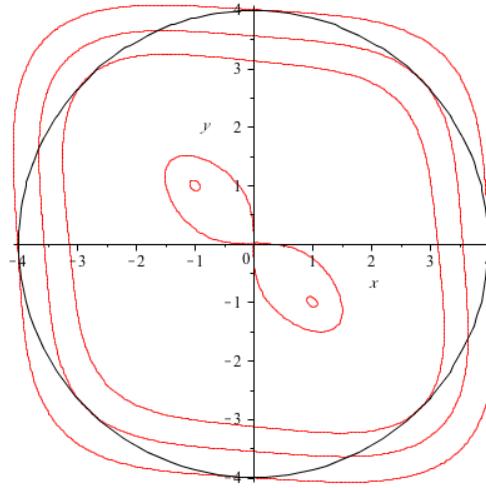
$$f(-2\sqrt{2}, 2\sqrt{2}) = f(2\sqrt{2}, -2\sqrt{2}) = 96 \quad \text{and} \quad f(2\sqrt{2}, 2\sqrt{2}) = f(-2\sqrt{2}, -2\sqrt{2}) = 160$$

Form the second group of four boundary points,

$$f(a, 1/a) = 258 \quad \text{for} \quad a = \pm\sqrt{8 \pm \sqrt{63}}.$$

Evidently, the maximum value of f in D is 258, and the corresponding maximizers are the four points in (5.14). Also evidently, the minimum value of f in D is -2 , and the corresponding maximizers are the two points $(-1, 1)$ and $(1, -1)$. Notice that the maximizers lie on the boundary, and the minimizers lie in the interior.

Here is a graph showing the boundary curve $g(x, y) = 0$, which is the circle of radius 4, and three contour curves of f , namely the contour curves at levels 258, 160, 96, 0, and -1.95 . These are the levels that showed up in our interrogation of the suspects, except that we have used the level -1.95 instead of -2 in the graphplot, since there are just two points, $(-1, 1)$ and $(1, -1)$, for which $f(x, y) = -2$, and that would not plot well.



Notice that the plot shows 4 points of tangency along the contour curve at altitude 258, and 2 points of tangency along each of the contour curves at altitudes 160 and 96, corresponding exactly to our computations.

Sometimes, we are only concerned with minimizers or maximizers on some curve. Then things are even simpler. As long as the curve is given by an equation $g(x, y) = 0$, where g is continuously differentiable and has no critical points on the level set at height 0, Lagrange's Theorem may be applied.

In our next example, we use Theorem 67 to compute the distance between a point and a parabola. The idea is to write the equation for the parabola in the form $g(x, y) = 0$. For instance, if the parabola is given by $y = x^2/2$, we can take $g(x, y) = x^2 - 2y$.

Suppose for example that we want to find the distance from $(3, -3/2)$ to this parabola. The square of the distance from any point (x, y) to $(3, -3/2)$ is

$$(x - 3)^2 + (y + 3/2)^2 .$$

To find the point on the parabola that is closest to $(3, -3/2)$, we use Theorem 1 to find the point (x_0, y_0) on the parabola that minimizes $(x - 3)^2 + (y + 3/2)^2$ – this is the point on the parabola that is closest to $(3, -3/2)$. Then, by definition, the distance from $(3, -3/2)$ to the parabola is the distance from $(3, -3/2)$ to this closest point.

Before proceeding to the calculations, note that the parabola is closed, but not bounded. Therefore, minima and maxima are not guaranteed to exist. Indeed, the parabola reaches upward for ever and ever, so there are points on the parabola that are arbitrarily far away from $(3, -3/2)$. That is, there is no furthest point.

But it is geometrically clear that there is a closest point. Indeed, since $(0, 0)$ is on the parabola, and the distance of this point from $(3, -3/2)$ is $3\sqrt{5}/2$, we only need to look for the minimum on the part of the parabola that lies in the closed disk of radius $3\sqrt{5}/2$ centered on $(3, -3/2)$. This is closed and bounded, so a minimizer will exist.

This particular problem can be solved using single variable methods: Substituting $y = x^2/2$ into $f(x, y)$ we see that the squared distance from $(x, x^2/2)$ to $(3, -3/2)$ is

$$(x - 3)^2 + (x^2/2 + 3/2)^2 = \frac{1}{4}(x^4 + 10x^2 - 24x + 45) .$$

Taking the derivative with respect to x , if x is a minimizer of this expression,

$$x^3 - 5x - 6 = 0 .$$

This is a cubic equation, but a very simple one: It is easy to see that $x = 1$ is one root, and hence $(x - 1)$ divides that cubic. Factoring it we find $x^3 - 5x - 6 = (x - 1)(x^2 + x + 6)$. Since $x^2 + x + 6 = (x + 1/2)^2 + 23/4 \geq 23/4$, the only root is $x = 1$. Since there is a minimizer, the distance is minimized at $x = 1$, and hence $y = 1/2$. Thus, the closest point on the parabola is $(1, 1/2)$, and the distance is $2\sqrt{2}$. We shall now look at this same problem from the point of view of Theorem 67:

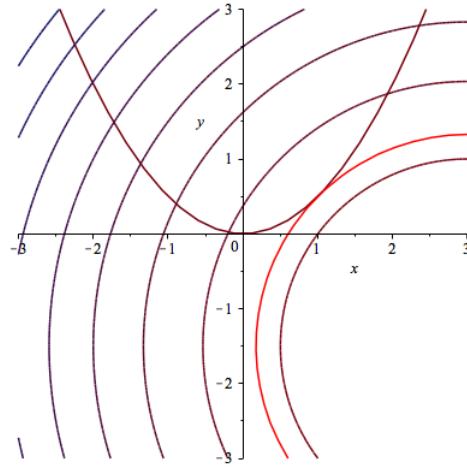
Example 85 (Finding the distance to a parabola). Consider the parabola $y = x^2/2$, and the point $(3, -3/2)$. As explained above, to find the point on the parabola that is closest to $(3, -3/2)$, we minimize

$$f(x, y) = (x - 3)^2 + (y + 3/2)^2$$

on the curve $g(x, y) = 0$ where

$$g(x, y) = x^2 - 2y .$$

Here is a graph showing the parabola, and some of the contour curves of f . As you can see, there is exactly one place at which the parabola is tangent to the contour curves. Therefore, when we set up and solve the system consisting of (5.9) and $g(x, y) = 0$, we will find only one solution, as we found in our single variable approach.



To apply Theorem 1, we compute $\nabla f(x, y) = 2(x - 3, y + 3/2)$ and $\nabla g(x, y) = 2(x, -1)$. Therefore, (5.9) reduces to

$$0 = \det \begin{pmatrix} x - 3 & y + 3/2 \\ x & -1 \end{pmatrix} = 3 - \frac{5}{2}x - xy .$$

Now using $y = x^2/2$, to eliminate y , we are left with

$$x^3 + 5x - 6 = 0 ,$$

which is the same cubic equation that we encountered above. It has only one real root, namely $x = 1$, and hence the closest point on the parabola is $(1, 1/2)$, and the distance is $2\sqrt{2}$, as we found above.

The reason we could use single variable methods to solve this problem was that it was easy to parameterize the parabola. But such explicit parameterizations can difficult to work with in general. The strength of Theorem 67 is that it allows us to work with implicit descriptions of the boundary. With more variables, this will be even more useful.

5.2.3 Optimization for regions with a piecewise smooth boundary

In our next example, we use Theorem 67 to solve an optimization problem in which the boundary is given in two pieces. This introduces one new feature into the method, as we shall see.

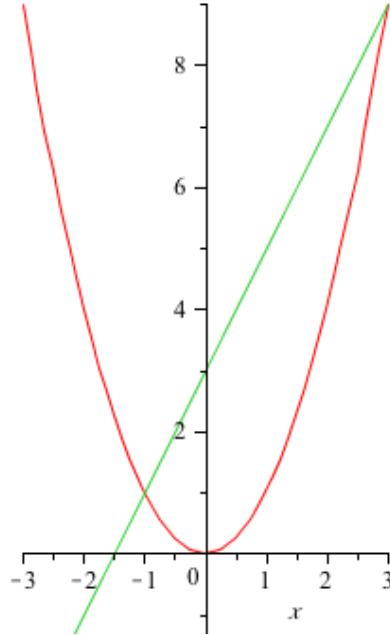
Example 86 (An optimization problem with a two piece boundary). Let D be the region consisting of all points (x, y) satisfying

$$x^2 \leq y \leq 3 + 2x .$$

Let $f(x, y) = x^2y - 3x$. We seek to find the minimum and maximum values of f on D , and find all minimizers and maximizers.

Note that D is a closed, bounded region, and hence minimizers and maximizers in D do exist. Next, we compute $\nabla f(x, y) = (2xy - 3, x^2)$. There are clearly no solutions of $\nabla f(x, y) = (0, 0)$. Hence, the minimizers and maximizers will lie on the boundary.

The first step towards finding them is to plot the two curves defining the boundary of the region D :



The region D lies above the parabola and below the line. Note that the boundary does not have a tangent direction where the line and parabola meet. Therefore, we have to include these points in our suspect list: Since there is no tangent at these points, the tangency condition cannot possibly exclude them.

To find the intersection points, eliminate y from $x^2 = y$ and $y = 3 + 2x$, obtaining

$$x^2 = y = 3 + 2x$$

and then since $x^2 - 2x - 3 = 0$ has the solutions $x = -1, 3$, we conclude:

- The boundary consists of the points of the parabola $y = x^2$ or on the line $y = 3 + 2x$ with $-1 \leq x \leq 3$ in either case. The points where the line meets the parabola, namely $(-1, 1)$ and $(3, 9)$, must be included in the suspect list.

We now proceed to check the tangency condition, finding points at which the tangent line to the contours curves of f are parallel to the tangent lines to the constraint curve. However, as explained above, the only relevant points are those whose x coordinate lies in the range $-1 < x < 3$.

Let us first deal with the parabola. We take $g(x, y) = x^2 - y$. Then the Lagrange condition $\nabla f = \lambda \nabla g$ implies that

$$2x^3 = 3 - 2xy$$

Combining this with $y = x^2$, we conclude $4x^3 = 3$. This equation has the unique real solution $x = (3/4)^{1/3}$. Hence one possible point to consider is

$$\mathbf{x}_1 = ((3/4)^{1/3}, (3/4)^{2/3}) .$$

Notice the x coordinate is in the relevant range.

Now let us consider the linear part of the boundary. This time we take $g(x, y) = y - 2x - 3$. Then the Lagrange condition $\nabla f = \lambda \nabla g$ implies that

$$-2x^2 = 2xy - 3 .$$

Substituting in $y = 3 + 2x$, this becomes $2x^2 + 2x = 1$, which has the two roots

$$x = -\frac{1 \pm \sqrt{3}}{2} .$$

Of these, only $(\sqrt{3} - 1)/2$ lies in the range of interest, $-1 \leq x \leq 3$. Using $y = 3 + 2x$, we find our next possible point:

$$\mathbf{x}_2 = ((\sqrt{3} - 1)/2, \sqrt{3} + 2) .$$

As explained above, we must included the points where the line and parabola meet, so we round out the list of suspects with

$$\mathbf{x}_3 = (-1, 1) \quad \text{and} \quad \mathbf{x}_4 = (3, 9) .$$

We now compute

$$\begin{aligned} f(\mathbf{x}_1) &= -\frac{9}{16}3^{1/3}4^{2/3} \approx -2.044260667 \\ f(\mathbf{x}_2) &= -2 - \frac{3}{2}\sqrt{3} \approx -0.598076212 \\ f(\mathbf{x}_3) &= 4 \\ f(\mathbf{x}_4) &= 72 \end{aligned}$$

We see that \mathbf{x}_1 is the minimizer, and \mathbf{x}_4 is the maximizer. Leaving the corners off the suspect list would have led to a gross underestimation of the maximum.

5.3 The Implicit Function Theorem via the Inverse Function Theorem

5.3.1 Inverting coordinate tranformations

Recall that in Example 80, we succeeded in parameterizing the Bernouli lemniscate by writing the equation that defines it in polar coordinates. We shall now prove the Implicit Function Theorem

by constructing a well-behaved system of coordinates in which one of the coordinate functions is a essentially f itself.

Here is the idea: Let f be a continuously differentiable function on some open set $U \subset \mathbb{R}^2$, and suppose that at $\mathbf{x}_0 \in U$, $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$. Define the unit vectors

$$\mathbf{u}_1 = \frac{1}{\|\nabla f(\mathbf{x}_0)\|} \nabla f(\mathbf{x}_0) \quad \text{and} \quad \mathbf{u}_2 = \mathbf{u}_1^\perp \quad (5.15)$$

so that $\{\mathbf{u}_1, \mathbf{u}_2\}$ is an orthonormal basis for \mathbb{R}^2 . We next define two functions

$$u(\mathbf{x}) = \frac{1}{\|\nabla f(\mathbf{x}_0)\|} (f(\mathbf{x}) - f(\mathbf{x}_0)) \quad \text{and} \quad v(\mathbf{x}) = \mathbf{u}_2 \cdot (\mathbf{x} - \mathbf{x}_0) . \quad (5.16)$$

Notice that

$$u(\mathbf{x}) = c \iff f(\mathbf{x}) = d \quad \text{where } d = c\|\nabla f(\mathbf{x}_0)\| + f(\mathbf{x}_0)$$

so that every level set of f is a level set of u , and *vice-versa*.

Also, by construction, $u(\mathbf{x}_0) = v(\mathbf{x}_0) = 0$, so if we define a transformation \mathbf{f} from U to \mathbb{R}^2 by $\mathbf{f}(\mathbf{x}) := \begin{bmatrix} u(\mathbf{x}) \\ v(\mathbf{x}) \end{bmatrix}$, we have $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$.

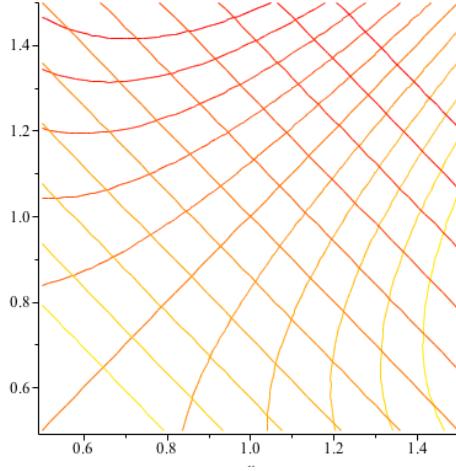
For example, let $f(x, y) = x^2y - xy^2$ and $\mathbf{x}_0 = (1, 1)$. Then $\nabla f(\mathbf{x}_0) = (1, -1)$ so that

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, -1) \quad \text{and} \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}(-1, -1) .$$

Then, since $f(\mathbf{x}_0) = 0$,

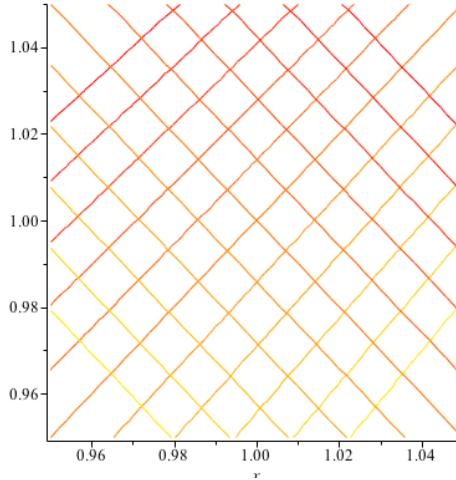
$$u(x, y) = \frac{1}{\sqrt{2}}(x^2y - xy^2) \quad \text{and} \quad v(x, y) = -\frac{1}{\sqrt{2}}(x + y - 2) .$$

Here is a plot showing contour lines of $u(\mathbf{x})$ and $v(\mathbf{x})$ in the unit square centered on $\mathbf{x} = \mathbf{x}_0$:



The contour curves of $v(\mathbf{x})$ are straight lines while the contour curves of $u(\mathbf{x})$ have curvature. However, if we “zoom in” more, the contour curves of $u(\mathbf{x})$ will look more and more like straight lines.

Here is a “zoomed in” contour plot showing contour curves of $u(\mathbf{x})$ and $v(\mathbf{x})$ in the square of side length 0.1 centered on \mathbf{x}_0 :



As you can see in these plots, the functions $u(\mathbf{x})$ and $v(\mathbf{x})$ define a system of coordinates in a neighborhood of \mathbf{x}_0 . Thus, we can think of the functions $u(\mathbf{x})$ and $v(\mathbf{x})$ as coordinate functions on a neighborhood of \mathbf{x}_0 . As we have noted above, every contour curve of $u(\mathbf{x})$ is also a contour curve of $f(\mathbf{x})$, though generally for a different altitude. Still, each coordinate curve $u = c$ in our coordinate system is a contour curve of f .

To make effective use of a coordinate system, we not only need to know the coordinate functions $u(x, y)$ and $v(x, y)$ that specify the new coordinates u and v in terms of x and y , we need the *inverse functions* $x(u, v)$ and $y(u, v)$ that express x and y in terms of u and v .

Indeed, when in Example 80 we used the polar coordinate system to parameterize the Bernoulli Lemiscate, we did this by substituting

$$x(r, \theta) = r \cos \theta \quad \text{and} \quad y(r, \theta) = r \sin \theta$$

into $f(x, y) = (x^2 + y^2)^2 - 2(x^2 - y^2) = 0$ to obtain an equation relating r and θ .

In the case of the polar coordinate system, it is easy to solve for x and y as a function of the alternate coordinates r and θ . But for other coordinate transformations $(u(x, y), v(x, y))$ it may not be possible to explicitly solve for $(x(u, v), y(u, v))$. In such a case, how can we know that such an inverse formula even exists?

The key to this is the linear approximation

$$(u, v) = \mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) , \quad (5.17)$$

valid for \mathbf{x} close to \mathbf{x}_0 .

Treating the approximate equality as if it were exact, as in Newton's method, we can now solve for \mathbf{x} , just as in Newton's method. We find, provided that $[D\mathbf{f}(\mathbf{x}_0)]$ is invertible,

$$\mathbf{x} \approx \mathbf{x}_0 + [D\mathbf{f}(\mathbf{x}_0)]^{-1}(u, v) . \quad (5.18)$$

Thus under the condition that $[D\mathbf{f}(\mathbf{x}_0)]$ is invertible, there is a good approximate inverse of our coordinate transformation, at least in a sufficiently small neighborhood of \mathbf{x}_0 such that (5.17) is a good approximation. One might therefore hope that under this condition, there is an *exact* inverse. When \mathbf{f} is continuously differentiable, this is indeed the case, as we shall show.

First of all, the way we have constructed the functions $u(\mathbf{x})$ and $v(\mathbf{x})$ guarantees that $[D_{\mathbf{f}}(\mathbf{x}_0)]$ is invertible. Indeed, the Jacobian of \mathbf{f} , $[D_{\mathbf{f}}]$ is then given by $[D_{\mathbf{f}}(\mathbf{x})] = \begin{bmatrix} \nabla u(\mathbf{x}) \\ \nabla v(\mathbf{x}) \end{bmatrix}$. We compute

$$\nabla u(\mathbf{x}) = \frac{1}{\|\nabla f(\mathbf{x}_0)\|} \nabla f(\mathbf{x})$$

so that $\nabla u(\mathbf{x}_0) = \mathbf{u}_1$. Also, $\nabla v(\mathbf{x}) = \mathbf{u}_2$ for all \mathbf{x} .

Therefore $[D_{\mathbf{f}}(\mathbf{x}_0)] = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$, and since \mathbf{u}_1 and \mathbf{u}_2 are orthonormal, $\text{rank}([D_{\mathbf{f}}(\mathbf{x}_0)]) = 2$, so that $[D_{\mathbf{f}}(\mathbf{x}_0)]$ is invertible. (Better yet, it is an orthogonal matrix, so its inverse is its transpose.) Since the set of invertible matrices is open, there is an $r > 0$ such that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq r \Rightarrow [D_{\mathbf{f}}(\mathbf{x})] \text{ is invertible}.$$

We now state an important theorem, namely the *Inverse function Theorem* in \mathbb{R}^n , which gives the n -dimensional analog of what we have been discussing in pictures for $n = 2$:

Theorem 68 (The Inverse Function Theorem in \mathbb{R}^n). *Let \mathbf{f} be a continuously differentiable function on some open set $U \subset \mathbb{R}^n$ with values in \mathbb{R}^n , and suppose that at $\mathbf{x}_0 \in U$, $[D_{\mathbf{f}}(\mathbf{x}_0)]$ is invertible. Then there is an open set V containing \mathbf{x}_0 and an open set W containing $\mathbf{f}(\mathbf{x}_0)$ such that \mathbf{f} is a one-to-one transformation of V onto W , and hence so that the inverse function \mathbf{f}^{-1} from W to V is well defined. Moreover, \mathbf{f}^{-1} is differentiable at each $\mathbf{u} \in W$, and $[D_{\mathbf{f}}(\mathbf{x})]$ is invertible everywhere on V , and*

$$[D_{\mathbf{f}^{-1}}(\mathbf{u})] = [D_{\mathbf{f}}(\mathbf{f}^{-1}(\mathbf{u}))]^{-1}.$$

Before proving the Inverse Function Theorem, we first show how it implies the Implicit Function Theorem.

5.3.2 From the Inverse Function Theorem to the Implicit Function Theorem

Proof of the Implicit Function Theorem for \mathbb{R}^2 , Theorem 65. Let f be continuously differentiable on $U \subset \mathbb{R}^2$. Given $\mathbf{x}_0 \in U$, construct $u(\mathbf{x})$ and $v(\mathbf{x})$ as in (5.16), and then form $\mathbf{f}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}))$. Note that by construction, $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$.

The Inverse Function Theorem provides an open set W containing $\mathbf{0}$ and an open set V containing \mathbf{x}_0 such that \mathbf{f} is one-to-one from V onto W , and hence invertible, and thus has an inverse transformation \mathbf{f}^{-1} from W to V , and moreover, \mathbf{f}^{-1} is differentiable everywhere on W . By construction $[D_{\mathbf{f}}(\mathbf{x}_0)] = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$ where \mathbf{u}_1 and \mathbf{u}_2 are given in (5.15). Since $\{\mathbf{u}_1, \mathbf{u}_2\}$ is orthonormal,

$[\mathbf{u}_1, \mathbf{u}_2]$ is an orthogonal matrix and by Theorem 57, $\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$ is its inverse. Therefore, by the Inverse Function Theorem, $[D_{\mathbf{f}^{-1}}(\mathbf{x}_0)] = [\mathbf{u}_1, \mathbf{u}_2]$.

Since W is open and contains $\mathbf{0}$, for some $a > 0$, W contains the open ball $\{\mathbf{u} : \|\mathbf{u}\| < a\}$. Since \mathbf{f} is continuous and $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$, there is an $r > 0$ such that

$$\|\mathbf{x} - \mathbf{x}_0\| < r \Rightarrow \mathbf{f}(\mathbf{x}) \in \{\mathbf{u} : \|\mathbf{u}\| < a\}. \quad (5.19)$$

Now define the curve

$$\mathbf{x}(v) := \mathbf{f}^{-1}(0, v) \quad \text{for } -a < v < a .$$

This is a differentiable curve since \mathbf{f}^{-1} is differentiable. Also, $\mathbf{x}(0) = \mathbf{f}^{-1}(\mathbf{0}) = \mathbf{x}_0$ since $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$.

Recall the definition (5.16) of $\mathbf{f}(\mathbf{x}) = (\mathbf{u}(\mathbf{x}), \mathbf{v}(\mathbf{x}))$:

$$u(\mathbf{x}) = \frac{1}{\|\nabla f(\mathbf{x}_0)\|} (f(\mathbf{x}) - f(\mathbf{x}_0)) \quad \text{and} \quad v(\mathbf{x}) = \mathbf{u}_2 \cdot (\mathbf{x} - \mathbf{x}_0) .$$

For all $-a < v < a$, $\mathbf{f}(\mathbf{x}(v)) = \mathbf{f}(\mathbf{f}^{-1}(0, v)) = (0, v)$, so that $u(\mathbf{x}(v)) = 0$, and hence $f(\mathbf{x}(v)) = f(\mathbf{x}_0)$: The function f is indeed constant on the curve $\mathbf{x}(v)$, $-a < v < a$ passing through \mathbf{x}_0 , so that item (1) of the theorem is proved. Next, *all* solutions of $f(\mathbf{x}) = f(\mathbf{x}_0)$ with $\|\mathbf{x} - \mathbf{x}_0\| < r$ lie on this curve: For \mathbf{x} such that $\|\mathbf{x} - \mathbf{x}_0\| < r$, \mathbf{x} solves $f(\mathbf{x}) = f(\mathbf{x}_0)$ if and only if $\mathbf{f}(\mathbf{x}) = (0, v)$ for some $v \in (-a, a)$, so that $\|(0, v)\| < a$. But this means that \mathbf{x} lies on our curve. Hence item (2) is proved. Finally, since \mathbf{f}^{-1} is one-to-one of W , which includes the points $(0, v)$ with $v \in (-a, a)$, $\mathbf{x}(v_1) = \mathbf{x}(v_2)$ for $v_1, v_2 \in (-a, a)$ if and only if $v_1 = v_2$. Moreover, by the chain rule,

$$\mathbf{x}'(0) = [D_{\mathbf{f}^{-1}}(\mathbf{0})](0, 1) = [\mathbf{u}_1, \mathbf{u}_2](0, 1) = \mathbf{u}_2 \neq \mathbf{0} ,$$

and hence item (3) is proved. \square

5.3.3 Proof of the Inverse Function Theorem.

We turn to the proof of the Inverse Function Theorem.

We first show that \mathbf{f} is one-to-one on a neighborhood of \mathbf{x}_0 , assuming for now that $[D_{\mathbf{f}}(\mathbf{x}_0)]$ has orthonormal rows, like the transformation \mathbf{f} that we constructed to prove the Implicit Function Theorem. Before stating the lemma, we recall some notation from Chapter 3: For any $r > 0$, $B_r(\mathbf{x}_0)$ denotes the open ball of radius r centered on \mathbf{x}_0 : $B_r(\mathbf{x}_0) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| < r\}$.

Lemma 15 (The one-to-one property). *Let \mathbf{f} be a continuously differentiable function on $U \subset \mathbb{R}^n$ with values in \mathbb{R}^n . Let $\mathbf{x}_0 \in U$, and suppose that $[D_{\mathbf{f}}(\mathbf{x}_0)]$ is invertible. Then there is an $r > 0$ so that*

$$\mathbf{x}, \tilde{\mathbf{x}} \in B_r(\mathbf{x}_0) \Rightarrow \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\tilde{\mathbf{x}})\| \geq \frac{1}{2\|[D_{\mathbf{f}}(\mathbf{x}_0)]^{-1}\|_{\mathbf{F}}} \|\mathbf{x} - \tilde{\mathbf{x}}\| \quad (5.20)$$

and

$$\mathbf{x} \in B_r(\mathbf{x}_0) \Rightarrow [D_{\mathbf{f}}(\mathbf{x})] \text{ is invertible .} \quad (5.21)$$

In particular, whenever \mathbf{x} and $\tilde{\mathbf{x}}$ belong to $B_r(\mathbf{x}_0)$ and $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\tilde{\mathbf{x}})$, then $\mathbf{x} = \tilde{\mathbf{x}}$.

Proof. Define $\mathbf{x}(t) := \mathbf{x} + t(\tilde{\mathbf{x}} - \mathbf{x})$, so that $\mathbf{x}(0) = \mathbf{x}$ and $\mathbf{x}(1) = \tilde{\mathbf{x}}$, and $\mathbf{x}'(t) = \tilde{\mathbf{x}} - \mathbf{x}$. By the Chain Rule and the Fundamental Theorem of Calculus,

$$\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x}) = \int_0^1 [D_{\mathbf{f}}(\mathbf{x}(t))](\tilde{\mathbf{x}} - \mathbf{x}) dt .$$

For \mathbf{x} and $\tilde{\mathbf{x}}$ close to \mathbf{x}_0 , we have $[D_{\mathbf{f}}(\mathbf{x} + t(\tilde{\mathbf{x}} - \mathbf{x})] \approx [D_{\mathbf{f}}(\mathbf{x}_0)]$ since \mathbf{f} is continuously differentiable. This approximation gives us

$$\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x}) \approx \int_0^1 [D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x}) dt = [D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x}) , \quad (5.22)$$

where we have used the fact that in this approximation, the integrand is constant. Thus, we would have $\tilde{\mathbf{x}} - \mathbf{x} \approx [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1}(\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x}))$. Applying Theorem 60, we would have

$$\|\tilde{\mathbf{x}} - \mathbf{x}\| \lesssim \| [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1} \|_{\text{F}} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\tilde{\mathbf{x}})\| .$$

Now we simply control the size of the errors made in this approximation, showing they cost us no more than a factor of 2. An exact form of (5.22) is

$$\begin{aligned} \mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x}) &= \int_0^1 ([D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x}) + ([D_{\mathbf{f}}(\mathbf{x}_0)](\mathbf{x}(t)) - [D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x}))) dt \\ &= [D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x}) + \int_0^1 ([D_{\mathbf{f}}(\mathbf{x}_0)](\mathbf{x}(t)) - [D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x})) dt . \end{aligned} \quad (5.23)$$

Hence, multiplying through by $[D_{\mathbf{f}}(\mathbf{x}_0)]^{-1}$ and using the triangle inequality and Theorem 60,

$$\begin{aligned} \|\mathbf{x} - \tilde{\mathbf{x}}\| &\leq \| [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1} \|_{\text{F}} \|\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\| \\ &\quad + \|\tilde{\mathbf{x}} - \mathbf{x}\| \int_0^1 \| [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1} \|_{\text{F}} \| [D_{\mathbf{f}}(\mathbf{x}_0)](\mathbf{x}(t)) - [D_{\mathbf{f}}(\mathbf{x}_0)](\tilde{\mathbf{x}} - \mathbf{x}) \| dt \end{aligned} \quad (5.24)$$

Since f is continuously differentiable, the real valued function $\varphi(\mathbf{x})$

$$\varphi(\mathbf{x}) := \| [D_{\mathbf{f}}(\mathbf{x}_0)](\mathbf{x}(t)) - [D_{\mathbf{f}}(\mathbf{x}_0)] \|_{\text{F}}$$

is continuous, and $\varphi(\mathbf{0}) = 0$. Hence there is an $r > 0$ such that

$$\|\mathbf{x} - \mathbf{x}_0\| < r \Rightarrow \varphi(\mathbf{x}) < \frac{1}{2\| [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1} \|_{\text{F}}} .$$

When \mathbf{x} and $\tilde{\mathbf{x}}$ are in $B_r(\mathbf{x}_0)$, then $\mathbf{x}(t) \in B_r(\mathbf{x}_0)$ for all $0 \leq t \leq 1$ (since balls are convex). Then, for $\mathbf{x}, \tilde{\mathbf{x}} \in B_r(\mathbf{x}_0)$, the integrand in (5.24) is no greater than $1/2$ for any t , and hence for $\mathbf{x}, \tilde{\mathbf{x}} \in B_r(\mathbf{x}_0)$, (5.24) gives

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \| [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1} \|_{\text{F}} \|\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\| + \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\| .$$

Cancelling $\frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\|$ from both sides, this gives us (5.20). Finally, since f is continuously differentiable, and since the set of invertible matrices is open, by further decreasing r we may arrange that (5.21) is valid. \square

Lemma 16 (The onto property). *Let f be a continuously differentiable function on $U \subset \mathbb{R}^n$ with values in \mathbb{R}^n . Let $\mathbf{x}_0 \in U$, and suppose that $[D_{\mathbf{f}}(\mathbf{x}_0)]$ is invertible. Let $r > 0$ be such that (5.20) and (5.21) are valid, and define*

$$r_0 = \frac{1}{2\| [D_{\mathbf{f}}(\mathbf{x}_0)]^{-1} \|_{\text{F}}} .$$

Then if \mathbf{y} belongs to the ball $B_{r_0/2}(\mathbf{y}_0)$, where $\mathbf{y}_0 := \mathbf{f}(\mathbf{x}_0)$, there exists $\mathbf{x}_ \in B_r(\mathbf{x}_0)$ such that $\mathbf{f}(\mathbf{x}_*) = \mathbf{y}$.*

Proof. Let $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$ and define the real valued function $g(\mathbf{x}) := \|\mathbf{f}(\mathbf{x}) - \mathbf{y}\|^2$. Let C be the closed unit ball of radius r about \mathbf{x}_0 in \mathbb{R}^n . Since C is compact, and since g is continuous, there is a minimizer \mathbf{x}_* of g on C .

Suppose we know that the minimizer \mathbf{x}_* cannot lie on the boundary of C . Then \mathbf{x}_* must be a critical point of g . Computing,

$$\mathbf{0} = \nabla g(\mathbf{x}_*) = 2[D_f(\mathbf{x}_*)](\mathbf{f}(\mathbf{x}_*) - \mathbf{y}) .$$

Since $\mathbf{x}_* \in C$ and since $[D_f(\mathbf{x})]$ is invertible for $\mathbf{x} \in C$, $[D_f(\mathbf{x}_*)]$ is invertible, and then $\mathbf{f}(\mathbf{x}_*) - \mathbf{y} = \mathbf{0}$, and hence $\mathbf{f}(\mathbf{x}_*) = \mathbf{y}$.

Hence, to prove the Lemma, we need only show that the minimizer of g on C cannot lie on the boundary of C . First, we estimate g at the center of C , \mathbf{x}_0 : When $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$,

$$g(\mathbf{x}_0) = \|\mathbf{f}(\mathbf{x}_0) - \mathbf{y}\|^2 = \|\mathbf{y}_0 - \mathbf{y}\|^2 < \frac{r_0^2}{4} .$$

Thereofre, if $g(\mathbf{x}) \geq r_0^2/4$ for all \mathbf{x} on the boundary of C , no such point can be a minimizer of g , since \mathbf{x}_0 does better. Hence the proof will be complete when we have shown that

$$\|\mathbf{x} - \mathbf{x}_0\| = r \Rightarrow g(\mathbf{x}) \geq \frac{r_0^2}{4} . \quad (5.25)$$

To see this, note that if $\|\mathbf{x} - \mathbf{x}_0\| = r$, (5.20) gives us

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{y}_0\| = \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| > \frac{r}{2\|[D_f(\mathbf{x}_0)]^{-1}\|_F} = r_0 .$$

By the triangle inequality, $\|\mathbf{f}(\mathbf{x}) - \mathbf{y}\| \geq \|\mathbf{f}(\mathbf{x}) - \mathbf{y}_0\| - \|\mathbf{y} - \mathbf{y}_0\|$. Thus for $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$, $\|\mathbf{f}(\mathbf{x}) - \mathbf{y}\| \geq r_0/2$ proving (5.25). \square

Proof of the Inverse Function Theorem, Theorem 68. Let \mathbf{x}_0 , \mathbf{y}_0 , r and r_0 be as in Lemma 16. By Lemmas 15 and 16, for each $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$, there is a unique $\mathbf{x} \in B_r(\mathbf{x}_0)$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. Hence on \mathbf{f}^{-1} is well defined on $B_{r_0/2}(\mathbf{y}_0)$. It remains to show that \mathbf{f}^{-1} is differentiable at \mathbf{y}_0 , and that the derivative there is $[D_f(\mathbf{x}_0)]^{-1}$.

To do this, fix $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$, and let $\mathbf{x} := \mathbf{f}^{-1}(\mathbf{y})$. We must show that for all $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that

$$\|\mathbf{y} - \mathbf{y}_0\| < \delta_\epsilon \Rightarrow \|\mathbf{f}^{-1}(\mathbf{y}) - \mathbf{f}^{-1}(\mathbf{y}_0) - [D_f(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{y}_0)\| < \epsilon \|\mathbf{y} - \mathbf{y}_0\| . \quad (5.26)$$

However,

$$\begin{aligned} \mathbf{f}^{-1}(\mathbf{y}) - \mathbf{f}^{-1}(\mathbf{y}_0) - [D_f(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{y}_0) &= \mathbf{x} - \mathbf{x}_0 - [D_f(\mathbf{x}_0)]^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)) \\ &= -[D_f(\mathbf{x}_0)]^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - [D_f(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)) . \end{aligned} \quad (5.27)$$

Hence by Theorem 60,

$$\begin{aligned} \|\mathbf{f}^{-1}(\mathbf{y}) - \mathbf{f}^{-1}(\mathbf{y}_0) - [D_f(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{y}_0)\| &\leq \\ &\quad \| [D_f(\mathbf{x}_0)]^{-1} \|_F \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - [D_f(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) \| . \end{aligned} \quad (5.28)$$

Therefore,

$$\frac{\|\mathbf{f}^{-1}(\mathbf{y}) - \mathbf{f}^{-1}(\mathbf{y}_0) - [D_f(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{y}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|}$$

is bounded above by

$$\|[D_f(\mathbf{x}_0)]^{-1}\|_F \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - [D_f(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} \frac{\|\mathbf{x} - \mathbf{x}_0\|}{\|\mathbf{y} - \mathbf{y}_0\|}$$

By Lemmas 15 and 16, there are $r_0, r > 0$ such that for all $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$, $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for a unique $\mathbf{x} \in B_r(\mathbf{x}_0)$, and moreover, for all $\mathbf{x} \in B_r(\mathbf{x}_0)$, $\|\mathbf{x} - \mathbf{x}_0\| \leq \frac{1}{2\|[D_f(\mathbf{x}_0)]^{-1}\|_F} \|\mathbf{y} - \mathbf{y}_0\|$ which is the same as

$$\frac{\|\mathbf{x} - \mathbf{x}_0\|}{\|\mathbf{y} - \mathbf{y}_0\|} \leq 2\|[D_f(\mathbf{x}_0)]^{-1}\|_F.$$

In particular, $\mathbf{x} \rightarrow \mathbf{x}_0$ as $\mathbf{y} \rightarrow \mathbf{y}_0$. Hence for all $\mathbf{y} \in B_{r_0/2}(\mathbf{y}_0)$,

$$\begin{aligned} \frac{\|\mathbf{f}^{-1}(\mathbf{y}) - \mathbf{f}^{-1}(\mathbf{y}_0) - [D_f(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{y}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|} &\leq \\ 2\|[D_f(\mathbf{x}_0)]^{-1}\|_F^2 \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - [D_f(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|}. \end{aligned} \quad (5.29)$$

Then since \mathbf{f} is differentiable at \mathbf{x}_0 , and since $\mathbf{x} \rightarrow \mathbf{x}_0$ as $\mathbf{y} \rightarrow \mathbf{y}_0$, the right hand side of (5.29) tends to 0 as $\mathbf{y} \rightarrow \mathbf{y}_0$, and then so does the left hand side. \square

5.4 The general Implicit Function Theorem

We have proved the Inverse Function Theorem for functions \mathbf{f} from \mathbb{R}^n to \mathbb{R}^n . So far, we have only discussed the Implicit Function Theorem for functions on \mathbb{R}^2 . As we now show, the method we have used to deduce the Implicit Function Theorem for functions on \mathbb{R}^2 from the Inverse Function Theorem readily generalizes to yield a general Implicit Function Theorem in all dimensions.

For the Implicit Function Theorem on \mathbb{R}^2 , recall that the basic idea is this: Given a continuously differentiable function $g(x, y)$ such that $\nabla g(x_0, y_0) \neq \mathbf{0}$, construct another continuously differentiable function $u(x, y)$ such that $\mathbf{h}(x, y) := (f(x, y), u(x, y))$ has an invertible Jacobian at (x_0, y_0) . Then apply the Inverse Function Theorem to \mathbf{h} .

Now consider a function \mathbf{g} defined on an open set U in \mathbb{R}^n with values in \mathbb{R}^k where $\mathbf{f}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_k(\mathbf{x}))$. If \mathbf{g} is continuously differentiable in U , then for $\mathbf{x} \in U$, the $k \times n$ matrix $[D_g(\mathbf{x})]$ is given by

$$[D_g(\mathbf{x})] = \begin{bmatrix} \nabla g_1(\mathbf{x}) \\ \vdots \\ \nabla g_k(\mathbf{x}) \end{bmatrix}.$$

In analogy with what we did in \mathbb{R}^2 , we would like to think of $\{g_1, \dots, g_k\}$ as the first k functions in a system of n coordinate functions $\{g_1, \dots, g_k, g_{k+1}, \dots, g_n\}$ on \mathbb{R}^n in some open set U containing some point \mathbf{x}_0 . Thus, we must “flesh out” the given functions $\{g_1, \dots, g_k\}$ with another $n - k$ continuously differentiable functions $\{g_{k+1}, \dots, g_n\}$ in such a way that if we define

$$\mathbf{h}(\mathbf{x}) := (g_1(\mathbf{x}), \dots, g_k(\mathbf{x}), g_{k+1}(\mathbf{x}), \dots, g_n(\mathbf{x})), \quad (5.30)$$

the Jacobian of \mathbf{h} at \mathbf{x}_0 is invertible; i.e., $[D_h(\mathbf{x}_0)]$ is invertible.

Whatever choice we make for $\{g_{k+1}, \dots, g_n\}$, when \mathbf{h} is defined by (5.30), $[D_{\mathbf{h}}(\mathbf{x}_0)]$ is invertible if and only if

$$\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0), \nabla g_{k+1}(\mathbf{x}_0), \dots, \nabla g_n(\mathbf{x}_0)\}$$

is linearly independent.

When $\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0)\}$ is linearly independent, we can always keep adding vectors to this set until we achieve a maximal linearly independent set of n vectors $\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0), \mathbf{u}_{k+1}, \dots, \mathbf{u}_n\}$. We even have several constructive methods for doing this.

However, we do it, if we define $g_j(\mathbf{x})$ by

$$f_g(\mathbf{x}) := \mathbf{u}_j \cdot (\mathbf{x} - \mathbf{x}_0) \quad \text{for } j = k + 1, \dots, n , \quad (5.31)$$

and then use these functions to define $\mathbf{h}(\mathbf{x})$, we shall have

$$[D_{\mathbf{h}}(\mathbf{x}_0)] = \begin{bmatrix} \nabla g_1(\mathbf{x}_0) \\ \vdots \\ \nabla g_k(\mathbf{x}_0) \\ \mathbf{u}_{k+1} \\ \vdots \\ \mathbf{u}_n \end{bmatrix} .$$

By construction, the rows are linearly independent, so $[D_{\mathbf{h}}(\mathbf{x}_0)]$ is invertible. This enables us to adapt the strategy we have used to prove the Implicit Function Theorem in \mathbb{R}^2 to prove a much more general result. As noted above, the key condition, besides the condition that \mathbf{g} is continuously differentiable near \mathbf{x}_0 , will be that $\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0)\}$ is linearly independent, which is the same as $[D_{\mathbf{f}}(\mathbf{x}_0)]$ having rank k .

Theorem 69 (Implicit Function Theorem in \mathbb{R}^n). *Let \mathbf{g} be an \mathbb{R}^k valued function defined on an open set $U \subset \mathbb{R}^n$, and suppose that \mathbf{g} is continuously differentiable at each point $\mathbf{x}_0 \in U$. Then, in case $[D_{\mathbf{g}}(\mathbf{x}_0)]$ has rank k , there is an $r_0 > 0$ and a continuously differentiable function $\mathbf{y}(\mathbf{x})$ defined on the open ball of radius r_0 in \mathbb{R}^{n-k} so that*

(1) *For all (u_{k+1}, \dots, u_n) with $\sum_{j=k+1}^n u_j^2 < r_0^2$, $\mathbf{g}(\mathbf{y}(u_{k+1}, \dots, u_n)) = \mathbf{g}(\mathbf{x}_0)$. That is, each $\mathbf{y}(u_{k+1}, \dots, u_n)$ solves the equation $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0)$.*

(2) *There is an $r > 0$ such that every \mathbf{x} satisfying $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0)$ and $\|\mathbf{x} - \mathbf{x}_0\| < r$ is of the form*

$$\mathbf{x} = \mathbf{y}(u_{k+1}, \dots, u_n)$$

for exactly one (u_{k+1}, \dots, u_n) such that $\sum_{j=k+1}^n u_j^2 < r^2$.

(3) $\text{rank}([D_{\mathbf{y}}(\mathbf{0})]) = n - k$.

The statement of the theorem may seem somewhat involved, although it does at least parallel the version we have studied earlier for $n = 2$. All the same, it is worthwhile to go over what the theorem says before going into the proof or applications.

Let $\tilde{\mathbf{g}}$ be the linear approximation to \mathbf{g} at \mathbf{x}_0 . That is,

$$\tilde{\mathbf{g}}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0) + [D_{\mathbf{g}}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0).$$

The equation $\tilde{\mathbf{g}}(\mathbf{x}) = \tilde{\mathbf{g}}(\mathbf{x}_0)$ is readily solved by the methods of Linear Algebra: Since $\tilde{\mathbf{g}}(\mathbf{x}_0) = \mathbf{g}(\mathbf{x}_0)$, this equation is equivalent to

$$[D_{\mathbf{g}}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) = \mathbf{0}. \quad (5.32)$$

By definition, $\mathbf{x} - \mathbf{x}_0$ solves this linear equation if and only if $\mathbf{x} - \mathbf{x}_0 \in \text{Null}([D_{\mathbf{g}}(\mathbf{x}_0)])$. Since the row space of $[D_{\mathbf{g}}(\mathbf{x}_0)]$ is k -dimensional, and since $\text{Null}([D_{\mathbf{g}}(\mathbf{x}_0)])$ is the orthogonal complement of the row space of $[D_{\mathbf{g}}(\mathbf{x}_0)]$, $\text{Null}([D_{\mathbf{g}}(\mathbf{x}_0)])$ is an $n - k$ dimensional subspace of \mathbb{R}^n . Let $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ be an orthonormal basis for $\text{Null}([D_{\mathbf{g}}(\mathbf{x}_0)])$. Then the general solution of (5.32) is of the form

$$\mathbf{x}_0 + \sum_{j=k+1}^n u_j \mathbf{u}_j$$

where $(u_{k+1}, \dots, u_n) \in \mathbb{R}^{n-k}$. Therefore, the function

$$\mathbf{y}(u_{k+1}, \dots, u_n) := \mathbf{x}_0 + \sum_{j=k+1}^n u_j \mathbf{u}_j$$

provides a parameterization of the solution set of the equation $\tilde{\mathbf{g}}(\mathbf{x}_0) = \mathbf{g}(\mathbf{x}_0)$. This function $\mathbf{y}(u_{k+1}, \dots, u_n)$ is the function provided by the Implicit Function Theorem in the case that $\mathbf{g}(\mathbf{x})$ is linear. The Implicit Function Theorem says that even if \mathbf{g} is not linear, but is continuously differentiable, there is a similar (non linear) parameterization of the solution set of the non linear equation $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0)$.

Proof of Theorem 69. Suppose that $[D_{\mathbf{g}}(\mathbf{x}_0)]$ has rank k . Then we can select vectors $\{\mathbf{u}_{k+1}, \dots, \mathbf{u}_n\}$ so that

$$\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0), \mathbf{u}_{k+1}, \dots, \mathbf{u}_n\}$$

is linearly independent. Define $g_j(\mathbf{x}) := \mathbf{u}_j \cdot (\mathbf{x} - \mathbf{x}_0)$, $j = k+1, \dots, n$, as in by (5.31), and define $\mathbf{h}(\mathbf{x})$ by

$$\mathbf{h}(\mathbf{x}) := (g_1(\mathbf{x}), \dots, g_k(\mathbf{x}), g_{k+1}(\mathbf{x}), \dots, g_n(\mathbf{x})). \quad (5.33)$$

Then

$$[D_{\mathbf{h}}(\mathbf{x}_0)] = [\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0), \mathbf{u}_{k+1}, \dots, \mathbf{u}_n]^T.$$

Since the rows of $[D_{\mathbf{h}}(\mathbf{x}_0)]$ are linearly independent, by the Fundamental Theorem of Linear Algebra, $[D_{\mathbf{h}}(\mathbf{x}_0)]$ is invertible.

Then by the Inverse Function Theorem, there is an open set $V \subset \mathbb{R}^n$ containing \mathbf{x}_0 and an open set $W \subset \mathbb{R}^n$ containing $\mathbf{h}(\mathbf{x}_0)$ such that \mathbf{h} is a one-to-one transformation from V onto W with a continuously differentiable inverse. By the construction of $\{g_{k+1}, \dots, g_n\}$, $g_j(\mathbf{x}_0) = 0$ for $j = k+1, \dots, n$. Thus,

$$\mathbf{h}(\mathbf{x}_0) = (g_1(\mathbf{x}_0), \dots, g_k(\mathbf{x}_0), 0, \dots, 0) \in W.$$

Since W is open, there is an $r_0 > 0$ so that

$$u_{k+1}^2 + \dots + u_n^2 < r_0 \Rightarrow (g_1(\mathbf{x}_0), \dots, g_k(\mathbf{x}_0), u_{k+1}, \dots, u_n) \in W.$$

Let $B_{r_0}(\mathbf{0})$ denote the set of vectors $(u_{k+1}, \dots, u_n) \in \mathbb{R}^{n-K}$ such that $u_{k+1}^2 + \dots + u_n^2 < r_0$. Define the function \mathbf{y} from $B_{r_0}(\mathbf{0})$ to \mathbb{R}^n by

$$\mathbf{y}(u_{k+1}, \dots, u_n) := \mathbf{h}^{-1}(g_1(\mathbf{x}_0), \dots, g_k(\mathbf{x}_0), u_{k+1}, \dots, u_n). \quad (5.34)$$

Then, by definition, $\mathbf{h}(\mathbf{y}(u_{k+1}, \dots, u_n)) = (g_1(\mathbf{x}_0), \dots, g_k(\mathbf{x}_0), u_{k+1}, \dots, u_n)$, and since the first k entries of \mathbf{h} are g_1, \dots, g_k , this means that $\mathbf{g}(\mathbf{y}(u_{k+1}, \dots, u_n)) = \mathbf{g}(\mathbf{x}_0)$, which proves (1).

Next, since \mathbf{h} is differentiable, it is continuous, and hence there is an $r > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| < r \Rightarrow \|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}_0)\| < r_0.$$

Since for $j > k$, $h_j(\mathbf{x}_0) = 0$, $\|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}_0)\| \geq \sum_{j=k+1}^n h_j^2(\mathbf{x})$. Therefore

$$\|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}_0)\| < r_0 \Rightarrow \sqrt{\sum_{j=k+1}^n h_j^2(\mathbf{x})} < r.$$

By the definition of \mathbf{h} , (5.33), this means that $(h_{k+1}(\mathbf{x}), \dots, h_n(\mathbf{x})) \in B_{r_0}(\mathbf{0})$.

Now suppose that $\|\mathbf{x} - \mathbf{x}_0\| < r$ and $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0)$. Then by the definition of \mathbf{h} , (5.33), and what we have just proved $\mathbf{h}(\mathbf{x}) = (g_1(\mathbf{x}_0), \dots, g_k(\mathbf{x}_0), u_{k+1}, \dots, u_n)$ for some $(u_{k+1}, \dots, u_n) \in B_{r_0}(\mathbf{0})$. By (5.34), this is the same as $\mathbf{h}(\mathbf{x}) = \mathbf{h}(\mathbf{y}(u_{k+1}, \dots, u_n))$, and therefore, since \mathbf{h} is one-to-one on $B_r(\mathbf{x}_0)$, $\mathbf{x} = \mathbf{y}(u_{k+1}, \dots, u_n)$ for some unique $(u_{k+1}, \dots, u_n) \in B_{r_0}(\mathbf{0})$. This proves (2).

Finally, recall that \mathbf{h}^{-1} is continuously differentiable with invertible derivative $[D_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{x}_0))]$ at $\mathbf{h}(\mathbf{x}_0)$. Since by definition, \mathbf{y} is obtained from \mathbf{h}^{-1} by “freezing” the values of the first k variables in \mathbf{h}^{-1} , its derivative in the remaining $n - k$ variables, u_{k+1}, \dots, u_n , is given by the $(n - k) \times n$ matrix $[D_{\mathbf{y}}(\mathbf{g}(\mathbf{x}_0))]$ consisting of the bottom $n - k$ rows of $[D_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{x}_0))]$. Since the rows of $[D_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{x}_0))]$ are linearly independent, as a consequence of the invertibility of $[D_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{x}_0))]$, the $n - k$ rows of $[D_{\mathbf{y}}(\mathbf{g}(\mathbf{x}_0))]$ are linearly independent. Thus, $[D_{\mathbf{y}}(\mathbf{g}(\mathbf{x}_0))]$ has rank $n - k$. This proves (3). \square

5.5 Lagrange’s Theorem in general

We are now ready to state and prove the general form of Lagrange’s Theorem.

Theorem 70 (Lagrange’s Theorem). *Let $U \subset \mathbb{R}^n$ be open, and let \mathbf{g} be a continuously differentiable function defined on U with values in \mathbb{R}^k ;*

$$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_k(\mathbf{x})).$$

Suppose that $\text{rank}([D_{\mathbf{g}}(\mathbf{x})]) = k$ for all $\mathbf{x} \in U$. Define $D := \{\mathbf{x} \in U : \mathbf{g}(\mathbf{x}) = \mathbf{c}\}$, and suppose that \mathbf{c} is chosen so that D is not empty. Let f be a continuously differentiable function on U with values in \mathbb{R} . Then a necessary condition for $\mathbf{x}_0 \in D$ to be a maximizer or a minimizer of f in D is that

$$\nabla f(\mathbf{x}_0) \in \text{Span}(\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0)\}).$$

Proof. Let $\mathbf{x}_0 \in D$ and suppose that \mathbf{x}_0 either maximizes or minimizes f in D . By the Implicit Function Theorem, there is a an $r_0 > 0$ and a continuously differentialbe function $\mathbf{y}(u_{k+1}, \dots, u_n)$ defined on

the centered open ball $B_{r_0}(\mathbf{0})$ in \mathbb{R}^{n-k} such that for all $(u_{k+1}, \dots, u_n) \in B_{r_0}(\mathbf{0})$, $\mathbf{g}(\mathbf{y}(u_{k+1}, \dots, u_n)) = \mathbf{g}(\mathbf{x}_0)$, so that $\mathbf{y}(u_{k+1}, \dots, u_n) \in D$.

For each $j = k+1, \dots, n$, consider the curve $\mathbf{u}(t) := \mathbf{y}(t\mathbf{e}_j)$, $-r_0 < t < r_0$. It is continuously differentiable since \mathbf{y} is continuously differentiable, and $\mathbf{u}(0) = \mathbf{x}_0$. By the Chain Rule

$$\frac{d}{dt} f(\mathbf{u}(t)) \Big|_{t=0} = \nabla f(\mathbf{x}_0) \cdot \mathbf{u}'(0) = \nabla f(\mathbf{x}_0) \cdot \frac{\partial}{\partial u_j} \mathbf{y}(\mathbf{0}) = 0 ,$$

where we have used the hypothesis that \mathbf{x}_0 either maximizes or minimizes f in D .

Notice that $\frac{\partial}{\partial u_j} \mathbf{y}(\mathbf{0})$ is column j of $[D_{\mathbf{y}}(\mathbf{g}(\mathbf{0})]$ which is an $n \times (n-k)$ matrix of rank $n-k$ by part (3) of the Implicit Function Theorem. Thus, its $n-k$ columns are linearly independent. Since $j \in \{k+1, \dots, n\}$ is arbitrary, we conclude that

$$\nabla f(\mathbf{x}_0) \cdot \frac{\partial}{\partial u_j} \mathbf{y}(\mathbf{0}) = 0 , \quad j = k+1, \dots, n . \quad (5.35)$$

In the same way, since \mathbf{g} is constant on D by definition, for each $\ell = 1, \dots, k$,

$$\nabla g_\ell(\mathbf{x}_0) \cdot \frac{\partial}{\partial u_j} \mathbf{y}(\mathbf{0}) = 0 , \quad j = k+1, \dots, n . \quad (5.36)$$

It follows from (5.36) that $\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0)\}$ is a subset of

$$V := \left\{ \frac{\partial}{\partial u_{k+1}} \mathbf{y}(\mathbf{0}), \dots, \frac{\partial}{\partial u_n} \mathbf{y}(\mathbf{0}) \right\}^\perp$$

and it is linearly independent since $[D_{\mathbf{g}}(\mathbf{x}_0)]$ has rank k . Hence $\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0)\}$ spans a k dimensional subspace of V , which has dimension k , since it is the orthogonal complement of a subspace of dimension $n-k$. Therefore, $\text{span}(\{\nabla g_1(\mathbf{x}_0), \dots, \nabla g_k(\mathbf{x}_0)\}) = V$, and (5.35) says that $\nabla f(\mathbf{x}_0) \in V$. \square

We now consider some examples of constrained optimization problems in \mathbb{R}^n or more variables with a constraint equation $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, where \mathbf{g} takes values in \mathbb{R}^k , some $1 \leq k \leq n-1$.

We begin with the case $n = 3$ and $k = 1$. Let g be a continuously differentiable function on \mathbb{R}^3 with values in \mathbb{R} . For example,

$$g(x, y, z) = x^2 + y^2 + z^2 .$$

In this case, the level set of f consisting of all the solutions of the equation $g(x, y, z) = 1$ is the unit sphere in \mathbb{R}^3 . The equation $g(x, y, z) = 1$ is the implicit definition of this surface.

The unit sphere is a particularly nice surface, so we can also explicitly parameterize it. Here is one way: Define

$$\mathbf{x}(s, t) = (\cos s \sin t, \sin s \sin t, \cos t) ,$$

where $0 \leq s < 2\pi$ and $0 \leq t < \pi$.

Now let us consider a constraint equation $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ on \mathbb{R}^3 where \mathbf{g} takes values in \mathbb{R}^2 . This may be considered as a system of two scalar valued constraint equations.

Let \mathbf{g} be a continuously differentiable function on \mathbb{R}^3 with values in \mathbb{R}^2 . For example,

$$\mathbf{g}(x, y, z) = \begin{bmatrix} x^2 + y^2 + z^2 - 1 \\ z \end{bmatrix} .$$

In this case, set of solutions of the equation $\mathbf{g}(x, y, z) = (0, 0)$ is the intersection of the unit sphere in \mathbb{R}^3 with the plane $z = 0$. This is nothing other than the unit circle in x, y plane, which can be explicitly parameterized by

$$\mathbf{x}(t) = (\sin(t), \cos(t), 0).$$

The Implicit Function Theorem says that whenever g is differentiable from \mathbb{R}^3 to \mathbb{R} , and $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$, the equation $f(\mathbf{x}) = f(\mathbf{x}_0)$ defines a differentiable parameterized surface in \mathbb{R}^3 passing through \mathbf{x}_0 , and including all of the solutions of $f(\mathbf{x}) = f(\mathbf{x}_0)$ in some neighborhood of \mathbf{x}_0 .

It also says that given a system of two such equations $g_1(\mathbf{x}) = g_1(\mathbf{x}_0)$ and $g_2(\mathbf{x}) = g_2(\mathbf{x}_0)$, with $\{\nabla g_1(\mathbf{x}_0), \nabla g_2(\mathbf{x}_0)\}$ linearly independent, and therefore non-zero, each of these individual equations will determine a surface passing through \mathbf{x}_0 , and the intersection of the two surfaces will be a continuously differentiable curve passing through \mathbf{x}_0 . The points on this curve will be the solutions to the system $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}_0)$ near \mathbf{x}_0 , where $\mathbf{g} = (g_1, g_2)$.

That is, just as with lines and planes in \mathbb{R}^3 , where $\mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ specifies a plane, but it takes a system of two such equations to specify a line, one equation $g(\mathbf{x}) = g(\mathbf{x}_0)$ specifies a surface in a neighborhood of \mathbf{x}_0 , provided $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$, however it takes two such equations $g_1(\mathbf{x}) = g_1(\mathbf{x}_0)$ and $g_2(\mathbf{x}) = g_2(\mathbf{x}_0)$ to specify a curve passing through \mathbf{x}_0 .

If the surfaces in question are more complicated than planes or spheres and such, it might not be possible to find explicit parameterizations, but the Implicit Function Theorem at least assures us that such parameterizations exist. We used this to prove the general form of Lagrange's Theorem. The really good news is the to *use* Lagrange's Theorem, you do not need to find the parameterizations. We now turn to some examples.

We begin with one constraint in \mathbb{R}^3 , that is, optimization on a surface. Let f and g be a continuously differentiable functions defined on \mathbb{R}^3 , and define the sets D , U and B by

$$\begin{aligned} D &:= \{(x, y, z) : g(x, y, z) \leq 0\} \\ U &:= \{(x, y, z) : g(x, y, z) < 0\} \\ B &:= \{(x, y, z) : g(x, y, z) = 0\}. \end{aligned} \tag{5.37}$$

Then U is the interior of D and B is its boundary.

Let us try to determine the minimizers and maximizers of f on D , if any. Just as before, the only points in U that can possibly be minimizers or maximizers are the critical points of f in U .

Next, we consider the boundary. By Lagrange's Theorem, at each point \mathbf{x}_0 of B , if $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$, we must have

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0) \quad \text{and} \quad g(\mathbf{x}_0) = 0. \tag{5.38}$$

This is a system of 4 equations in the 4 variables x, y, z and λ . Since we are in \mathbb{R}^3 , we can immediately eliminate λ by taking the cross product: $\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0)$ for some λ if and only if $\nabla f(\mathbf{x}_0) \times \nabla g(\mathbf{x}_0) = \mathbf{0}$. Thus we have the equivalent formulation of (5.38):

$$\nabla f(\mathbf{x}_0) \times \nabla g(\mathbf{x}_0) = \mathbf{0} \quad \text{and} \quad g(\mathbf{x}_0) = 0. \tag{5.39}$$

Example 87 (One constraint in three variables). Let $f(x, y, z) = xyz$, and let $g(x, y, z) = x^2 + y^2 + z^2 - 1$. Let us find the minimizers and maximizers of f in the set D given by $g(\mathbf{x}) \leq 0$, which is the closed unit ball.

We first compute $\nabla f(x, y, z) = (yz, xz, xy)$. Clearly $\mathbf{0}$ is one critical point. Next, suppose (x, y, z) is any critical point with $x \neq 0$. Then dividing $(yz, xz, xy) = (0, 0, 0)$ through by x , we find

$$(yz/x, z/y) = (0, 0, 0).$$

Thus, $y = 0$ and $z = 0$. Hence all of the points

$$(x, 0, 0) \quad \text{with} \quad -1 \leq x \leq 1$$

are critical points of f in D . By symmetry in x , y and z , so are the points $(0, y, 0)$ with $-1 \leq y \leq 1$ and $(0, 0, z)$ with $-1 \leq z \leq 1$. Thus we have infinite many critical points. However, f takes on the same value, namely 0, at each of them, so this is not so bad.

Next, to look for possible minimizers and maximizers on the boundary B , we compute

$$\nabla g(x, y, z) = 2(x, y, z),$$

and then, to use (5.39),

$$\nabla f(\mathbf{x}) \times \nabla g(\mathbf{x}) = 2((y^2 - z^2)x, (z^2 - x^2)y, (x^2 - y^2)z).$$

Thus if $\mathbf{x} = (x, y, z)$ is on the boundary and is a minimizer or a maximizer, then

$$\begin{aligned} (y^2 - z^2)x &= 0 \\ (z^2 - x^2)y &= 0 \\ (x^2 - y^2)z &= 0 \\ x^2 + y^2 + z^2 &= 1. \end{aligned}$$

From the first equation we see that either $x = 0$ or $y^2 = z^2$. Suppose $x = 0$. Then from the second and third equations, we have that either $y = 0$ or $z = 0$. Suppose $y = 0$. Then from the last equation, we have $z^2 = 1$ so $z = \pm 1$. Thus we find the solution $(0, 0, \pm 1)$. If instead we took $z = 0$, we would find the solutions $(0, \pm 1, 0)$.

On the other hand, if $y^2 = z^2$, either $y = z = 0$, in which case the fourth equation tells us $x = \pm 1$, giving us the solutions $(\pm 1, 0, 0)$, or else $y^2 = z^2 > 0$, and then the second equation tells us $x^2 = z^2$ as well. Thus, using the fourth equation,

$$x^2 = y^2 = z^2 = \frac{1}{3}.$$

This gives us 8 more solutions, namely

$$(\pm 3^{-1/2}, \pm 3^{-1/2}, \pm 3^{-1/2}).$$

By symmetry, we have found all of the solutions, so now let us evaluate f at each of them. Only the last 8 give non-zero values.

$$f(\pm 3^{-1/2}, \pm 3^{-1/2}, \pm 3^{-1/2}) = \pm 3^{-3/2}$$

with the plus sign in the right if there are an even number of minus signs on the left, and the minus sign on the right otherwise. This gives all of the minimizers and maximizers.

Next, let us consider the problem of finding minmizers and maximizers along an implicitly defined curve in \mathbb{R}^3 given by the system of equations

$$\mathbf{g}(\mathbf{x}) = \mathbf{0} \quad \text{where} \quad \mathbf{g}(\mathbf{x}) = \begin{bmatrix} \nabla g_1(\mathbf{x}) \\ \nabla g_2(\mathbf{x}) \end{bmatrix} .$$

Let us assume that $\nabla g_1(\mathbf{x}_0) \times \nabla g_2(\mathbf{x}_0) \neq \mathbf{0}$ so that the Implicit Function Theorem does ensure that the solution of $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ near \mathbf{x}_0 is given by a differentiable curve $\mathbf{x}(t)$ passing through \mathbf{x}_0 at $t = 0$.

By Lagrange's Theorem, if \mathbf{x}_0 is a maximizer or a minimizer on the curve,

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g_1(\mathbf{x}_0) + \mu \nabla g_2(\mathbf{x}_0) , \quad (5.40)$$

which is the traditional Lagrange multiplier formulation of the necessary condition for \mathbf{x}_0 to be a maximizer or a minimizer of f on the level set given by $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.

Since we are working in \mathbb{R}^3 , we can use the cross product to eliminate the two Lagrange multipliers as follows: The condition (5.40) is equivalent to

$$\nabla f(\mathbf{x}_0) \cdot \nabla g_1(\mathbf{x}_0) \times \nabla g_2(\mathbf{x}_0) = 0 . \quad (5.41)$$

Example 88 (Two constraints in three variables). Let $f(x, y, z) := xyz$, and let $g_1(x, y, z) := x^2 + y^2 + z^2 - 1$ and $g_2(x, y, z) = 2x + 2y - z$. Let us find the minimizers and maximizers of f on the curve in \mathbb{R}^3 given implicitly by $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.

Writing out $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ explicitly as a system of equations, it becomes

$$\begin{aligned} x^2 + y^2 + z^2 &= 1 \\ 2x + 2y - z &= 0 \end{aligned}$$

Thus, the solution set is the intersection of the unit sphere and a plane through the origin, and it is therefore a great circle (geodesic) on the unit sphere. We could of course easily parameterize it. But let us instead solve the problem by means of Theorem ??.

We compute

$$\nabla g_1(\mathbf{x}) \times \nabla g_2(\mathbf{x}) = 2(x, y, z) \times (2, 2, -1) = 2((-y - 2z, 2z + x, 2x - 2y)) .$$

Next, we compute $\nabla f(\mathbf{x}) = (yz, xz, xy)$, and so

$$\nabla f(\mathbf{x}) \cdot \nabla g_1(\mathbf{x}) \times \nabla g_2(\mathbf{x}) = 2(-yz(y + 2z) + xz(2z + x) + xy(2x - 2y)) .$$

Thus, after some grouping, we have the system of equations

$$\begin{aligned} z(x^2 - y^2) + 2z^2(x - y) + 2xy(x - y) &= 0 \\ x^2 + y^2 + z^2 &= 1 \\ 2x + 2y - z &= 0 \end{aligned}$$

Substituting $z = 2(x+y)$ from the third equation into the first equation, the first equation becomes

$$10(x+y)^2(x-y) = 2xy(x-y) .$$

Thus, either $x = y$, or

$$10(x+y)^2 - 2xy = 0 .$$

The latter is a quadratic equation, and its only solution is $x = y = 0$, which is a special case of $x = y$. Thus, we must have $x = y$.

Now the last two equations in our system simplify to $z = 4x$ and $2x^2 + z^2 = 1$. Eliminating z , $18x^2 = 1$ so we have $y = x = \pm \frac{2}{3\sqrt{2}}$ and then from $z = 4x$, we find our candidates

$$\pm \frac{2}{3\sqrt{2}}(1, 1, 4) .$$

We then evaluate to find $f\left(\pm \frac{2}{3\sqrt{2}}(1, 1, 4)\right) = \pm \frac{1}{2}3^{-3/2}$, and hence $\frac{2}{3\sqrt{2}}(1, 1, 4)$ is the maximizer and $-\frac{2}{3\sqrt{2}}(1, 1, 4)$ is the minimizer.

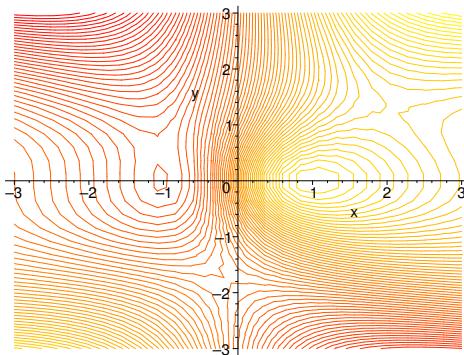
Comparing with Example 87, we see that imposing the additional constraint raised the minimum value and lowered the maximum value, as one would expect.

5.6 Exercises

5.1: Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^3y + 2y - 3y^2x$.

(a) Compute the gradient of f , and find all points (x, y) at which the tangent plane to the graph of f is horizontal.

(b) Could the following be a contour plot of f ? Explain your answer.

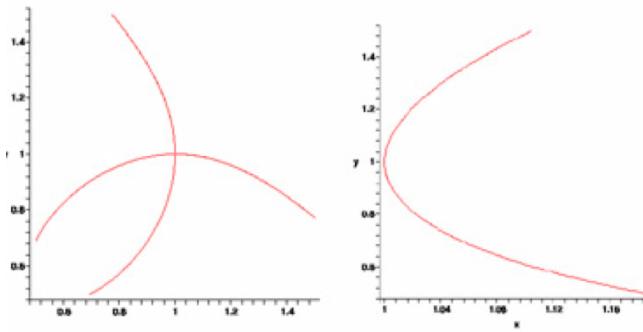


5.2: Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^2y + yx - xy^2$.

(a) Compute the gradient of f , and find all critical points of f .

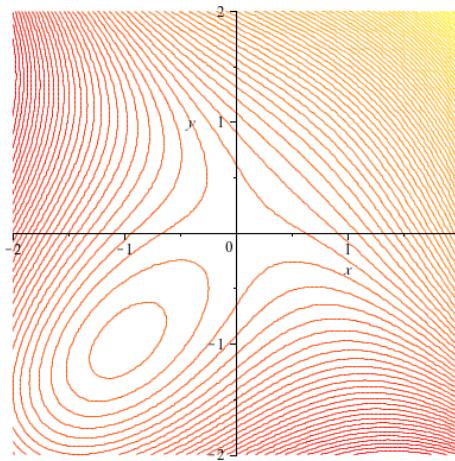
(b) Find a parameterization of the tangent line to the level curve of f through the point $(1, 1)$.

(c) Could either of the following be a contour plot of f ? If so, which could, and which could not? Explain your answer.



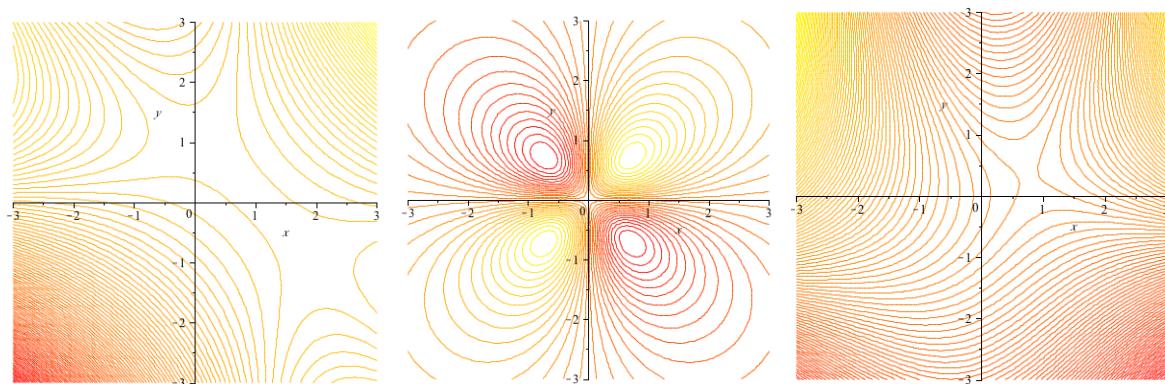
5.3: Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^3 + y^3 + 3xy$.

- (a) Compute the gradient of f , and find all points (x, y) at which the tangent plane to the graph of f is horizontal.
- (b) Find the equation of the tangent line to the level curve of f passing through the point $(1, 1)$.
- (c) Could the following be a contour plot of f ? Explain your answer.



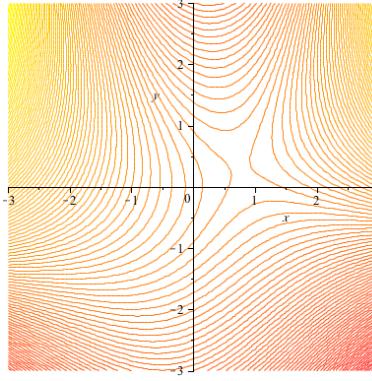
5.4: Let $f(x, y) = \frac{xy}{(1+x^2+y^2)^2}$.

- (a) Find all of the critical points of f , and find the value of f at each of the critical points.
- (b) One of the following is a contour plot for f . Which one is it? Explain your answer to receive credit.



5.5: Let $f(x, y) = xy^2 - xy$.

- (a) Find all of points at which the tangent plane to the graph of f is horizontal.
- (b) Find the equation of the tangent line to the contour curve of f through the point $(3/2, 1/3)$.
- (c) Could the following be a contour plot for f ? Explain your answer to receive credit.



5.6: Let $f(x, y) = (x + y)^4 + (x - y)^2$. Find the minimum and maximum values of f on the unit circle $x^2 + y^2 = 1$, and all of the places on the circle at which f takes on these values.

5.7: Let $f(x, y) = xy$. Let D denote the region in the plane consisting of all of the points (x, y) such that $x^2 + 4y^2 \leq 6$. Find the minimum and maximum values of f in this region. Also, find all of the minimizers and maximizers in this region.

5.8: Let $f(x, y) = \frac{xy}{(1 + x^2 + y^2)^2}$. Find the minimum and maximum values of f in the set where

$$|x| + |y| \leq 1 .$$

Also, find all of the minimizers and maximizers.

5.9: Let $f(x, y) = xy$. Find the minimum and maximum values of f the set $D \subset \mathbb{R}^2$ given by

$$(x^2 + y^2)^2 \leq 2x^2 - 2y^2 ,$$

and all of the points in D at which f takes on these values. Note that D is the region inside the Bernoulli lemniscate.

5.10 Let \mathcal{S} be closed upper hemisphere of the unit sphere in \mathbb{R}^3 . Let $f(x, y, z) = xyz$. Find the minimum and maximum values of f on \mathcal{S} , and all of the points at which f takes on these values. Explain how you are taking into account both of the constraints $x^2 + y^2 + z^2 = 1$ and $z \geq 0$.

5.11: Let $f(x, y) = x^2 + y^2 - xy$. Find the minimum and maximum values of f the closed unit disk $D \subset \mathbb{R}^2$. That is, D is given by

$$x^2 + y^2 \leq 1 .$$

5.12: Let $f(x, y) = x^2 + 2y$. Find the minimum and maximum values of f in the set $D \subset \mathbb{R}^2$ that lies below the parabola $y = 2x - x^2$ and inside the circle $(x - 1)^2 + y^2 = 1$.

5.13: Let $f(x, y) = xy + 2x - 2y$. Find the minimum and maximum values of f on the region $D \subset \mathbb{R}^2$ that lies below the parabola $y = 1 - x^2$ and above the x -axis, $y = 0$.

5.14: Let D be the region consisting of all points (x, y) satisfying

$$x^2 \leq y \leq 3 + 2x .$$

Let $f(x, y) = x^2y - 3x$. We seek to find the minimum and maximum values of f on D , and find all minimizers and maximizers.

5.15: Let $f(x, y) = x^2 + y^2 - 2xy$. Find the minimum and maximum values of f the region D defined by

$$\frac{x^2}{2} \leq y \leq 2 .$$

Also, find all points in D at which f takes on these values; i.e., find all of the minimizers and maximizers.

5.16: Find the points on the ellipsoid given by $x^2 + y^2 + xy = 1$ that minimize and maximize the distance to the line $y = 3 - 2x$. (Draw a picture, and think about the geometric idea behind the idea of Lagrange multipliers before plunging into computation.)

5.17: Find the maximum and minimum values of $f(x, y, z) = 3x + y - z$ on the set C of points in \mathbb{R}^3 satisfying

$$\begin{aligned} x + y + z &= 3 \\ x^2 + y^2 &= 1 . \end{aligned}$$

Notice that C is the intersection of a plane and a cylinder about the z -axis.

5.18: Find the points on the sphere $x^2 + y^2 + z^2 = 1$ that are closest and farthest from the point $(1, 2, 3)$.

5.19: Find the point on the paraboloid $z = x^2 + y^2$ that is closest and farthest from the point $(1, 3, 4)$.

Chapter 6

CURVATURE AND QUADRATIC APPROXIMATION

6.1 Quadratic functions

6.1.1 The matrix form of a purely quadratic function

So far we have developed a number of computational methods that rely on finding the best linear approximation to a function of several variables. To go further, we need to qualitatively analyze how much, for example, the graph of $z = f(x, y)$ “curves away” from its tangent plane at a point (x_0, y_0) . The fundamental subject of this chapter is the curvature of surfaces, such as the graph of $z = f(x, y)$, and related concepts for functions of more variables. The fundamental method is “best quadratic approximation”

A purely quadratic function f on \mathbb{R}^2 is a function of the form $f(x, y) = \alpha x^2 + 2\beta xy + \gamma y^2$ for some numbers α, β and γ . If we introduce the matrix $A = \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}$, we can write f in its matrix form

$$f(x, y) = (x, y) \cdot A(x, y) .$$

More generally, a *purely quadratic function* $f(\mathbf{x})$ on \mathbb{R}^n has the form

$$f(x_1, \dots, x_n) = \sum_{i,j=1}^n \alpha_{i,j} x_i x_j ,$$

for some numbers $\alpha_{i,j}$, $1 \leq i, j \leq n$. Fix any k, ℓ with $1 \leq k < \ell \leq n$. Since $x_\ell x_k = x_k x_\ell$, the coefficient of $x_k x_\ell$ in f is $\alpha_{k,\ell} + \alpha_{\ell,k}$. Therefore, if for each $1 \leq i, j \leq n$ we replace $\alpha_{i,j}$ by $(\alpha_{i,j} + \alpha_{j,i})/2$, we do not change the value of $f(\mathbf{x})$ at any \mathbf{x} . We shall always assume that the coefficients $\alpha_{i,j}$ that specify f satisfy

$$\alpha_{i,j} = \alpha_{j,i}$$

for all $1 \leq i, j \leq n$. Let A be the $n \times n$ matrix whose i, j entry is $\alpha_{i,j}$. Then we may write f in its matrix form $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$.

Notice that the matrix A in this expression is *symmetric*. That is, $A_{i,j} = A_{j,i}$ for all $1 \leq i, j \leq n$, or equivalently, $A = A^T$.

Definition 66. An $n \times n$ matrix A is symmetric in case $A = A^T$.

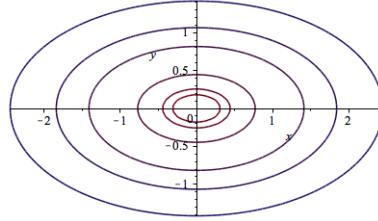
The correspondence between purely quadratic functions and symmetric matrices is *one-to-one*. Given a symmetric matrix, define $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$. Conversely, let the coefficients $\alpha_{i,j}$ that specify f satisfy the symmetry requirement, $\alpha_{i,j} = \alpha_{j,i}$ and then define A by $A_{i,j} = \alpha_{i,j}$.

6.1.2 Purely quadratic functions as sums of squares

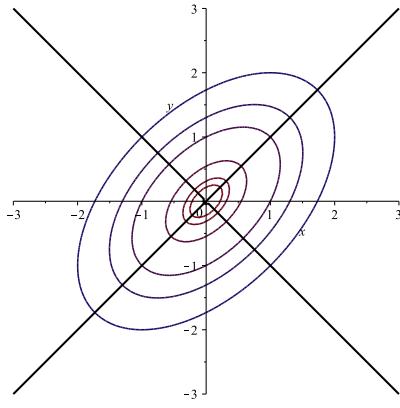
Consider the purely quadratic function $f(x, y) = x^2 - xy + y^2$. Define new variables u and v by $u = (x+y)/\sqrt{2}$ and $v = (x-y)/\sqrt{2}$. Then $x = (u+v)/\sqrt{2}$ and $y = (u-v)/\sqrt{2}$. Substituting these expressions into f , we find

$$f(x, y) = f\left(\frac{u+v}{\sqrt{2}}, \frac{u-v}{\sqrt{2}}\right) = \frac{(u+v)^2 - (u^2 - v^2) + (u-v)^2}{2} = \frac{u^2 + 3v^2}{2}.$$

In terms of the u, v variables, we see that f is a sum of positive multiples of u^2 and v^2 , and therefore f is never negative, even though xy can have either sign. The expression in terms of u and v is much more revealing than the expression in terms of x and y . In addition to the non-negativity, notice that for any constant $c > 0$, the curve in the u, v plane given by $\frac{u^2 + 3v^2}{2} = c$ is an ellipse with the major axis running along the u axis, and the minor axis running along the v axis. Here is a plot of the level curves for $c = 0.05, 0.1, 0.3, 1, 1.7, 3$.



To “transplant” this plot to the x, y plane, and get a contour plot of f , note that with $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1)$, $u = \mathbf{u}_1 \cdot \mathbf{x}$ and $v = \mathbf{u}_2 \cdot \mathbf{x}$, where of course $\mathbf{x} = (x, y)$. Thus, $\mathbf{x} = u\mathbf{u}_1 + v\mathbf{u}_2$, and (u, v) is the coordinate vector for \mathbf{x} relative to the orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2\}$. The u axis is the line $v = 0$, which is the line through the origin and \mathbf{u}_1 . The v axis is the line $u = 0$, which is the line through the origin and \mathbf{u}_2 . To draw the level curves of f in the x, y plane, simply draw in the u and v axes in the x, y plane, and then transplant the figure by drawing it in relative to the u, v axes. Here is the result:



We see that the $\mathbf{x} = (0, 0)$ minimizes f , and the graph $z = f(x, y)$ “curves upward” from its minimum.

It turns out that *every* quadratic function in *any number of variables* has a preferred orthonormal basis such that in coordinates based on that basis, it is given by a sum of multiples of squares. This makes it easy to understand the nature of its level sets.

We now explain how to find this preferred orthonormal basis. The methods we now develop are extremely useful in a great many problems; their utility goes *far* beyond the problem at hand.

6.1.3 Eigenvalues and eigenvectors of a symmetric matrix

Definition 67 (Eigenvectors and eigenvalues). *Let A be any $n \times n$ matrix. A number μ is an eigenvalue of A in case there is some non-zero vector \mathbf{v} such that*

$$A\mathbf{v} = \mu\mathbf{v}.$$

In this case, the vector \mathbf{v} is an eigenvector of A .

There are many kinds of problems involving an $n \times n$ matrix in which it is very helpful to find all of the eigenvectors of A . This is because A has a very simple effect on eigenvectors: If \mathbf{v} is an eigenvector of A , $A\mathbf{v}$ is simply a scalar multiple of \mathbf{v} .

Some $n \times n$ matrices have no eigenvectors in \mathbb{R}^n at all: Consider the 2×2 rotation matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$: $A(a, b) = (b, -a)$. $A\mathbf{v}$ is the counter-clockwise rotation of \mathbf{v} through the angle $\pi/2$, and hence for any non-zero \mathbf{v} , $A\mathbf{v}$ cannot be a multiple of \mathbf{v} .

However, as the following theorem says, if A is symmetric, there will always be eigenvectors, and plenty of them: If A is an $n \times n$ matrix, there exists an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n consisting of eigenvectors of A . It is one of the most important theorems in Linear Algebra.

Theorem 71 (The Spectral Theorem). *Let A be any $n \times n$ symmetric matrix. Then there is an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n and a set of n real numbers $\{\mu_1, \dots, \mu_n\}$ such that*

$$A\mathbf{u}_j = \mu_j\mathbf{u}_j$$

for each $j = 1, \dots, n$. That is, the orthonormal basis $\{\mu_1, \dots, \mu_n\}$ consists of eigenvectors of A .

The set of numbers $\{\mu_1, \dots, \mu_n\}$ is called the *Spectrum* of A , hence the name of the theorem. One way we shall use it is as follows: Given $\mathbf{x} \in \mathbb{R}^n$, define

$$\mathbf{y} = (y_1, \dots, y_n) = (\mathbf{x} \cdot \mathbf{u}_1, \dots, \mathbf{x} \cdot \mathbf{u}_n)$$

so that \mathbf{y} is the coordinate vector of \mathbf{x} with respect to the orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$. That is,

$$\mathbf{x} = \sum_{j=1}^n y_j \mathbf{u}_j .$$

Therefore,

$$A\mathbf{x} = \sum_{j=1}^n y_j A\mathbf{u}_j = \sum_{j=1}^n y_j \mu_j \mathbf{u}_j ,$$

and hence

$$\mathbf{x} \cdot A\mathbf{x} = \left(\sum_{j=1}^n y_j \mathbf{u}_j \right) \cdot \left(\sum_{j=1}^n y_j \mu_j \mathbf{u}_j \right) = \sum_{j=1}^n \mu_j y_j^2 , \quad (6.1)$$

which displays $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$ as a sum of squares. From here it is easy to prove the following:

Theorem 72. *Let A be a symmetric $n \times n$ matrix, and let f be the purely quadratic function defined by $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$. Then*

$$f(\mathbf{x}) > f(\mathbf{0}) = 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}$$

if and only if all of the eigenvalues of A are strictly positive. Likewise,

$$f(\mathbf{x}) < f(\mathbf{0}) = 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}$$

if and only if all of the eigenvalues of A are strictly negative.

Proof. Suppose that $f(\mathbf{x}) > f(\mathbf{0}) = 0$ for all non-zero \mathbf{x} . Taking $\mathbf{x} = \mathbf{u}_j$, we see that

$$\mu_j = \mathbf{u}_j \cdot A\mathbf{u}_j = f(\mathbf{u}_j) > 0$$

so in this case each eigenvalue is strictly positive. On the other hand, if each eigenvalue is strictly positive, the identity (6.1) says that $f(\mathbf{x}) \geq 0$ with equality if and only if each $y_j = 0$. But this is the case if and only if $\mathbf{x} = \mathbf{0}$. This proves the first assertion; the proof of the second is similar. \square

Lemma 17. *Let A be a symmetric $n \times n$ matrix, and define $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$. Then f is continuously differentiable, and*

$$\nabla f(\mathbf{x}) = 2A\mathbf{x} . \quad (6.2)$$

Proof. Since $f(\mathbf{x})$ is a quadratic polynomial in the variables x_1, \dots, x_n , it is continuously differentiable. By the chain rule, for any $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$,

$$\nabla f(\mathbf{x}) \cdot \mathbf{v} = \lim_{t \rightarrow 0} \frac{1}{t} (f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})) .$$

By definition, $(f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})) = (\mathbf{x} + t\mathbf{v}) \cdot A(\mathbf{x} + t\mathbf{v}) - \mathbf{x} \cdot A\mathbf{x} = t(\mathbf{x} \cdot A\mathbf{v} + \mathbf{v} \cdot A\mathbf{x}) + t^2 \mathbf{v} \cdot A\mathbf{v}$. Therefore,

$$\nabla f(\mathbf{x}) \cdot \mathbf{v} = \mathbf{x} \cdot A\mathbf{v} + \mathbf{v} \cdot A\mathbf{x} . \quad (6.3)$$

Recall the identity $\mathbf{y} \cdot A\mathbf{x} = (A^T \mathbf{y} \cdot \mathbf{x})$ proved in Theorem 58. Since A is symmetric, $\mathbf{x} \cdot A\mathbf{v} = A\mathbf{x} \cdot \mathbf{v}$, and then (6.3) becomes $\nabla f(\mathbf{x}) \cdot \mathbf{v} = 2A\mathbf{x} \cdot \mathbf{v}$. Since \mathbf{v} is arbitrary in \mathbb{R}^n , this implies (6.2). \square

Lemma 17 says, in particular, that $\mathbf{0}$ is always a critical point of the quadratic function $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$, and it is the only critical point if and only if A is invertible. More generally, the set of critical points of f is precisely the null space of A , so that when A is not invertible, there is a continuum of infinitely many critical points.

The next lemma again makes use of the identity $\mathbf{y} \cdot A\mathbf{x} = (A^T\mathbf{y} \cdot \mathbf{x})$ proved in Theorem 58.

Lemma 18 (Orthogonality lemma). *Let A be a symmetric matrix, and suppose that μ and λ are two distinct eigenvalues of A , and that $A\mathbf{v} = \mu\mathbf{v}$ and $A\mathbf{w} = \lambda\mathbf{w}$. Then \mathbf{v} and \mathbf{w} are orthogonal.*

Proof: By Theorem 58 and the hypotheses on \mathbf{v} and \mathbf{w} ,

$$\mu\mathbf{w} \cdot \mathbf{v} = \mathbf{w} \cdot (\mu\mathbf{v}) = \mathbf{w} \cdot A\mathbf{v} = A\mathbf{w} \cdot \mathbf{v} = \lambda\mathbf{w} \cdot \mathbf{v} .$$

Therefore, $0 = (\mu - \lambda)\mathbf{w} \cdot \mathbf{v}$, and since $\mu \neq \lambda$, $\mathbf{w} \cdot \mathbf{v} = 0$ □

The next lemma is the heart of the matter:

Lemma 19 (Maximum values and eigenvalues). *The function $f(\mathbf{x})$ on \mathbb{R}^n defined by*

$$f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$$

has maximizers \mathbf{u} on the unit sphere $S^{n-1} := \{\mathbf{x} : \|\mathbf{x}\| = 1\}$. The maximum value $\lambda_1 := f(\mathbf{u})$ is an eigenvalue of A , and every maximizer is an eigenvector with this eigenvalue.

Proof. The function f is a continuous function on the unit sphere S^{n-1} , and the unit sphere is closed, and bounded and non-empty. Thus, f has a maximizer on S^{n-1} .

To see that any maximizer is an eigenvector of A , with eigenvalue λ_1 , we use the method of Lagrange multipliers. Let $g(\mathbf{x}) = \|\mathbf{x}\|^2 - 1$ so that $g(\mathbf{x}) = 0$ is the equation specifying the unit sphere. By Lagrange's Theorem, Theorem 70, any maximizer \mathbf{u} of f subject to the constraint $g = 0$ satisfies the system of equations

$$\nabla f(\mathbf{u}) = \lambda \nabla g(\mathbf{u}) \tag{6.4}$$

as well as the constraint equation $g(\mathbf{x}) = 0$.

Evidently $\nabla g(\mathbf{u}) = 2\mathbf{u}$, and by Lemma 17 to see that $\nabla f(\mathbf{u}) = 2A\mathbf{u}$. (One could also apply Lemma 17 to g with $I_{n \times n}$ in place of A to see that $\nabla g(\mathbf{u}) = 2\mathbf{u}$, but since $g(\mathbf{x}) = \sum_{j=1}^n x_j^2$, this is not really needed.) Therefore, Lagrange's equation (6.4) becomes: $A\mathbf{u} = \lambda\mathbf{u}$. This shows that \mathbf{u} is an eigenvector of A . Taking the dot product of both sides with \mathbf{u} , and using the constraint equation, we obtain

$$f(\mathbf{u}) = \mathbf{u} \cdot A\mathbf{u} = \lambda\mathbf{u} \cdot \mathbf{u} = \lambda .$$

Therefore, since \mathbf{u} is a maximizer, λ is the maximum value of f . □

Proof of the Spectral Theorem. Suppose A is an $n \times n$ symmetric matrix and introduce the function $f(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x}$ on \mathbb{R}^n . We have already seen that f has maximizers on the unit sphere S^{n-1} , and that any maximizer is an eigenvector of A whose eigenvalue is the maximum value of f on S^{n-1} . Let \mathbf{u}_1 be any such maximizer, and let λ_1 be the corresponding eigenvalue.

We now consider a constrained optimization problem with *two* constraints. Namely, we seek to maximize f subject to the constraints $\|\mathbf{x}\| = 1$ and $\mathbf{x} \cdot \mathbf{u}_1 = 0$. The set of points \mathbf{x} satisfying these two

constraints is the intersection of the unit sphere and a hyperplane through the origin. In particular, it is closed, bounded, non-empty set for all $n \geq 2$. Therefore, f will have a maximum on this set. To find a maximizer, we introduce

$$g_0(\mathbf{x}) = \|\mathbf{x}\|^2 - 1 \quad \text{and} \quad g_1(\mathbf{x}) = 2\mathbf{u}_1 \cdot \mathbf{x}.$$

By Lagrange's Theorem, any maximizer \mathbf{u} of f subject to these constraints satisfies

$$\nabla f(\mathbf{u}) = \lambda \nabla g_0(\mathbf{u}) + \mu_1 \nabla g_1(\mathbf{u})$$

for some number λ and μ_1 . The explicit form of this is $A\mathbf{u} = \lambda\mathbf{u} + \mu\mathbf{u}_1$. Now, take the dot product of both sides with respect to \mathbf{u}_1 , and use the fact that \mathbf{u} is orthogonal to \mathbf{u}_1 to obtain $\mathbf{u}_1 \cdot A\mathbf{u} = \mu_1$. By Theorem 58,

$$\mathbf{u}_1 \cdot A\mathbf{u} = (A\mathbf{u}_1 \cdot \mathbf{u}) = \lambda\mathbf{u}_1 \cdot \mathbf{u} = 0.$$

Therefore, $\mu_1 = 0$, and our equation reduces to $A\mathbf{u} = \lambda\mathbf{u}$. Thus, \mathbf{u} is an eigenvector of A , and is a unit vector that is orthogonal to \mathbf{u}_1 . Call it \mathbf{u}_2 , and call the corresponding eigenvalue λ_2 .

Now that we have constructed the orthonormal set $\{\mathbf{u}_1, \mathbf{u}_2\}$ consisting on eigenvectors of A , we are done in case $n = 2$. We now show how to iterate this procedure to find the rest of the orthonormal basis for $n \geq 3$. Suppose that for $k < n$, and that we have already found an orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ consisting of eigenvectors of A with eigenvalues $\{\lambda_1, \dots, \lambda_k\}$ such that $\lambda_1 \geq \dots \geq \lambda_k$. Let $g_0(\mathbf{x}) = \|\mathbf{x}\|^2 - 1$ as above, and define

$$g_j(\mathbf{x}) = 2\mathbf{x} \cdot \mathbf{u}_j \quad \text{for } j = 1, \dots, k.$$

The set of points satisfying the constraint equation $g_j(\mathbf{x}) = 0$, $j = 0, 1, \dots, k$ is a closed and bounded subset of \mathbb{R}^n . For $k < n$, it is also non-empty. Hence f will have a maximizer \mathbf{u} on this set. By Lagrange's Theorem, there are numbers λ and μ_1, \dots, μ_k such that

$$\nabla f(\mathbf{u}) = \lambda \nabla g(\mathbf{u}) + \sum_{j=1}^k \mu_j \nabla g_j(\mathbf{u})$$

which has the explicit form

$$A\mathbf{u} = \lambda\mathbf{u} + \sum_{j=1}^k \mu_j \mathbf{u}_j. \tag{6.5}$$

Now take the dot product of both sides with \mathbf{u}_ℓ for $1 \leq \ell \leq k$. By the orthogonality relations, we obtain $\mathbf{u}_\ell \cdot A\mathbf{u} = \mu_\ell$. By Theorem 58, $\mathbf{u}_\ell \cdot A\mathbf{u} = (A\mathbf{u}_\ell \cdot \mathbf{u}) = \lambda_\ell \mathbf{u}_\ell \cdot \mathbf{u} = 0$. Therefore (6.5) reduces to $A\mathbf{u} = \lambda\mathbf{u}$, so that \mathbf{u} is an eigenvector of A . Define $\mathbf{u}_{k+1} = \mathbf{u}$ and $\lambda_{k+1} = \lambda$, and note that λ_{k+1} , being the maximum value of f subject to one additional constraint, is no larger than λ_k . By construction, $\{\mathbf{u}_1, \dots, \mathbf{u}_{k+1}\}$ is an orthonormal set in \mathbb{R}^n consisting of eigenvectors of A , and with the corresponding sequence of eigenvalues in non-increasing order.

Clearly we may repeat the procedure to produce the orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$. We cannot continue the procedure beyond this point, since the only vector \mathbf{x} satisfying $\mathbf{u}_j \cdot \mathbf{x} = 0$ for $j = 1, \dots, n$ is the zero vector, but it does not satisfy $g_0(\mathbf{x}) = 0$. Hence, the set of vectors satisfying the $n+1$ constraints is empty. \square

A *diagonal matrix* is a square ($n \times n$) matrix Λ such that $\Lambda_{i,j} = 0$ if $i \neq j$. $\Lambda_{j,j}$ is called the j th *diagonal entry* of Λ . Diagonal matrices are very simple to work with; for example, it is a trivial matter to solve $\Lambda\mathbf{x} = \mathbf{b}$ when Λ is diagonal.

The Spectral Theorem allows us to factor any $n \times n$ symmetric matrix as $A = U\Lambda U^T$ where U is an $n \times n$ orthogonal matrix.

To see this, let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be a orthonormal basis of \mathbb{R}^n consisting of eigenvectors of A , and let $A\mathbf{u}_j = \mu_j \mathbf{u}_j$, $j = 1, \dots, n$. Then for any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j$. Hence

$$A\mathbf{x} = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) A\mathbf{u}_j = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mu_j \mathbf{u}_j = [\mathbf{u}_1, \dots, \mathbf{u}_n] (\mu_1 \mathbf{x} \cdot \mathbf{u}_1, \dots, \mathbf{x} \cdot \mu_n \mathbf{u}_n) = U(\mu_1 \mathbf{x} \cdot \mathbf{u}_1, \dots, \mathbf{x} \cdot \mu_n \mathbf{u}_n) . \quad (6.6)$$

Now note that $U^T = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix}$ and so $(\mathbf{x} \cdot \mathbf{u}_1, \dots, \mathbf{x} \cdot \mathbf{u}_n) = U^T \mathbf{x}$. Then defining Λ to be the $n \times n$ diagonal matrix whose j th diagonal entry is μ_j ,

$$(\mu_1 \mathbf{x} \cdot \mathbf{u}_1, \dots, \mathbf{x} \cdot \mu_n \mathbf{u}_n) = \Lambda(\mathbf{x} \cdot \mathbf{u}_1, \dots, \mathbf{x} \cdot \mathbf{u}_n) = \Lambda U^T \mathbf{x} .$$

Combining this with (6.6), since \mathbf{x} is any vector in \mathbb{R}^n , we obtain $A = U\Lambda U^T$. We have proved

Theorem 73 (Diagonalization of Symmetric Matrices). *Let A be an $n \times n$ symmetric matrix. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be an orthonormal basis of \mathbb{R}^n such that $A\mathbf{u}_j = \mu_j \mathbf{u}_j$ for $j = 1, \dots, n$. Let $U := [\mathbf{u}_1, \dots, \mathbf{u}_n]$ and let Λ be the diagonal matrix whose j th diagonal entry is μ_j . Then*

$$A = U\Lambda U^T \quad \text{and} \quad \Lambda = U^T A U . \quad (6.7)$$

Our first application of this is to 3×3 symmetric matrices A ; we shall show if Λ is the 3×3 diagonal matrix associated to A by Theorem 73, then $\det(A) = \det(\Lambda)$. Later on when we study determinants in general, we shall prove that for any two $n \times n$ matrices B and C , $\det(BC) = \det(B)\det(C)$. Using this identity repeatedly,

$$\det(A) = \det(U) \det(\Lambda) \det(U^T) = \det(\Lambda) \det(U^T U) = \det(\Lambda) \det(I_{n \times n}) = \det(\Lambda) .$$

We have the means to readily prove the 3×3 version now. Let $B = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ be any 3×3 matrix. Then $\det(B) = \mathbf{v}_1 \times \mathbf{v}_2 \cdot \mathbf{v}_3$. Theorem 11, expressed in matrix language, says that if $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right handed orthonormal basis, and $U := [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$, then $U\mathbf{v}_1 \times U\mathbf{v}_2 = U(\mathbf{v}_1 \times \mathbf{v}_2)$, and then

$$(U\mathbf{v}_1 \times U\mathbf{v}_2) \cdot U\mathbf{v}_3 = U(\mathbf{v}_1 \times \mathbf{v}_2) \cdot U\mathbf{v}_3 = (\mathbf{v}_1 \times \mathbf{v}_2) \cdot U^T U\mathbf{v}_3 = (\mathbf{v}_1 \times \mathbf{v}_2) \cdot \mathbf{v}_3 .$$

That is, $\det[U\mathbf{v}_1, U\mathbf{v}_2, U\mathbf{v}_3] = \det[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$, which is the same as $\det(UB) = \det(B)$. By the same reasoning, if $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a left handed orthonormal basis, $\det(UB) = \det(B)$. Recall also from Exercise 4.20 that for all 3×3 matrices C , $\det(C) = \det(C^T)$; that is the triple product of the rows of C is the same as the triple product of the columns of C .

Now let A be a 3×3 symmetric matrix, and let $A = U\Lambda U^T$ as in Theorem 73. Define $B := \Lambda U^T$. By what we have noted just above, $\det(A) = \pm \det(B)$ where the plus sign is valid if the columns of U are a right handed orthonormal basis, and the minus sign is valid if the columns of U are a left handed orthonormal basis. But the $\det(B) = \det(B^T) = \det(U\Lambda) = \pm \det(\Lambda)$ with the same rule determining the sign. Altogether, $\det(A) = \det(\Lambda)$, regardless of whether the columns of U are a right or left handed orthonormal basis since the same sign comes in twice. Evidently, $\det(\Lambda) = \mu_1\mu_2\mu_3$, and so

$$\det(A) = \mu_1\mu_2\mu_3 . \quad (6.8)$$

By “embedding” the general 2×2 symmetric matrix $\begin{bmatrix} a & c \\ c & b \end{bmatrix}$ in the 3×3 symmetric matrix $\begin{bmatrix} a & c & 0 \\ c & b & 0 \\ 0 & 0 & 1 \end{bmatrix}$, we may apply our conclusions to 2×2 symmetric matrices to conclude that if A is such a matrix with eigenvalues μ_1 and μ_2 , then

$$\det(A) = \mu_1\mu_2 . \quad (6.9)$$

6.1.4 Computing eigenvectors and eigenvalues

Now that we know orthonormal bases consisting of eigenvectors of symmetric matrices exist, how do we find them? This is not hard for $n = 2$ or $n = 3$, and we shall now explain a general method that is applicable for all n , but is usually only feasible for small values of n .

Given an $n \times n$ matrix A , the *eigenvalue problem* for A is to find all eigenvalues and all eigenvectors of A . In the equation that defines eigenvalues and eigenvectors, $A\mathbf{v} = \mu\mathbf{v}$, only A is given and both μ and \mathbf{v} are unknown, apart from the requirement that $\mathbf{v} \neq \mathbf{0}$. This last statement may not seem like much, but it unlocks everything: If μ is an eigenvalue of A , then there is a non-zero vector \mathbf{v} satisfying $(A - \mu I_{n \times n})\mathbf{v} = \mathbf{0}$, and then $(A - \mu I_{n \times n})$ is not invertible. Conversely, if $(A - \mu I_{n \times n})$ is not invertible, $\text{Null}(A) \neq \{\mathbf{0}\}$, and hence there exists a non-zero \mathbf{v} such that $(A - \mu I_{n \times n})\mathbf{v} = \mathbf{0}$, which is the same as $A\mathbf{v} = \mu\mathbf{v}$. We have proved:

Theorem 74 (Eigenvalues and invertibility). *Let A be an $n \times n$ matrix. Then μ is an eigenvalue of A if and only if $A - \mu I$ is not invertible.*

Notice that this theorem is valid whether or not A is symmetric. As we have seen, for $n = 2$ and $n = 3$, $A - \mu I$ is not invertible if and only if $\det(A - \mu I) = 0$. (Later we will define the determinant of an $n \times n$ matrix for all n , and then the statement will be true for all n .) Hence, to find all of the eigenvalues of a 2×2 or 3×3 matrix A , compute

$$p(t) := \det(A - tI) ,$$

which will be a polynomial of degree $n = 2$ or $n = 3$ in t , and then solve the equation $p(t) = 0$. *The solutions; i.e., the roots of this polynomial, are the eigenvalues of A .*

Example 89 (Eigenvalues of a 2×2 matrix). Let $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Then

$$\det(A - tI) = \det\left(\begin{bmatrix} 2-t & 1 \\ 1 & 2-t \end{bmatrix}\right) = t^2 - 4t + 3.$$

The two roots of the quadratic polynomial are $t = 3$ and $t = 1$. Thus the eigenvalues are

$$\mu_1 = 3 \quad \text{and} \quad \mu_2 = 1.$$

Now that you have the eigenvalues, finding the eigenvectors is really easy, at least for $n = 2$ or $n = 3$. Indeed, suppose μ is an eigenvalue of A . Form the matrix $A - \mu I$. We must then find a non-zero vector \mathbf{v} such that $(A - \mu I)\mathbf{v} = \mathbf{0}$. By the dot product formulation of matrix-vector multiplication, this is equivalent to \mathbf{v} being orthogonal to each row of $A - \mu I$.

In particular, if $n = 2$, and the first row of $A - \mu I$ is $(a, b) \neq (0, 0)$, then \mathbf{v} must be a multiple of $(a, b)^\perp = (-b, a)$. The same reasoning applies to the second row. If both rows are zero, it means $A = \mu I$, so every non-zero vector is an eigenvector of A .

Moreover, when the two eigenvalues are distinct, Lemma 18 say the second eigenvector must be orthogonal to the first, so you do not need to solve for the second eigenvector once you have the first – any non-zero vector orthogonal to the first eigenvector will do.

Example 90 (Eigenvectors of a 2×2 matrix). Let $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. We have seen above that the two eigenvalues are $\mu_1 = 3$ and $\mu_2 = 1$. Form

$$A - 3I = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The first row is $(-1, 1)$ so $(-1, 1)^\perp = (-1, -1)$ is an eigenvector with eigenvalue 3, as you can check. Any non-zero multiple of this, so $(1, 1)$ is also an eigenvector with eigenvalue 3.

To get the second eigenvector, we simply use Lemma 18 which says that this must be orthogonal to the one we have already found. Since the first eigenvector is orthogonal to the first row of $A - 3I$, the first row of $A - 3I$ will be an eigenvector with the second eigenvalue. We do not even have to write down $A - I$ in this case. Thus, as you can check, $(-1, 1)$ is an eigenvector with eigenvalue 1. Normalizing these, we get our orthonormal basis:

$$\{\mathbf{u}_1, \mathbf{u}_2\} = \left\{ \frac{1}{\sqrt{2}}(1, 1), \frac{1}{\sqrt{2}}(-1, 1) \right\}.$$

The situation for $n = 3$ is not much more complicated. If \mathbf{v} is orthogonal to each row of $A - \mu I$, and any two of these rows are not multiples of one another, \mathbf{v} must be a multiple of the cross product of these two rows. On the other hand, if all of the rows are multiples of the same vector \mathbf{r} , we know how to get an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ such that \mathbf{u}_1 is a multiple of \mathbf{r} . Then \mathbf{u}_2 and \mathbf{u}_3 are eigenvectors with eigenvalue μ .

The challenge in $n = 3$ is in finding the eigenvalues, since this involves solving for the roots of a cubic polynomial, and the formulas for this are a bit cumbersome. However, if one root can be found “by inspection” the other two can be found by solving a quadratic equation.

6.2 The best quadratic approximation

6.2.1 Higher order directional derivatives and repeated partial differentiation

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Given \mathbf{x}_0 and $\mathbf{v} \in \mathbb{R}^n$, we then define, as usual, the “slice” function $g : \mathbb{R} \rightarrow \mathbb{R}$ by $g(t) = f(\mathbf{x}_0 + t\mathbf{v})$. *The higher order derivatives of g at $t = 0$, if they exists, are the higher order directional derivatives of f at \mathbf{x}_0 in the direction \mathbf{v} .*

By the chain rule

$$g'(t) = \mathbf{v} \cdot \nabla f(\mathbf{x}_0 + t\mathbf{v}) . \quad (6.10)$$

Taking higher derivatives will tell us how the graph of $z = f(\mathbf{x})$ “curves away” from its tangent plane at a given point \mathbf{x}_0 . Does it “curve up”, or “curve down”, or does it “curve up in some directions and down in others”? In many applications of multivariable calculus, it is important to be able to answer these questions. There are two issues before us: We need to develop an efficient means of computing higher order directional derivatives, and then, once we know how to compute them, we need to understand what they are telling us about the geometry of the graph of $z = f(\mathbf{x})$. In the rest of this subsection, we focus on the issue of how to compute.

To take a second derivative, it helps to write the dot product in (6.11) as an explicit sum:

$$g'(t) = \sum_{i=1}^n v_i \frac{\partial}{\partial x_i} f(\mathbf{x}_0 + t\mathbf{v}) . \quad (6.11)$$

Then since the v_i do not depend on t , and since the derivative of a sum is the sum of the derivatives, we can apply the chain rule once more to compute

$$\begin{aligned} g''(t) &= \sum_{i=1}^n v_i \frac{d}{dt} \left(\frac{\partial}{\partial x_i} f(\mathbf{x}_0 + t\mathbf{v}) \right) \\ &= \sum_{i=1}^n v_i \left(\mathbf{v} \cdot \nabla \frac{\partial}{\partial x_i} f \right) (\mathbf{x}_0 + t\mathbf{v}) \\ &= \sum_{i,j=1}^n v_i v_j \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} f(\mathbf{x}_0 + t\mathbf{v}) . \end{aligned} \quad (6.12)$$

To be clear about the notation we are using, $\frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} f$ denotes the function you get by first taking the x_i partial derivative of f , and then taking the x_j partial derivative of that. The following more compact notation is often used to denote the same thing: $\frac{\partial^2}{\partial x_j \partial x_i} f(x_1, \dots, x_n)$, and in case $j = i$, it is common to write $\frac{\partial^2}{\partial x_i^2} f(x_1, \dots, x_n)$. We shall use these notations interchangeably, which is the usual practice.

Example 91. Let $f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1x_2$. Then $\frac{\partial}{\partial x_1} f(x_1, x_2) = 3x_1^2 - 3x_2$ and so

$$\frac{\partial^2}{\partial x_1^2} f(x_1, x_2) = 6x_1 \quad \text{and} \quad \frac{\partial^2}{\partial x_2 \partial x_1} f(x_1, x_2) = -3 .$$

Likewise, $\frac{\partial}{\partial x_2} f(x_1, x_2) = 3x_2^2 - 3x_1$ and so

$$\frac{\partial^2}{\partial x_2^2} f(x_1, x_2) = 6x_2 \quad \text{and} \quad \frac{\partial^2}{\partial x_1 \partial x_2} f(x_1, x_2) = -3 .$$

Notice that computing second partial derivatives is only a matter of computing one variable derivatives with respect to various pairs of variables.

The formula

$$\left. \frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}) \right|_{t=0} = \sum_{i,j=1}^n v_i v_j \frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x}_0) \quad (6.13)$$

can be written in matrix notation, and this turns out to much more useful than one might first guess.

Definition 68 (Hessian matrix). Let f be a function defined on an open set $U \subset \mathbb{R}^n$ with values in \mathbb{R} such that all of the second order partial derivatives of f exist and are continuous in U . Then at any $\mathbf{x} \in U$, the Hessian matrix of f at \mathbf{x} is the $n \times n$ matrix $[\text{Hess}_f(\mathbf{x})]$ whose i,j th entry is

$$[\text{Hess}_f(\mathbf{x})]_{i,j} := \frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x}) .$$

With this definition, $([\text{Hess}_f(\mathbf{x})]\mathbf{v})_i = \sum_{j=1}^n [\text{Hess}_f(\mathbf{x})]_{i,j} v_j = \sum_{j=1}^n v_j \frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x})$, and so

$$\mathbf{v} \cdot [\text{Hess}_f(\mathbf{x})]\mathbf{v} = \sum_{i,j=1}^n v_i v_j \frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x}_0) .$$

Therefore, we can rewrite our formula (6.13) as

$$\left. \frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}) \right|_{t=0} = \mathbf{v} \cdot [\text{Hess}_f(\mathbf{x}_0)]\mathbf{v} . \quad (6.14)$$

Example 92. Let $f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1x_2$ as in Example 91. Let us compute the second order directional derivative $\left. \frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}) \right|_{t=0}$ for $\mathbf{x}_0 := (-1, -1)$ and $\mathbf{v} = (u, v)$.

By our computations in Example 91, $[\text{Hess}_f(\mathbf{x})] = \begin{bmatrix} 6x_1 & -3 \\ -3 & 6x_2 \end{bmatrix}$. Therefore,

$$[\text{Hess}_f(-1, -1)] = -3 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} .$$

Now we compute

$$\begin{aligned} \mathbf{v} \cdot [\text{Hess}_f(-1, -1)]\mathbf{v} &= -3(u, v) \cdot \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} (u, v) \\ &= -3(u, v) \cdot (2u + v, u + 2v) \\ &= -6(u^2 + uv + v^2) . \end{aligned}$$

In the previous example, the Hessian turned out to be a symmetric matrix. We next prove Clairault's Theorem which says that this was no coincidence.

6.2.2 Clairault's Theorem

At each \mathbf{x} where the second order partial derivatives of f all exist and are continuous, the the Hessian of f is a symmetric matrix:

Theorem 75 (Clairault's Theorem). *$f(x, y)$ be a function of the two variables x and y such that all partial derivatives of f order 2 are continuous in a neighborhood of (x_0, y_0) . Then*

$$\frac{\partial}{\partial x} \frac{\partial}{\partial y} f(x_0, y_0) = \frac{\partial}{\partial y} \frac{\partial}{\partial x} f(x_0, y_0) . \quad (6.15)$$

Proof. The basic idea of the proof is to take some small $h > 0$, and then compute a formula

$$f(x_0 + h, y_0 + h) - f(x_0, y_0)$$

in terms of the mixed second order partial derivatives of f two different ways: First, we keep track of how $f(x, y)$ changes along the path going straight from (x_0, y_0) to $(x_0 + h, y_0)$, and then straight from $(x_0 + h, y_0)$ to $(x_0 + h, y_0 + h)$. Second, we keep track of how $f(x, y)$ changes along the path going straight from (x_0, y_0) to $(x_0, y_0 + h)$, and then straight from $(x_0, y_0 + h)$ to $(x_0 + h, y_0 + h)$.

Let's begin: By the Fundamental Theorem of Calculus,

$$f(x_0 + h, y_0) - f(x_0, y_0) = \int_0^h \frac{\partial}{\partial x} f(x_0 + t, y_0) dt$$

and

$$f(x_0 + h, y_0 + h) - f(x_0 + h, y_0) = \int_0^h \frac{\partial}{\partial y} f(x_0 + h, y_0 + t) dt .$$

Together we have

$$\begin{aligned} f(x_0 + h, y_0 + h) - f(x_0, y_0) &= \\ &\int_0^h \frac{\partial}{\partial x} f(x_0 + t, y_0) dt + \int_0^h \frac{\partial}{\partial y} f(x_0 + h, y_0 + t) dt . \end{aligned} \quad (6.16)$$

Going along the other path we have

$$f(x_0, y_0 + h) - f(x_0, y_0) = \int_0^h \frac{\partial}{\partial y} f(x_0, y_0 + t) dt$$

and

$$f(x_0 + h, y_0 + h) - f(x_0, y_0 + h) = \int_0^h \frac{\partial}{\partial x} f(x_0 + t, y_0 + h) dt .$$

Together we have

$$\begin{aligned} f(x_0 + h, y_0 + h) - f(x_0, y_0) &= \\ &\int_0^h \frac{\partial}{\partial x} f(x_0 + t, y_0 + h) dt + \int_0^h \frac{\partial}{\partial y} f(x_0 + h, y_0 + t) dt . \end{aligned} \quad (6.17)$$

The since (6.16) and (6.17) give two different formulas for the same thing, the differences between the right hand sides must be zero. After a bit of algebra, we conclude that

$$\begin{aligned} \int_0^h \left(\frac{\partial}{\partial y} f(x_0 + h, y_0 + t) - \frac{\partial}{\partial y} f(x_0, y_0 + t) \right) dt &= \\ \int_0^h \left(\frac{\partial}{\partial x} f(x_0 + t, y_0 + h) - \frac{\partial}{\partial x} f(x_0 + t, y_0) \right) dt . \end{aligned}$$

By the Mean Value Theorem, for some $0 \leq t_1 \leq h$

$$\begin{aligned} \int_0^h \left(\frac{\partial}{\partial y} f(x_0 + h, y_0 + t) - \frac{\partial}{\partial y} f(x_0, y_0 + t) \right) dt = \\ h \left(\frac{\partial}{\partial y} f(x_0 + h, y_0 + t_1) - \frac{\partial}{\partial y} f(x_0, y_0 + t_1) \right) . \end{aligned}$$

Now on the right hand side only x changes, and by the Mean Value Theorem once more, this time applied to the variation in x , for some $0 \leq s_1 \leq h$,

$$h \left(\frac{\partial}{\partial y} f(x_0 + h, y_0 + t_1) - \frac{\partial}{\partial y} f(x_0, y_0 + t_1) \right) = h^2 \frac{\partial}{\partial x} \frac{\partial}{\partial y} f(x_0 + s_1, y_0 + t_1) .$$

In summary, for some $0 \leq s_1, t_1 \leq h$,

$$\frac{\partial}{\partial x} \frac{\partial}{\partial y} f(x_0 + s_1, y_0 + t_1) = \frac{1}{h^2} \int_0^h \left(\frac{\partial}{\partial y} f(x_0 + h, y_0 + t) - \frac{\partial}{\partial y} f(x_0, y_0 + t) \right) dt . \quad (6.18)$$

In exactly the same way we deduce that for some $0 \leq s_2, t_2 \leq h$

$$\frac{\partial}{\partial y} \frac{\partial}{\partial x} f(x_0 + s_2, y_0 + t_2) = \frac{1}{h^2} \int_0^h \left(\frac{\partial}{\partial x} f(x_0 + t, y_0 + h) - \frac{\partial}{\partial x} f(x_0 + t, y_0) \right) dt . \quad (6.19)$$

Combining (6.18), (6.18) and (6.19), we see that for some $0 \leq s_1, t_1 \leq h$ and some $0 \leq s_2, t_2 \leq h$,

$$\frac{\partial}{\partial x} \frac{\partial}{\partial y} f(x_0 + s_1, y_0 + t_1) = \frac{\partial}{\partial y} \frac{\partial}{\partial x} f(x_0 + s_2, y_0 + t_2) .$$

Now take the limit $h \rightarrow 0$, along which s_1, s_2, t_1, t_2 all tend to zero. By the continuity of the second order partial derivatives, we conclude that (6.15) is true. \square

Theorem 76 (Symmetry of the Hessian). *f be a function defined on an open set $U \subset \mathbb{R}^n$ with values in \mathbb{R} . Suppose that all partial derivatives of f order 2 are defined and continuous in U. Then for all $\mathbf{x} \in U$, $[\text{Hess}_f(\mathbf{x})]$ is a symmetric $n \times n$ matrix.*

Proof. When we compute $\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$ and $\frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x})$ we are only concerned with the two variables x_i and x_j , so this is an immediate consequence of Clairault's Theorem. \square

6.2.3 A multivariable second order Taylor expansion

Let f a a real valued function defined on a open, convex set $U \subset \mathbb{R}^n$, and suppose that all of the second order partial derivatives of f exist in U and are continuous there. Let $\mathbf{x}_0 \in U$, and $\mathbf{v} \in \mathbb{R}^n$ be such that $\mathbf{x}_0 + \mathbf{v} \in U$. Then since U is convex, $\mathbf{x}_0 + t\mathbf{v} \in U$ for all $t \in (0, 1)$.

Then by the Chain Rule and (6.14) if we define $g(t) := f(\mathbf{x}_0 + t\mathbf{v})$, then

$$g(0) = f(\mathbf{x}_0), \quad g'(0) = \nabla f(\mathbf{x}_0) \cdot \mathbf{v} \quad \text{and} \quad g''(t) = \mathbf{v} \cdot [\text{Hess}_f(\mathbf{x}_0 + t\mathbf{v})]\mathbf{v} . \quad (6.20)$$

Recall that by Taylor's Theorem with Remainder,

$$g(t) = g(0) + g'(0)t + \frac{1}{2}t^2 g''(0) + \int_0^t (t-s)(g''(s) - g''(0))ds . \quad (6.21)$$

(This is proved by using the Fundamental Theorem of Calculus to write $g(t) = g(0) + \int_0^t g'(s)ds$, and using the identity $1 = \frac{d}{ds}(s - t)$ together with integration by parts to bring in the second derivative of g , as in any text on single variable calculus.) Using (6.20) in (6.21),

$$f(\mathbf{x}_0 + t\mathbf{v}) = f(\mathbf{x}_0) + t\nabla f(\mathbf{x}_0) \cdot \mathbf{v} + \frac{t^2}{2}\mathbf{v} \cdot [\text{Hess}_f(\mathbf{x}_0 + t\mathbf{v})]\mathbf{v} \quad (6.22)$$

$$+ \int_0^t (t-s)\mathbf{v} \cdot A(t)\mathbf{v} \quad (6.23)$$

where $A(t) := [\text{Hess}_f(\mathbf{x}_0 + t\mathbf{v})] - [\text{Hess}_f(\mathbf{x}_0)]$. Since all second order partial derivatives of f exist and are continuous, all of the entries of $A(t)$ are continuous and equal to 0 at $t = 0$. Hence $\lim_{t \rightarrow 0} \|A(t)\|_F = 0$. It follows that for any $\epsilon > 0$, there is a $\delta > 0$ so that for $|t| < \delta$, $\|A(t)\|_F < \epsilon$, and then by the Cauchy-Schwarz inequality, $\mathbf{v} \cdot A(t)\mathbf{v} \leq \epsilon \|\mathbf{v}\|^2$ for $|t| < \delta$. For such t ,

$$\left| \int_0^t (t-s)\mathbf{v} \cdot A(t)\mathbf{v} \right| \leq \epsilon \|\mathbf{v}\|^2 \int_0^t |t-s| ds = \frac{t^2}{2} \epsilon \|\mathbf{v}\|^2$$

Altogether, we have proved:

Lemma 20. *Let f be a real valued function define an open convex set $U \subset \mathbb{R}^n$, where all of its second order partial derivatives exists and are continuous. Let $\mathbf{x}_0 \in U$, and let $\mathbf{v} \in \mathbb{R}$ by such that $\mathbf{x}_0 + \mathbf{v}$ in U , and then, since U is convex, $\mathbf{x}_0 + t\mathbf{v} \in U$ for all $t \in (0, 1)$. Then for all $\epsilon > 0$, there is a $\delta \in (0, 1)$ such that for all $t \in (0, \delta)$,*

$$\left| f(\mathbf{x}_0 + t\mathbf{v}) - f(\mathbf{x}_0) - t\nabla f(\mathbf{x}_0) \cdot \mathbf{v} - \frac{t^2}{2}\mathbf{v} \cdot [\text{Hess}_f(\mathbf{x}_0)]\mathbf{v} \right| < \epsilon t^2 \|\mathbf{v}\|^2. \quad (6.24)$$

We now rewrite this as follows. Let f , U and \mathbf{x}_0 be given as in Lemma 20. Let $\mathbf{x} \in U$, and define $t := \|\mathbf{x} - \mathbf{x}_0\|$ and $\mathbf{v} := \frac{\mathbf{x} - \mathbf{x}_0}{\|\mathbf{x} - \mathbf{x}_0\|}$ so that $t\mathbf{v} = \mathbf{x} - \mathbf{x}_0$, $\mathbf{x} = \mathbf{x}_0 + t\mathbf{v}$. Then, with δ and ϵ as in Lemma 20,

$$\left| f(\mathbf{x}) - f(\mathbf{x}_0) - \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot [\text{Hess}_f(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) \right| < \epsilon \|\mathbf{x} - \mathbf{x}_0\|^2. \quad (6.25)$$

Definition 69 (Twice differentiable function on \mathbb{R}^n). *Let U be an open subset of \mathbb{R}^n and $\mathbf{x}_0 \in U$. A real valued function f on U is twice differentiable at \mathbf{x}_0 in case for all $\epsilon > 0$, there is a vector $\mathbf{v} \in \mathbb{R}^n$ and a symmetric $n \times n$ matrix A and a $\delta > 0$ such that*

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow \left| f(\mathbf{x}) - f(\mathbf{x}_0) - \mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) \right| < \epsilon \|\mathbf{x} - \mathbf{x}_0\|^2. \quad (6.26)$$

An equivalent way to express (6.26), which “hides” the ϵ and δ in the definition of the limit, is

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|f(\mathbf{x}) - f(\mathbf{x}_0) - \mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|^2} = 0. \quad (6.27)$$

Note that function satisfying (6.27) is automatically differentiable at \mathbf{x}_0 since

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{x}_0) - \mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0)| &\leq \left| f(\mathbf{x}) - f(\mathbf{x}_0) - \mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) \right| \\ &+ \left| \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) \right|, \end{aligned}$$

and if we divide through by $\|\mathbf{x} - \mathbf{x}_0\|$, both terms on the right tend to 0 as $\mathbf{x} \rightarrow \mathbf{x}_0$. Hence a twice differentiable function is differentiable, and in turn, continuous.

Theorem 77. Let f be twice differentiable at \mathbf{x}_0 . Then there is exactly one vector \mathbf{v} and one symmetric matrix A for which (6.27) is valid.

Proof. Suppose that (6.27) is also valid with \mathbf{w} in place of \mathbf{v} and B in place of A . Define

$$\begin{aligned} h(\mathbf{x}) &:= f(\mathbf{x}_0) + \mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) \\ g(\mathbf{x}) &:= f(\mathbf{x}_0) + \mathbf{w} \cdot (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot B(\mathbf{x} - \mathbf{x}_0). \end{aligned}$$

Then by the triangle inequality

$$\frac{|h(\mathbf{x}) - g(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|^2} \leq \frac{|f(\mathbf{x}) - h(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|^2} + \frac{|f(\mathbf{x}) - g(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|^2},$$

and therefore $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|h(\mathbf{x}) - g(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|^2} = 0$, which is the same as

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \left| (\mathbf{v} - \mathbf{w}) \cdot \frac{\mathbf{x} - \mathbf{x}_0}{\|\mathbf{x} - \mathbf{x}_0\|^2} + \frac{1}{2} \frac{\mathbf{x} - \mathbf{x}_0}{\|\mathbf{x} - \mathbf{x}_0\|} \cdot (A - B) \frac{\mathbf{x} - \mathbf{x}_0}{\|\mathbf{x} - \mathbf{x}_0\|^2} \right| = 0. \quad (6.28)$$

First, suppose that $\mathbf{v} \neq \mathbf{w}$, and define $\mathbf{x} = \mathbf{x}_0 + t(\mathbf{v} - \mathbf{w})$. Then (6.28) says that

$$\lim_{t \rightarrow 0} \left| \frac{1}{t} + \frac{1}{2} \frac{\mathbf{v} - \mathbf{w}}{\|\mathbf{v} - \mathbf{w}\|} \cdot (A - B) \frac{\mathbf{w} - \mathbf{w}}{\|\mathbf{w} - \mathbf{w}\|^2} \right| = 0.$$

which is impossible. Hence $\mathbf{w} = \mathbf{v}$. Given this, and choosing any non-zero $\mathbf{y} \in \mathbb{R}^n$, and then taking $\mathbf{x} = \mathbf{x}_0 + t\mathbf{y}$, (6.28) says that

$$\lim_{t \rightarrow 0} \left| \frac{\mathbf{y}}{\|\mathbf{y}\|} \cdot (A - B) \frac{\mathbf{y}}{\|\mathbf{y}\|^2} \right| = 0,$$

and hence $\mathbf{y} \cdot A\mathbf{y} = \mathbf{y} \cdot B\mathbf{y}$ for all $\mathbf{y} \in \mathbb{R}^n$. Choosing $\mathbf{y} = \mathbf{e}_j$, $j = 1, \dots, n$, we have $A_{j,j} = B_{j,j}$ for each j . Next, choosing $\mathbf{y} = \mathbf{e}_i + \mathbf{e}_j$, we see that

$$A_{i,i} + A_{i,j} + A_{j,i} + A_{j,j} = B_{i,i} + B_{i,j} + B_{j,i} + B_{j,j}.$$

By what we just proved, and the fact that A and B are symmetric, this means that $A = B$. \square

Therefore, if f satisfies (6.27), we may regard \mathbf{v} as the first derivative of f as \mathbf{x}_0 , and A as the second derivative of f as \mathbf{x}_0 .

Theorem 78. Let f be a real valued function on a open set $U \subset \mathbb{R}^n$ such that all of the second order partial derivatives of f exist on U and are continuous at \mathbf{x}_0 . Then f is twice differentiable and the unique vector \mathbf{v} and the unique symmetric matrix A such that (6.27) is valid, or equivalently, (6.26) is valid for each $\epsilon > 0$, are given by

$$\mathbf{v} = \nabla f(\mathbf{x}_0) \quad \text{and} \quad A = [\text{Hess}_f(\mathbf{x}_0)].$$

Proof. The fact that (6.27) is valid with this choice of \mathbf{v} and A follows from Lemma (6.27), with the conclusion written in the form (6.25), one we observe that since U is open and contains \mathbf{x}_0 , it also contains $B_r(\mathbf{x}_0)$ for some $r > 0$, and $B_r(\mathbf{x}_0)$ is convex. The uniqueness of \mathbf{v} and A follows from Theorem 77. \square

Definition 70 (Quadratic function). *A quadratic function f on \mathbb{R}^n is a polynomial in (x_1, \dots, x_n) each term of which is of degree at most two. A purely quadratic function f on \mathbb{R}^n is a polynomial in (x_1, \dots, x_n) each term of which is of degree exactly two, the kind we have studied in the first section of this chapter.*

According to the definition, the general quadratic function has the form

$$q(\mathbf{x}) = a + \sum_{j=1}^n a_j x_j + \frac{1}{2} \sum_{i,j=1}^n A_{i,j} x_i x_j$$

for some coefficients $a, a_j, A_{i,j}$, $1 \leq i, j \leq n$. We can express this in vector notation as

$$q(\mathbf{x}) = a + \mathbf{a} \cdot \mathbf{x} + \mathbf{x} \cdot A\mathbf{x}$$

for some $a \in \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$ and A a symmetric $n \times n$ matrix.

Notice that for any a , \mathbf{a} and A as above, and any \mathbf{x}_0 , the function $h(\mathbf{x})$ given by

$$h(\mathbf{x}) = a + \mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0)$$

is also a quadratic function: By expanding the terms and regrouping, we also have

$$h(\mathbf{x}) = [a - b\mathbf{a} \cdot \mathbf{x}_0 + \mathbf{x}_0 \cdot A\mathbf{x}_0] + [\mathbf{a} - 2A\mathbf{x}_0] \cdot \mathbf{x} + \mathbf{x} \cdot A\mathbf{x} ,$$

which has the form of a quadratic function. This form is especially useful when studying the behavior of functions at \mathbf{x} near to \mathbf{x}_0 .

Definition 71 (Best quadratic approximation). *Let f be a function defined on an open set $U \subset \mathbb{R}^n$ with values in \mathbb{R} such that every second order partial derivatives of f exists and is continuous everywhere in U . For any $\mathbf{x}_0 \in U$, define the function $h(\mathbf{x})$ by*

$$h(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot [\text{Hess}_f(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) . \quad (6.29)$$

Then $h(\mathbf{x})$ is the best quadratic approximation to f at \mathbf{x}_0 .

The quadratic function $h(\mathbf{x})$ define by (6.29) is the “best” quadratic approximation in the sense that, according to Theorem 77, it is the *only* quadratic function such that

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|f(\mathbf{x}) - h(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|^2} = 0 . \quad (6.30)$$

6.2.4 Principal curvatures at a critical point

Let f be twice continuously differentiable on an open set $U \subset \mathbb{R}^n$, and suppose that \mathbf{x}_0 is a critical point of f . For any non-zero $\mathbf{v} \in \mathbb{R}^n$, define the function $g(t) = f(\mathbf{x}_0 + t\mathbf{v})$. By the Chain Rule, $g'(0) = 0$. Recall from single variable calculus that if $g''(0) > 0$, then $t = 0$ is a local minimizer of g , so that the graph of $y = g(t)$ “curves upward” as it passes through $(0, g(0))$. Likewise, if $g''(0) < 0$, then $t = 0$ is a local maximizer of g , so that the graph of $y = g(t)$ “curves downward” as it passes through $(0, g(0))$.

As we vary the vector \mathbf{v} , we will get different values for $g''(0)$, which specifies the curvature of the graph of the “one dimensional slice” of the graph of $f(\mathbf{x}_0 + t\mathbf{v})$ at $t = 0$. To compare different directions on an equal footing, we restrict our attention to the case in which \mathbf{v} is a unit vector \mathbf{u} . By (6.14)

$$\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{u}) \Big|_{t=0} = \mathbf{u} \cdot [\text{Hess}_f(\mathbf{x}_0)]\mathbf{u}. \quad (6.31)$$

We now ask: *What choice of \mathbf{u} maximizes $\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{u}) \Big|_{t=0}$, and what choice of \mathbf{u} minimizes $\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{u}) \Big|_{t=0}$?*

We have seen in the proof of the Spectral Theorem that the maximum value of $\mathbf{u} \cdot [\text{Hess}_f(\mathbf{x}_0)]\mathbf{u}$ as \mathbf{u} ranges of the set of all unit vectors is the largest eigenvalue of $[\text{Hess}_f(\mathbf{x}_0)]$, which we shall denote by μ_{\max} . In the same way, the minimum value of $\mathbf{u} \cdot [\text{Hess}_f(\mathbf{x}_0)]\mathbf{u}$ as \mathbf{u} ranges of the set of all unit vectors is the least eigenvalue of $[\text{Hess}_f(\mathbf{x}_0)]$, which we shall denote by μ_{\min} . Then by eqref{maininfo2B}, for all unit vectors \mathbf{u} ,

$$\mu_{\min} \leq \frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{u}) \Big|_{t=0} \leq \mu_{\max}, \quad (6.32)$$

and there is equality on the left if and only if $[\text{Hess}_f(\mathbf{x}_0)] = \mu_{\min}\mathbf{u}$, and there is equality on the right if and only if $[\text{Hess}_f(\mathbf{x}_0)] = \mu_{\max}\mathbf{u}$.

If $\mu_{\min} = \mu_{\max}$, then all of the eigenvalues of A are the same number, say c . But the $A = cI_{n \times n}$. To see this, let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be an orthonormal basis of \mathbb{R}^n consisting of eigenvectors of A . Then for any $\mathbf{x} \in \mathbb{R}^n$,

$$A\mathbf{x} = A \left(\sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j \right) = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) A\mathbf{u}_j = c \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j = c\mathbf{x} = cI_{n \times n}\mathbf{x}.$$

Otherwise, if $\mu_{\min} < \mu_{\max}$ then every eigenvector of $[\text{Hess}_f(\mathbf{x}_0)]$ with eigenvalue μ_{\min} is orthogonal to every eigenvector with eigenvalue μ_{\max} . Unit vectors \mathbf{u} such that either $[\text{Hess}_f(\mathbf{x}_0)] = \mu_{\min}\mathbf{u}$ or $[\text{Hess}_f(\mathbf{x}_0)] = \mu_{\max}\mathbf{u}$ are called *directions of principal curvature at the critical point \mathbf{x}_0* .

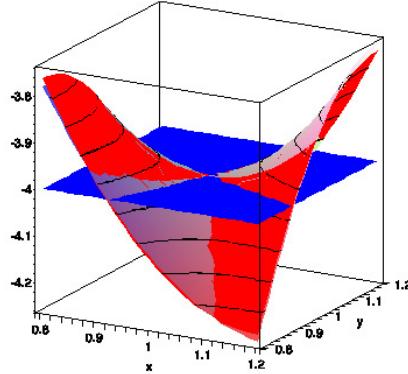
Example 93 (Directions of principal curvature at a critical point). Let $f(x, y) = x^3 + y^3 - 3xy$. We have seen in Example 91 that $\mathbf{x}_0 := (-1, -1)$ is a critical point, and that $[\text{Hess}_f(\mathbf{x}_0)] = -3 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

We have seen in Example 89 that the two eigenvalues of $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ are 3 and 1, so the two eigenvalues of $[\text{Hess}_f(\mathbf{x}_0)]$ are -9 and -3 . Hence $\mu_{\min} = -9$, and $\mu_{\max} = -3$. The two unit vectors \mathbf{u} for which $[\text{Hess}_f(\mathbf{x}_0)] = \mu_{\min}\mathbf{u}$ are $\mathbf{u} = \pm 2^{-1/2}(1, 1)$, and the two unit vectors \mathbf{u} for which $[\text{Hess}_f(\mathbf{x}_0)] = \mu_{\max}\mathbf{u}$ are $\mathbf{u} = \pm 2^{-1/2}(1, -1)$. By (6.32), for all init vectors \mathbf{u} ,

$$-9 \leq \frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{u}) \Big|_{t=0} \leq -3.$$

Hence the graph $z = f(x, y)$ “curves downward” in every direction at \mathbf{x}_0 . As we discuss next, this indicates that \mathbf{x}_0 is a local maximizer of f .

When $\mu_{\max} > 0$ and $\mu_{\min} < 0$ at a critical point \mathbf{x}_0 , the surface $z = f(x, y)$ near \mathbf{x}_0 curves upward at some directions, and downward in others. Here is a graph for another function that illustrates this:



The same sort of reasoning shows that the curved surface given by $z = f(x, y)$ “curves downward” from the tangent plane at \mathbf{x}_0 in case both μ_+ and μ_- are strictly negative. The plot above suggests the name introduced in the following definition:

Definition 72 (Saddle point). *Let f be twice continuously differentiable on an open set $U \subset \mathbb{R}^n$, and suppose that \mathbf{x}_0 is a critical point of f . If $\mu_{\max} > 0$ and $\mu_{\min} < 0$ at \mathbf{x}_0 , then \mathbf{x}_0 is a saddle point.*

6.2.5 Contour plots near critical points

What we have learned so far enables us to draw a contour plot of a twice continuously differentiable function $f(x, y)$ near any of its critical points \mathbf{x}_0 , at least when none of the eigenvalues of $[\text{Hess}_f(\mathbf{x}_0)]$ is zero.

Example 94 (Drawing a contour plot). *Let $f(x, y) = 2yx^2 + 4x^2 + xy + 8y^2 + 8x - y$. Then*

$$\nabla f(x, y) = (4xy + 8x + y + 8, 2x^2 + x - 1 + 16y) .$$

At a critical point (x, y) , we must have

$$4xy + 8x + y + 8 = 0 \quad \text{and} \quad 2x^2 + x - 1 + 16y = 0 .$$

From the first equation we see that $x = -\frac{y+8}{4y+8}$. Substituting this into the second, we see, after some algebra, that $y(476 + 503y + 128y^2) = 0$. This has three roots, one of course being $y = 0$. The corresponding x value is -1 . Hence one of the three critical points is $\mathbf{x}_0 := (-1, 0)$. Let us compute the principal curvatures at \mathbf{x}_0 , and sketch a contour plot for f near \mathbf{x}_0 .

We compute:

$$\frac{\partial^2}{\partial x^2} f(x, y) = 8 + 4y \quad \frac{\partial^2}{\partial x \partial y} f(x, y) = 1 + 4x \quad \text{and} \quad \frac{\partial^2}{\partial y^2} f(x, y) = 16 .$$

Evaluating at $\mathbf{x}_0 = (-1, 0)$, we find $[\text{Hess}_f(\mathbf{x}_0)] = \begin{bmatrix} 8 & -3 \\ -3 & 16 \end{bmatrix}$.

The quadratic equation we must solve to find the eigenvalues is $(8-t)(16-t) = 9$, which reduces to $t^2 - 24t = -119$. Completing the square, $(t-12)^2 = 25$, and so the two eigenvalues are

$$\mu_1 = 17 \quad \text{and} \quad \mu_2 = 7 .$$

We then get \mathbf{u}_1 by normalizing $(3, 8-17) = 3(1, -3)$ which yields $\mathbf{u}_1 = \frac{1}{\sqrt{10}}(1, -3)$. Once we have \mathbf{u}_1 , we obtain \mathbf{u}_2 from $\mathbf{u}_2 = (\mathbf{u}_1)^\perp$: $\mathbf{u}_2 = \frac{1}{\sqrt{10}}(3, 1)$.

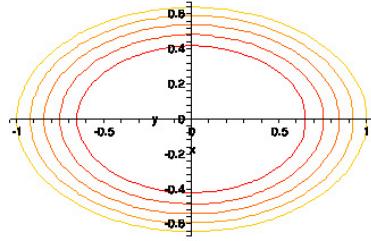
In the corresponding coordinate system, the best quadratic approximation to f at \mathbf{x}_0 is

$$f((-1, 0) + u\mathbf{u}_1 + v\mathbf{u}_2) \approx -4 + \frac{1}{2}(17u^2 + 7v^2) .$$

Let us make a contour plot of the right hand side in the u, v plane. Setting

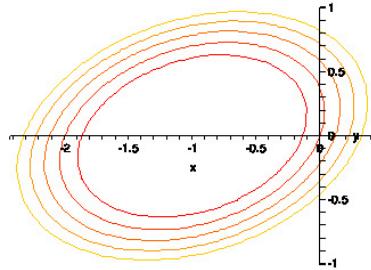
$$17u^2 + 7v^2 = c ,$$

we get the equation of an ellipse whose major axis is the v -axis, and whose minor axis is the u -axis: Here is a plot in the u, v plane for 5 values of c :

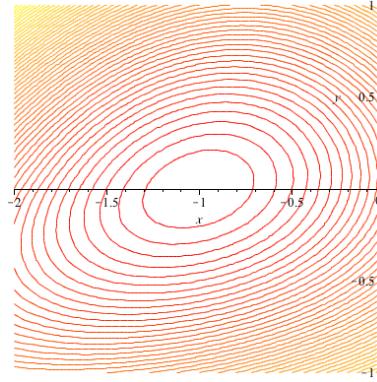


Now let us shift and rotate this plot so it fits into the original x, y plane. The point \mathbf{x}_0 is the center of the u, v coordinate system, and the positive u axis runs along the \mathbf{u}_1 direction and the positive v -axis runs along the \mathbf{u}_2 direction.

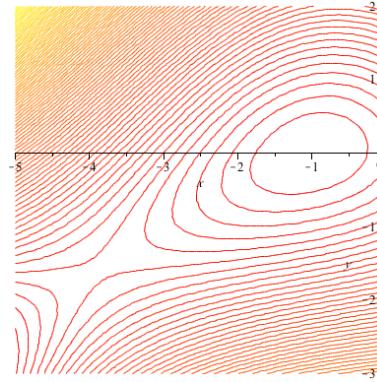
So shifting and rotating the ellipses accordingly, we get our plot:



Here is a contour plot of the actual function $f(x, y)$ over the range $-2 \leq x \leq 0$, $-1 \leq y \leq 1$.



As you can see, the contour plot corresponds quite closely to the contour plot for the quadratic approximation over this whole region, which is not even so small. If we look at a larger region, we see that the approximation breaks down farther away from $(-1, 0)$:



Indeed, you see there is another critical point near $(-4, -3/2)$, and that the contour lines near this critical point are hyperbolas. What is impressive however, is the size of the region around the critical point where there contour plots of f and its quadratic approximation are virtually indistinguishable.

Example 95 (Drawing a contour plot). Let $f(x, y) = 3x^2y - 6x - y^3$. Then

$$\nabla f(x, y) = (6xy - 6, 3x^2 - 3y^2).$$

We have a critical point if and only if $xy = 1$ and $x^2 = y^2$. The latter equation says $x = \pm y$. But if $x = -y$, the $xy = 1$ is impossible, so $x = y$, and then we must have $x = -1$ or $x = 1$. Hence the two critical points are $(-1, -1)$ and $(1, 1)$.

Let us first take $\mathbf{x}_0 = (1, 1)$. We compute:

$$\frac{\partial^2}{\partial x^2} f(x, y) = 6y \quad \frac{\partial^2}{\partial x \partial y} f(x, y) = 6x \quad \text{and} \quad \frac{\partial^2}{\partial y^2} f(x, y) = 6y.$$

Evaluating at $\mathbf{x}_0 := (1, 1)$, we find $[\text{Hess}_f(\mathbf{x}_0)] = \begin{bmatrix} 6 & 6 \\ 6 & -6 \end{bmatrix}$.

The quadratic equation we must solve to find the eigenvalues is $(6-t)(-6-t) = 36$. which reduces to $t^2 = 72$. The two roots are

$$\mu_1 = 6\sqrt{2} \quad \text{and} \quad \mu_2 = -6\sqrt{2}.$$

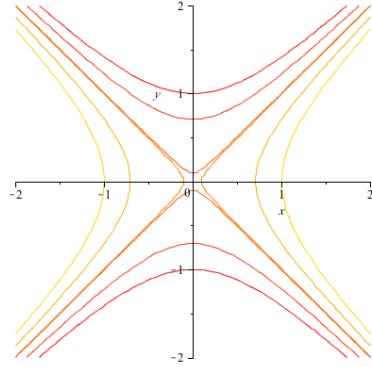
We then get \mathbf{u}_1 by normalizing $(6, 6 - 6\sqrt{2}) = 6(1, 1 - \sqrt{2})$ and then $\mathbf{u}_2 = \mathbf{u}_1^\perp$:

$$\mathbf{u}_1 = \frac{1}{4 - 2\sqrt{2}}(1, 1 - \sqrt{2}) \quad \text{and} \quad \mathbf{u}_2 = \frac{1}{4 - 2\sqrt{2}}(\sqrt{2} - 1, 1).$$

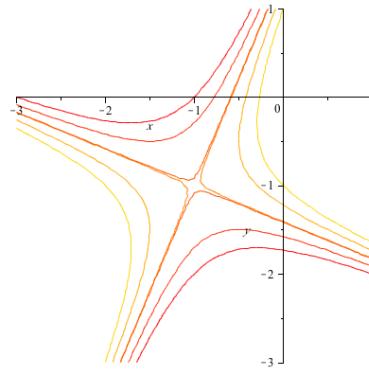
Since $f(\mathbf{x}_0) = -2$, the best quadratic approximation near \mathbf{x}_0 in the u, v coordinates is

$$f((1, 1) + u\mathbf{u}_1 + v\mathbf{u}_2) = -2 + 3\sqrt{2}u^2 - 3\sqrt{2}v^2.$$

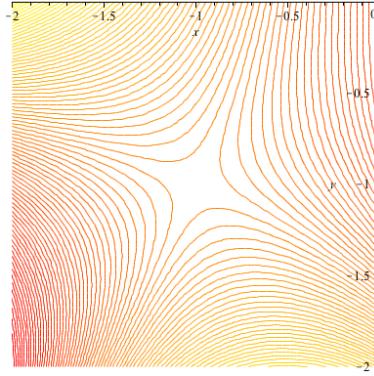
The equation obtained by setting the right hand side is equal to a constant is equivalent to the equation $\tilde{u}^2 - \tilde{v}^2 = c$, for some other constant c , which is the equation of an hyperbola. Here is a plot in the u, v plane for 6 values of c :



Now let us shift and rotate this plot so it fits into the original x, y plane. The center of the u, v coordinate system is at \mathbf{x}_0 , the positive u -axis runs along the \mathbf{u}_1 direction, and the positive v -axis runs along the \mathbf{u}_2 direction. Shifting and rotating the hyperbolas accordingly, we obtain:



To the extent that the quadratic approximation is valid, this should be a good match to the contour plot of the original function $f(x, y)$ near to the critical point $(-1, -1)$. For comparison, here is a computer generated contour plot of the actual function $f(x, y)$ over the range $-2 \leq x, y \leq 0$.



As you can see, the actual contour plot corresponds quite closely to the contour plot for the quadratic approximation over this whole region, which is not even so small.

6.2.6 Types of critical points for real valued functions on \mathbb{R}^n

The “second derivative test” in single variable calculus says that if f is a twice continuously differentiable function on an interval (a, b) , and for some $x_0 \in (a, b)$, $f'(x_0) = 0$, then x_0 is a local maximum of f in case $f''(x_0) < 0$, and is a local minimum of f in case $f''(x_0) > 0$. In multivariable calculus, it is the Hessian matrix that plays the role of f'' , but what is the relevant sense of positivity or negativity of an $n \times n$ matrix? As we shall see, it concerns the signs of the eigenvalues of the matrix.

Definition 73 (Positive definite matrices). *Let A be an $n \times n$ symmetric matrix. Then A is positive definite in case for every non-zero vector \mathbf{x} in \mathbb{R}^n ,*

$$\mathbf{x} \cdot A\mathbf{x} > 0 .$$

We say A is negative definite if $-A$ is positive definite. We say that A is positive semi-definite in case $\mathbf{x} \cdot A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and that A is negative semi-definite in case $-A$ is positive semi-definite.

Theorem 79 (Eigenvalues and matrix positivity). *A symmetric $n \times n$ matrix A is positive definite if and only if all of its eigenvalues are strictly positive, and in this case, with $c := \min_{j=1,\dots,n}\{\mu_j\}$*

$$\mathbf{x} \cdot A\mathbf{x} \geq c\|\mathbf{x}\|^2 . \quad (6.33)$$

Proof. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be an orthonormal basis of \mathbb{R}^n consisting of eigenvectors of A . Let $\{\mu_1, \dots, \mu_n\}$ be the corresponding eigenvalues. Since $\mu_j = \mathbf{u}_j \cdot A\mathbf{u}_j$, when A is positive definite, then each eigenvalue of A is strictly positive.

On the other hand, define $\mu_{\min} := \min_{j=1,\dots,n}\{\mu_j\}$, and suppose that $\mu_{\min} > 0$. We can express any $\mathbf{x} \in \mathbb{R}^n$ as $\mathbf{x} = \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) \mathbf{u}_j$, and then, $\mathbf{x} \cdot A\mathbf{x} = \sum_{j=1}^n \mu_j (\mathbf{x} \cdot \mathbf{u}_j)^2 \geq \mu_{\min} \sum_{j=1}^n (\mathbf{x} \cdot \mathbf{u}_j) = \mu_{\min} \|\mathbf{x}\|^2$. Thus, A is positive definite, and (6.33) is valid. \square

Now consider a real valued function f that is twice continuously differentiable on an open set U . Suppose that $\mathbf{x}_0 \in U$, and $\nabla f(\mathbf{x}_0) = 0$, so that \mathbf{x}_0 is a critical point. Let $A := [\text{Hess}_f(\mathbf{x}_0)]$, and suppose that A is positive definite. Then by Theorem 79, if μ_{\min} is defined to be the least eigenvalue of A , $\mu_{\min} > 0$, and for all $\mathbf{x} \in \mathbb{R}^n$, (6.33) is valid.

By Theorem 78, choosing $\epsilon = \mu_{\min}/4$, there is a $\delta > 0$ so that

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow \left| f(\mathbf{x}) - f(\mathbf{x}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) \right| < \frac{\mu_{\min}}{4} \|\mathbf{x} - \mathbf{x}_0\|^2.$$

In particular,

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow f(\mathbf{x}) &\geq f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) - \frac{\mu_{\min}}{4} \|\mathbf{x} - \mathbf{x}_0\|^2 \\ &\geq f(\mathbf{x}_0) + \frac{\mu_{\min}}{4} \|\mathbf{x} - \mathbf{x}_0\|^2, \end{aligned}$$

where we have used the fact that $(\mathbf{x} - \mathbf{x}_0) \cdot A(\mathbf{x} - \mathbf{x}_0) \geq \mu_{\min} \|\mathbf{x} - \mathbf{x}_0\|^2$ for all \mathbf{x} . That is, at all points \mathbf{x} such that $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, $f(\mathbf{x}) \geq f(\mathbf{x}_0)$ with equality only in case $\mathbf{x} = \mathbf{x}_0$. This brings us to the following definition:

Definition 74 (Local maximizers and minimizers). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. Then $\mathbf{x}_0 \in \mathbb{R}^n$ is a strict local maximizer of f in case there is some $r > 0$ so that*

$$0 < \|\mathbf{x} - \mathbf{x}_0\| < r \Rightarrow f(\mathbf{x}_0) > f(\mathbf{x}).$$

We say that \mathbf{x}_0 is a strict local minimizer of $-f$ in case there is some $r > 0$ so that

$$0 < \|\mathbf{x} - \mathbf{x}_0\| < r \Rightarrow f(\mathbf{x}_0) < f(\mathbf{x}).$$

Now notice that \mathbf{x}_0 is a strict local maximizer of f if and only if \mathbf{x}_0 is a strict local minimizer of $-f$. Hence the following theorem is an immediate consequence of the discussion that precedes Definition 74

Theorem 80 (Criteria for local maxima and minima). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose that \mathbf{x}_0 is a critical point of f , and that for some $R > 0$, all of the second order partial derivatives of f exists and are continuous at each \mathbf{z} with $\|\mathbf{z} - \mathbf{x}_0\| \leq R$. Then:*

- (1) If $[\text{Hess}_f(\mathbf{x}_0)]$ is negative definite, \mathbf{x}_0 is a strict local maximizer of f .
- (2) If $[\text{Hess}_f(\mathbf{x}_0)]$ positive definite, \mathbf{x}_0 is a strict local minimizer of f .

6.2.7 Sylvester's Criterion

Suppose \mathbf{x}_0 is a critical point of a twice continuously differentiable function \mathbf{x}_0 . If we want to apply Theorem 80 to determine whether or not \mathbf{x}_0 might be a strict local minimizer or maximizer, one way to proceed would be to compute all of the eigenvalues of $[\text{Hess}_f(\mathbf{x}_0)]$. With 3 or more variables, it is often difficult to do this. Fortunately, there is another way: One can determine whether all of the eigenvalues are strictly positive simply by computing certain determinants. We now explain how this works for $n = 2$ and $n = 3$, in which case we have already defined the determinant.

Recall from (6.9) that if A is a 2×2 symmetric matrix and $\{\mathbf{u}_1, \mathbf{u}_2\}$ is an orthonormal basis of \mathbb{R}^2 such that $A\mathbf{u}_j = \mu_j \mathbf{u}_j$, $j = 1, 2$, then $\det(A) = \mu_1 \mu_2$. Therefore, $\det(A) > 0$ if and only if either both eigenvalues of A are strictly positive, or both eigenvalues of A are strictly negative, and $\det(A) < 0$ if and only if one eigenvalue of A is strictly positive, and the other is strictly negative. Hence if $\det(A) > 0$, A is either positive definite or negative definite. Since $A_{1,1} = \mathbf{e}_1 \cdot A \mathbf{e}_1$, A will

then be positive definite in case $A_{1,1} > 0$ and negative definite in case $A_{1,1} < 0$. (You could also use $A_{2,2}$ for the same purpose.) We summarize:

(1) Both eigenvalues are strictly positive if and only if $\det(A) > 0$ and $A_{1,1} > 0$.

(2) Both eigenvalues are strictly negative if and only if $\det(A) > 0$ and $A_{1,1} < 0$.

(3) If $\det(A) < 0$, one eigenvalue is strictly positive, and the other is strictly negative.

This is the 2×2 version of *Sylvester's Criterion*. The 3×3 case is only slightly more complicated.

Let

$$A = \begin{bmatrix} a & u & v \\ u & b & w \\ v & w & c \end{bmatrix} \quad \text{and} \quad B := \begin{bmatrix} a & u \\ u & b \end{bmatrix} .$$

Notice that B is the 2×2 block in the upper left corner of A .

The 3×3 version of Sylvester's Criterion states that

(1) All three eigenvalues are strictly positive if and only if $\det(A) > 0$, $\det(B) > 0$ and $A_{1,1} > 0$.

(2) All three eigenvalues are strictly negative if and only if $\det(A) < 0$, $\det(B) > 0$ and $A_{1,1} < 0$.

(3) Suppose none of $\det(A) \neq 0$ is zero, but neither of the conditions in (1) and (2) is satisfied. Then at least one eigenvalue is strictly positive, and at least one eigenvalue is strictly negative.

To see this, let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be an orthonormal basis of \mathbb{R}^3 such that $A\mathbf{u}_j = \mu_j \mathbf{u}_j$, $j = 1, 2, 3$. Then by (6.8), $\det(A) = \mu_1 \mu_2 \mu_3$. If $\det(A) > 0$, then there are two possibilities: Either each eigenvalue is strictly positive, or else two are strictly negative, and one is strictly positive. We now show that when two eigenvalues of A are strictly negative, then B cannot be positive definite.

Suppose that μ_1 and μ_2 are strictly negative and μ_3 is strictly positive. Consider the plane parameterized by $\mathbf{x}(s, t) = s\mathbf{u}_1 + t\mathbf{u}_2$. Since every pair of planes that pass through the origin in \mathbb{R}^3 intersect at least in a line, There is a unit vector in this parameterized plane that also lies in the x, y plane. That is, for some s and t with $s^2 + t^2 = 1$, and some x and y with $x^2 + y^2 = 1$,

$$s\mathbf{u}_1 + t\mathbf{u}_2 = (x, y, 0) .$$

We then compute

$$(\mathbf{s}\mathbf{u}_1 + \mathbf{t}\mathbf{u}_2) \cdot A(\mathbf{s}\mathbf{u}_1 + \mathbf{t}\mathbf{u}_2) = \mu_1 s^2 + \mu_2 t^2 < 0 ,$$

and

$$(x, y, 0) \cdot A(x, y, 0) = (x, y) \cdot B(x, y) . \quad (6.34)$$

We conclude that there is a unit vector (x, y) such that $(x, y) \cdot B(x, y) < 0$, and hence B is not positive definite, nor even positive-semidefinite. However, by the 2×2 version of Sylvester's Criterion, if $\det(B) > 0$ and $B_{1,1} > 0$, B is positive definite. Since $A_{1,1} = B_{1,1}$, the condition that $\det(B) > 0$ and $A_{1,1} < 0$ is incompatible with A having two negative eigenvalues so all of the eigenvalues of A must be strictly positive when $\det(A) > 0$, $\det(B) > 0$ and $A_{1,1} > 0$.

This proves that $\det(A) > 0$, $\det(B) > 0$ and $A_{1,1} > 0$ is a necessary and sufficient condition for A to be positive definite. Then $\det(-A) > 0$, $\det(-B) > 0$ and $-A_{1,1} > 0$ is a necessary and sufficient condition for $-A$ to be positive definite, or what is the same thing, $\det(A) < 0$, $\det(B) > 0$ and $A_{1,1} < 0$ is a necessary and sufficient condition for A to be negative definite.

If $\det(A) \neq 0$, then $\mu_1\mu_2\mu_3 \neq 0$, and hence none of the eigenvalues is zero. If they were all positive, condition (1) would be satisfied. If they were all negative, condition (2) would be satisfied. Hence both signs must be present.

There is an $n \times n$ version of Sylvester's Criterion, whose form you can probably guess. However, since have not introduced higher dimensional determinant in this course, we stop here.

6.3 Curvature of surfaces in \mathbb{R}^3

6.3.1 Parameterized surfaces in \mathbb{R}^3

Definition 75 (Parameterized surface). A parameterized surface \mathcal{S} in \mathbb{R}^3 is a continuously differentiable function

$$\mathbf{X}(u, v) = (x(u, v), y(u, v), z(u, v))$$

from an open set U in \mathbb{R}^2 to \mathbb{R}^3 such that the columns of $[D_{\mathbf{X}}(u, v)]$ are linearly independent for each $(u, v) \in U$. We also require that for distinct (u_1, v_1) and (u_2, v_2) in U , $\mathbf{X}(u_1, v_1) \neq \mathbf{X}(u_2, v_2)$ so that the function sending (u, v) to $\mathbf{X}(u, v)$ is an invertible transformation from U to \mathcal{S} . The inverse function, which sends a point $\mathbf{p} \in \mathcal{S}$ to its coordinate vector $(u(\mathbf{p}), v(\mathbf{p})) \in U$, is called the coordinate function of the parameterization.

Differentiable surfaces in \mathbb{R}^3 are the two dimensional analogs of differentiable curves $\mathbf{x}(t)$ in \mathbb{R}^3 . The requirement that the columns of the Jacobian matrix be linearly independent ensures that the surface \mathcal{S} is really two dimensional, and not a one dimensional object, like a curve, in disguise. For example, let $\mathbf{x}(t)$ be any continuously differentiable function from the interval $(0, 1)$ to \mathbb{R}^3 . Define

$$\mathbf{X}(u, v) := \mathbf{x}(uv)$$

on $U = (0, 1) \times (0, 1) \subset \mathbb{R}^2$. The image of this function is not a surface, but a curve. Moreover, one computes in this case that

$$[D_{\mathbf{X}}(u, v)] = [v\mathbf{x}'(uv), u\mathbf{x}'(uv)] ,$$

so that the two columns of $[D_{\mathbf{X}}(u, v)]$ are multiples of one another.

Since we shall be doing a number of computations with the 3×2 Jacobian matrix $[D_{\mathbf{X}}(u, v)]$, let us pause to become familiar with it and introduce some useful notation. Since $\mathbf{X}(u, v) = (x(u, v), y(u, v), z(u, v))$, we have that

$$[D_{\mathbf{X}}(u, v)] = \begin{bmatrix} \nabla x(u, v) \\ \nabla y(u, v) \\ \nabla z(u, v) \end{bmatrix} = \begin{bmatrix} \partial x(u, v)/\partial u & \partial x(u, v)/\partial v \\ \partial y(u, v)/\partial u & \partial y(u, v)/\partial v \\ \partial z(u, v)/\partial u & \partial z(u, v)/\partial v \end{bmatrix} .$$

The two columns of $[D_{\mathbf{X}}(u, v)]$ are the two vectors $\frac{\partial \mathbf{X}}{\partial u}(u, v)$ and $\frac{\partial \mathbf{X}}{\partial v}(u, v)$, which come up frequently in what follows, and it will be convenient to have a more compact notation for them. We define

$$\mathbf{X}_u(u, v) = \frac{\partial \mathbf{X}}{\partial u}(u, v) \quad \text{and} \quad \mathbf{X}_v(u, v) = \frac{\partial \mathbf{X}}{\partial v}(u, v) . \quad (6.35)$$

Then we have the *column representation* for the Jacobian matrix $[D_{\mathbf{X}}(u, v)]$:

$$[D_{\mathbf{X}}(u, v)] = [\mathbf{X}_u(u, v), \mathbf{X}_v(u, v)] . \quad (6.36)$$

One example of a parameterized surface is provided by the graph of a continuously differentiable function f on an open set $U \subset \mathbb{R}^2$. Then we take $u = x$, $v = y$ and $z = f(x, y) = f(u, v)$, and we have $\mathbf{X}(u, v) = (u, v, f(u, v))$. Then

$$[D_{\mathbf{X}}(u, v)] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\partial f(u, v)}{\partial u} & \frac{\partial f(u, v)}{\partial v} \end{bmatrix} .$$

In this case the columns are clearly linearly independent no matter what f is.

Not all parameterized surfaces are graphs: As we have seen early on in Chapter One, (see also Example 97 below), the unit sphere in \mathbb{R}^3 can be parameterized, except for the poles, by

$$\mathbf{X}(u, v) = (\sin u \cos v, \sin u \sin v, \cos u)$$

for $-\pi < u < \pi$ and $0 < v < \pi$. The Jacobian matrix is given by

$$[D_{\mathbf{X}}(u, v)] = \begin{bmatrix} \cos u \cos v & -\sin u \cos v \\ \cos u \sin v & \sin u \sin v \\ -\sin u & 0 \end{bmatrix} .$$

Since $\sin v > 0$ for $0 < v < \pi$, the two columns of $[D_{\mathbf{X}}(u, v)]$ are never multiples of one another. Thus the parameterization of the unit sphere that we met in Chapter One is, as we would expect, a parameterized surface.

Now let \mathcal{S} be parameterized surface given by the function $\mathbf{X}(u, v)$ defined on an open set $U \subset \mathbb{R}^2$ with values in \mathbb{R}^3 , and let $(u_0, v_0) \in U$. Let \mathbf{p} denote the point $\mathbf{X}(u_0, v_0)$.

Let $\mathbf{y}(t) = (u(t), v(t))$ be a continuously differentiable curve in U defined for $t \in (-a, a)$, some $a > 0$, and such that $\mathbf{y}(0) = (u_0, v_0)$. Then the composite function $\mathbf{x}(t) = \mathbf{X}(u(t), v(t))$ is a curve in \mathbb{R}^3 such that $\mathbf{x}(0) = \mathbf{p}$, and such that $\mathbf{x}(t) \in \mathcal{S}$ for all $t \in (-a, a)$. The tangent vector to this curve is tangent to \mathcal{S} at (u_0, v_0) in a natural sense. By the chain rule, we find, using (6.36),

$$\mathbf{x}'(0) = [D_{\mathbf{X}}(u_0, v_0)](u'(0), v'(0)) = u'(0)\mathbf{X}_u(u_0, v_0) + v'(0)\mathbf{X}_v(u_0, v_0) .$$

Every linear combination of the columns of $[D_{\mathbf{X}}(u_0, v_0)]$ is a tangent vector: To see this, fix $a, b \in \mathbb{R}$, and consider the curve $\mathbf{x}(t) = \mathbf{X}(u_0 + at, v_0 + bt)$. Then $\mathbf{x}'(0) = a\mathbf{X}_u(u_0, v_0) + b\mathbf{X}_v(u_0, v_0)$, showing that the vector on the right is tangent to \mathcal{S} at \mathbf{p} .

Next consider the vector

$$\mathbf{X}_u(u_0, v_0) \times \mathbf{X}_v(u_0, v_0) .$$

The two vectors in the cross product are linearly independent, the cross product is not zero. By the orthogonality property of the cross product, it is orthogonal to any linear combination of $\mathbf{X}_u(u_0, v_0)$ and $\mathbf{X}_v(u_0, v_0)$. Hence it is orthogonal to every vector that this tangent to \mathcal{S} at \mathbf{p} . Normalizing the cross product, which we can do since it is not zero, we obtain a unit vector known as the *unit normal to the surface \mathcal{S}* .

Definition 76 (Unit normal and tangent plane). *Let \mathcal{S} be a parameterized surface. The unit normal vector $\widehat{\mathbf{N}}(u, v)$ is given by*

$$\widehat{\mathbf{N}}(u, v) = \frac{1}{\|\mathbf{n}(u, v)\|} \mathbf{n}(u, v) , \quad (6.37)$$

where

$$\mathbf{n}(u, v) = \mathbf{X}_u \times \mathbf{X}_v(u, v) . \quad (6.38)$$

(We put the hat on this unit vector to distinguish it from the unit normal vector to the curve, which we shall denote by $\mathbf{N}(t)$ as usual.) The tangent plane to \mathcal{S} at a point $\mathbf{X}(u_0, v_0)$ in \mathcal{S} is the plane consisting of all $\mathbf{x} \in \mathbb{R}^3$ that satisfy

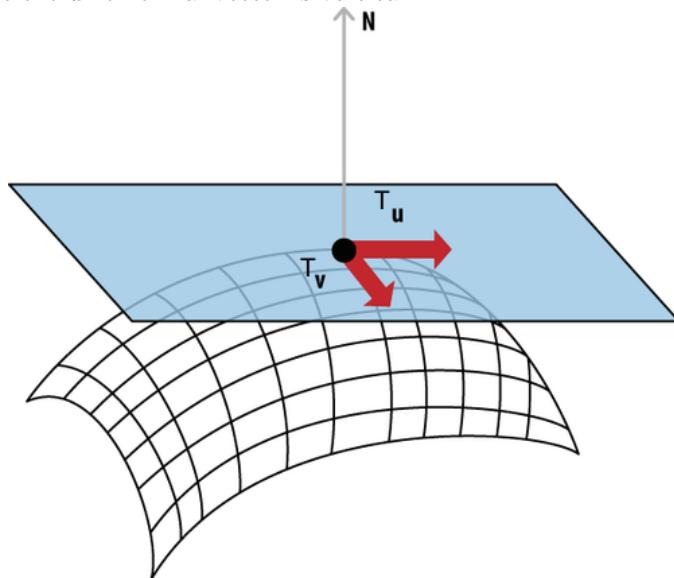
$$\widehat{\mathbf{N}}(u_0, v_0) \cdot (\mathbf{x} - \mathbf{X}(u_0, v_0)) = 0 .$$

The tangent space to \mathcal{S} at $\mathbf{X}(u_0, v_0)$ is the span of the vectors $\{\mathbf{X}_u(u_0, v_0), \mathbf{X}_v(u_0, v_0)\}$.

Notice that if we changed the order of u and v , the sign of $\widehat{\mathbf{N}}(u_0, v_0)$ would change. The side of the surface towards which $\widehat{\mathbf{N}}(u_0, v_0)$ points is often called the “positive” side of the surface. Surfaces that can be parameterized in the simple manner described here always have two sides. Later we shall see that there are sided surfaces, such as a Möbius band, cannot be parameterized in this manner. (We shall discuss such surfaces in Chapter 9). The choice of one side or the other as “the positive side” is known as a choice of “orientation” of the surface.

Also note the distinction between the *tangent plane* and the *tangent space*. The tangent plane to \mathcal{S} at a point $\mathbf{p} \in \mathcal{S}$ is a plane passing through \mathbf{p} that need not contain $\mathbf{0}$. The tangent space \mathcal{S} at \mathbf{p} is however a subspace of \mathbb{R}^3 , which like all subspaces of a vector space, contains $\mathbf{0}$. As a subset of \mathbb{R}^3 , the tangent space is the plane through $\mathbf{0}$ that is parallel to the tangent plane.

Here is a picture showing a surface \mathcal{S} showing its tangent plane \mathbf{p} at a point. Also in the diagram you see a number of coordinate curves on the surface, the tangent vectors to the coordinate curves crossing at \mathbf{p} , and the unit normal to \mathbf{p} . The perspective is such that the tangent plane appears horizontal, and hence the unit normal vector is vertical.



Then for fixed u_0 and v_0 , the functions sending u to $\mathbf{X}(u, v_0)$ and v to $\mathbf{X}(u_0, v)$ are *coordinate curves* on the surface. A number of these are shown in the figure above, for various values of u_0 and v_0 to produce a “coordinate grid” on the surface. The vectors \mathbf{X}_u and \mathbf{X}_v are tangent vectors to the coordinate curves.

Example 96 (Tangent planes for graphs). *Let $f(x, y)$ be a continuously differentiable function on \mathbb{R}^2 . Consider the parameterized surface S given by $\mathbf{X}(u, v) = (u, v, f(u, v))$. As we have seen above,*

$$[D\mathbf{X}(u, v)] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\partial f(u, v)}{\partial u} & \frac{\partial f(u, v)}{\partial v} \end{bmatrix}.$$

Therefore,

$$\mathbf{n}(u, v) = \left(-\frac{\partial f(u, v)}{\partial u}, -\frac{\partial f(u, v)}{\partial v}, 1 \right).$$

Defining $\mathbf{X}_0 = \mathbf{X}(u_0, v_0)$, we have that $\mathbf{n}(u_0, v_0) \cdot (\mathbf{x} - \mathbf{X}_0) = 0$ can be written as

$$\left(\frac{\partial f(u, v)}{\partial u}, \frac{\partial f(u, v)}{\partial v}, -1 \right) \cdot (\mathbf{x} - \mathbf{X}_0) = 0,$$

which is the equation of the tangent plane to the graph of $z = f(x, y)$ at $(x, y) = (u_0, v_0)$.

Hence the new notion of “tangent plane” agrees with the old one for surfaces that are graphs, as it should.

Example 97 (Two parameterizations of the unit sphere). *There are many ways to parameterize S^2 , the unit sphere in \mathbb{R}^3 . The first is the “latitude and longitude” parameterization: Given \mathbf{p} in S^2 , define $\cos u = \mathbf{p} \cdot \mathbf{e}_3$ so that $0 \leq u \leq \pi$. Then u is essentially the “latitude” except it is measured from the North Pole, and not the Equator. Let $\mathbf{p}_\perp = \mathbf{p} - (\mathbf{p} \cdot \mathbf{e}_3)\mathbf{e}_3$, so that $\|\mathbf{p}_\perp\| = \sin u$, and define v by*

$$\mathbf{p}_\perp = \|\mathbf{p}_\perp\|(\cos v, \sin v, 0) = (\sin u \cos v, \sin u \sin v, 0).$$

Reassembling the components of the general point \mathbf{p} , we have our first parameterization:

$$\mathbf{X}(u, v) = (\sin u \cos v, \sin u \sin v, \cos u). \quad (6.39)$$

Here, we take $0 < u < \pi$ and $0 < v < 2\pi$. This parameterization does not cover the whole sphere; it leaves out the half-circle running from the North Pole to the South Pole along the coordinate curve with $v = 0$ (or, equivalently, $v = 2\pi$). However, we require the parameterization to be defined on an open set. We see something even more significant when we compute the coordinate tangent vectors \mathbf{X}_u and \mathbf{X}_v :

$$\mathbf{X}_u = (\cos u \cos v, \cos u \sin v, -\sin u) \quad \text{and} \quad \mathbf{X}_v = (-\sin u \sin v, \sin u \cos v, 0).$$

Note that $\mathbf{X}(0, v) = \mathbf{X}(\pi, v) = \mathbf{0}$ for all v , so that at $u = 0$ and at $u = \pi$, $\{\mathbf{X}_u(u, v), \mathbf{X}_v(u, v)\}$ are not linearly independent, as we require.

Next we compute $\mathbf{n} = \mathbf{X}_u \times \mathbf{X}_v$ and find $\mathbf{n}(u, v) = \sin u (\sin u \cos v, \sin u \sin v, \cos u)$. Normalizing, $\|\mathbf{n}(u, v)\| = \sin u$ and so

$$\widehat{\mathbf{N}}(u, v) = (\sin u \cos v, \sin u \sin v, \cos u),$$

which you will notice says that $\hat{\mathbf{N}}(u, v) = \mathbf{X}(u, v)$, which is geometrically obvious for the unit sphere.

Consider $u_0 = v_0 = \pi/4$. Then $\mathbf{X}(\pi/4, \pi/4) = \hat{\mathbf{N}}(\pi/4, \pi/4) = (1/2, 1/2, 1/\sqrt{2})$. Hence the equation for the tangent plane at $\mathbf{p} = \mathbf{X}(\pi/4, \pi/4)$ is $(1/2, 1/2, 1/\sqrt{2}) \cdot (1/2 - x, 1/2 - y, 1/\sqrt{2} - z) = 0$ which simplifies to

$$x + y + \sqrt{2}z = 2 .$$

There is another nice way to parameterize the sphere using the stereographic projection. Let \mathbf{p} be a point in S^2 other than the South Pole $-\mathbf{e}_3$. Then there is a uniquely determined line through $-\mathbf{e}_3$ and \mathbf{p} that intersects the x, y plane in a uniquely determined point $\mathbf{x}(\mathbf{p}) \in \mathbb{R}^2$. The line is parameterized by $\mathbf{x}(t) = -\mathbf{e}_3 + t(\mathbf{p} + \mathbf{e}_3)$ and it intersects the x, y plane exactly when $\mathbf{e}_3 \cdot \mathbf{x}(t) = 0$. Solving for t we find $0 = -1 + t(\mathbf{p} \cdot \mathbf{e}_3 + 1)$, so that $t = 1/(\mathbf{p} \cdot \mathbf{e}_3 + 1)$. Substituting this t into $\mathbf{x}(t)$ we find

$$\mathbf{x}(\mathbf{p}) = \frac{1}{1 + p_3} (p_1, p_2)$$

where $\mathbf{p} = (p_1, p_2, p_3)$. The functions $u(\mathbf{p}) = p_1/(1 + p_3)$ and $v(\mathbf{p}) = p_2/(1 + p_3)$ are the stereographic coordinate functions on S^2 . That is, we write $\mathbf{x}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$.

To get the corresponding parameterization, we invert to express \mathbf{p} as a function of u and v . Note that

$$u^2 + v^2 = u^2(\mathbf{p}) + v^2(\mathbf{p}) = \frac{p_1^2 + p_2^2}{(1 + p_3)^2} = \frac{1 - p_3^2}{(1 + p_3)^2} = \frac{1 - p_3}{1 + p_3} .$$

Solving for p_3 , we find $(u^2 + v^2)(1 + p_3) = (1 - p_3)$ so that

$$p_3 = \frac{1 - u^2 - v^2}{1 + u^2 + v^2} .$$

Then $p_1 = u(1 + p_3) = 2u/(1 + u^2 + v^2)$ and likewise, $p_2 = v(1 + p_3) = 2v/(1 + u^2 + v^2)$. The expresses the point \mathbf{p} as a function of u and v , and we have our second parameterization:

$$\tilde{\mathbf{X}}(u, v) = \frac{1}{1 + u^2 + v^2} (2u, 2v, 1 - u^2 - v^2) . \quad (6.40)$$

You should check that the right hand side is, in fact, a unit vector.

This parameterization covers the wholes of S^2 except for the South Pole. Interchanging the roles of the North and South Poles, one get another parameterization that covers everything except the North Pole. Using the two together, one has the whole sphere covered.

We now compute the coordinate tangent vectors for this parameterization:

$$\begin{aligned} \tilde{\mathbf{X}}_u(u, v) &= \frac{-2}{(1 + u^2 + v^2)^2} (u^2 - v^2 - 1, 2uv, 2u) \\ \tilde{\mathbf{X}}_v(u, v) &= \frac{-2}{(1 + u^2 + v^2)^2} (2uv, u^2 - v^2 + 1, 2v) . \end{aligned}$$

Taking the cross product, we find

$$\mathbf{n}(u, v) = \frac{4}{(1 + u^2 + v^2)^3} (2u, 2v, 1 - u^2 - v^2) .$$

Once more, we see that the unit normal vector at $\tilde{\mathbf{X}}(u, v)$ is $\tilde{\mathbf{X}}(u, v)$ itself. We can of course compute the tangent plane at $\mathbf{p} = (1/2, 1/2, 1/\sqrt{2})$, and of course we get the same equations for it.

6.3.2 The arclength of curves on a parameterized surface

Consider the unit sphere S^2 with the “latitude and longitude” parameterization, (6.39). Let $\mathbf{p} = (1/2, 1/2, 1/\sqrt{2}) \in S^2$, and $\mathbf{q} = (0, 1, 0) \in S^2$. The coordinate of \mathbf{p} are $u(\mathbf{p}) = v(\mathbf{p}) = \pi/4$, while the coordinates of \mathbf{q} are $u(\mathbf{q}) = v(\mathbf{q}) = \pi/2$.

Consider the curve $\mathbf{u}(t)$ in the coordinate plane given by

$$\mathbf{u}(t) = (u(t), v(t)) = (1+t)(\pi/4, \pi/4).$$

Think of this curve as specifying the latitude and longitude at time t , as a point moves across the surface of the sphere from \mathbf{p} to \mathbf{q} between times $t = 0$ and $t = 1$.

Corresponding to this curve in the coordinate plane is the curve $\mathbf{x}(t) = \mathbf{X}(\mathbf{u}(t))$ on the sphere. We can compute the arclength of this geometric, or physical, curve in \mathbb{R}^3 (which happens to “live” on $S^2 \subset \mathbb{R}^3$) in terms of the coordinate curve in \mathbb{R}^2 . Here us how:

We first compute the velocity vector $\mathbf{x}'(t)$. By the chain rule,

$$\mathbf{x}'(t) = [D_{\mathbf{X}}(\mathbf{u}(t))] \mathbf{u}'(t) = \mathbf{X}_u(\mathbf{u}(t)) u'(t) + \mathbf{X}_v(\mathbf{u}(t)) v'(t).$$

Therefore, by Theorem 58, which says in particular that for all $m \times n$ matrices A and all $\mathbf{u} \in \mathbb{R}^n$, $A\mathbf{u} \cdot A\mathbf{u} = \mathbf{u} \cdot A^T A\mathbf{u}$. Applying this with $n = 2$ and $m = 3$, we have

$$\begin{aligned} \|\mathbf{x}'(t)\|^2 &= [D_{\mathbf{X}}(\mathbf{u}(t))] \mathbf{u}'(t) \cdot [D_{\mathbf{X}}(\mathbf{u}(t))] \mathbf{u}'(t) \\ &= \mathbf{u}'(t) \cdot ([D_{\mathbf{X}}(\mathbf{u}(t))]^T [D_{\mathbf{X}}(\mathbf{u}(t))] \mathbf{u}'(t)). \end{aligned} \quad (6.41)$$

Definition 77 (First Fundamental Matrix). *Let $\mathbf{X}(u, v)$ be a continuously differentiable parameterization of a surface \mathcal{S} in \mathbb{R}^3 . Then the First Fundamental Matrix of the parameterized surface at the point $\mathbf{p} = \mathbf{X}(u_0, v_0)$ is the 2×2 matrix*

$$[I]_p := [D_{\mathbf{X}}(u_0, v_0)]^T [D_{\mathbf{X}}(u_0, v_0)] \quad (6.42)$$

The notation is traditional, and will be used here, despite the risk that $[I]_{(u,v)}$ might suggest the identity matrix. By definition, if $(u(t), v(t))$ is a continuously differentiable curve in the u, v plane passing through (u_0, v_0) at $t = 0$, and if $\mathbf{x}(t) = \mathbf{X}(u(t), v(t))$, then the first fundamental matrix relates speed in physical 3 dimensional space, $\|\mathbf{x}'(0)\|$ to the velocity $\mathbf{u}'(0) := (u'(0), v'(0))$ in parameter space through

$$\|\mathbf{x}'(0)\|^2 = \mathbf{u}'(0) \cdot [I]_p \mathbf{u}'(0). \quad (6.43)$$

Integrating the speed, we can compute arc-length of curves on the surface in \mathbb{R}^3 doing computations with curves in the two-dimensional parameter space.

Since $[D_{\mathbf{X}}(u, v)] = [\mathbf{X}_u(u, v), \mathbf{X}_v(u, v)]$,

$$[I]_p = [\mathbf{X}_u(u_0, v_0), \mathbf{X}_v(u_0, v_0)]^T \cdot [\mathbf{X}_u(u_0, v_0), \mathbf{X}_v(u_0, v_0)] = \begin{bmatrix} \mathbf{X}_u \cdot \mathbf{X}_u(u_0, v_0) & \mathbf{X}_v \cdot \mathbf{X}_u(u_0, v_0) \\ \mathbf{X}_u \cdot \mathbf{X}_v(u_0, v_0) & \mathbf{X}_v \cdot \mathbf{X}_v(u_0, v_0) \end{bmatrix}.$$

To write $[I]_p$ more conveniently, define the three functions E , F and G by

$$E(u, v) = \|\mathbf{X}_u(u, v)\|^2, \quad F(u, v) = \mathbf{X}_u(u, v) \cdot \mathbf{X}_v(u, v) \quad \text{and} \quad G(u, v) = \|\mathbf{X}_v(u, v)\|^2. \quad (6.44)$$

Once the functions E , F and G are known, one can compute the speed of the three dimensional curve $\mathbf{x}'(t)$ entirely in terms of the two dimensional coordinate curve $\mathbf{u}(t)$, and then the arc length of curves on the surface in \mathbb{R}^3 in terms of their 2-variable parametric description.

Example 98 (Computing arclength using coordinates). *Consider the parameterization of S^2 in terms of latitude and longitude (6.39) parameterization. To compute the first fundamental matrix for this parameterization, we compute the dot specified products to find*

$$E(u, v) = 1 \quad F(u, v) = 0 \quad \text{and} \quad G(u, v) = \sin^2(u, v) .$$

Hence

$$[I]_{(u,v)} = \begin{bmatrix} 1 & 0 \\ 0 & \sin^2 u \end{bmatrix} .$$

Now consider the coordinate curve $\mathbf{u}(t) = (u(t), v(t)) = (1+t)(\pi/4, \pi/4)$. and the corresponding curve $\mathbf{x}(t) = \mathbf{X}(\mathbf{u}(t))$ on S^2 . We will now calculate the arclength of this curve. We have from (??) that and the fact that $\mathbf{u}(t) = (\pi/4, \pi/4)$ for all t ,

$$\begin{aligned} \|\mathbf{x}'(t)\|^2 &= (\pi/4, \pi/4) \cdot \begin{bmatrix} 1 & 0 \\ 0 & \sin^2((1+t)\pi/4) \end{bmatrix} (\pi/4, \pi/4) \\ &= \frac{\pi^2}{16} (1 + \sin^2((1+t)\pi/4)) . \end{aligned}$$

Taking the square root and integrating, we find that the arclength of the path on the sphere is

$$\frac{\pi}{4} \int_0^1 \sqrt{1 + \sin^2((1+t)\pi/4)} dt .$$

This integral can be expressed in terms of a special function known as an elliptic integral of the second kind. The numerical value is 1.058095501.... We have seen in Chapter 2 that the shortest path on the sphere connecting \mathbf{p} and \mathbf{q} the shorter of the two great circle segments joining them, and that this path has length $\arccos(\mathbf{p} \cdot \mathbf{q})$. applying this with $\mathbf{p} = \mathbf{u}(0) = (1/2, 1/2, 1/\sqrt{2})$ and $\mathbf{q} = \mathbf{u}(1) = (0, 1, 0)$, we find that the length of the shortest path is $\arccos(1/2) = \pi/3 = 1.047197551....$ Hence linear interpolation of the coordinates of \mathbf{p} and \mathbf{q} , which is what the path $\mathbf{u}(t)$ is, does not provide that shortest path on the sphere from \mathbf{p} to \mathbf{q} .

Example 99 (First Fundamental Matrix for a graph). *Let $f(x, y)$ be a continuously differentiable function on \mathbb{R}^2 . Consider the parameterized surface \mathcal{S} given by*

$$\mathbf{X}(u, v) = (u, v, f(u, v)) .$$

Then $\mathbf{X}_u(u, v) = \left(1, 0, \frac{\partial f(u, v)}{\partial u}\right)$ and $\mathbf{X}_v(u, v) = \left(0, 1, \frac{\partial f(u, v)}{\partial v}\right)$, and therefore,

$$E(u, v) = 1 + \left(\frac{\partial f(u, v)}{\partial u}\right)^2 , \quad F(u, v) = \left(\frac{\partial f(u, v)}{\partial u}\right) \left(\frac{\partial f(u, v)}{\partial v}\right) \quad \text{and} \quad G(u, v) = 1 + \left(\frac{\partial f(u, v)}{\partial v}\right)^2 .$$

For example, for $f(x, y) = x^2 + y^2$, the First Fundamental Matrix is $\begin{bmatrix} 1 + 4u^2 & 4uv \\ 4uv & 1 + 4v^2 \end{bmatrix}$.

Note also that if (u_0, v_0) is a critical point of f , so that $\nabla f(u_0, v_0) = \mathbf{0}$, then from (6.55), we have that

$$\begin{bmatrix} E(u_0, v_0) & F(u_0, v_0) \\ F(u_0, v_0) & G(u_0, v_0) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (6.45)$$

There are some other useful formulas involving the First Fundamental Matrix $[I]_{\mathbf{p}}$. Note that $[I]_{\mathbf{p}}$ may be written in terms of the Jacobian matrix $[D_{\mathbf{X}}(u, v)]$ as

$$[I]_{(u,v)} = [D_{\mathbf{X}}(u, v)]^T [D_{\mathbf{X}}(u, v)].$$

Then

$$(a, b) \cdot [I]_{(u,v)}(a, b) = \| [D_{\mathbf{X}}(u, v)](a, b) \|^2 = \| a\mathbf{X}_u(u, v) + b\mathbf{X}_v(u, v) \|^2.$$

Since \mathbf{X}_u and \mathbf{X}_v are linearly independent, $a\mathbf{X}_u(u, v) + b\mathbf{X}_v(u, v) \neq \mathbf{0}$ unless both a and b are zero. Therefore $[I]_{(u,v)}(a, b) = \mathbf{f}0$ if and only if $(a, b) = \mathbf{0}$. By the Fundamental Theorem of Linear Algebra, $[I]_{(u,v)}$ is invertible.

We know that every tangent vector \mathbf{T} at $\mathbf{p} = \mathbf{X}(u_0, v_0)$ can be written as a linear combination of $\mathbf{X}_u(u_0, v_0)$ and $\mathbf{X}_v(u_0, v_0)$. That is,

$$\mathbf{T} = a\mathbf{X}_u(u_0, v_0) + b\mathbf{X}_v(u_0, v_0) = [D_{\mathbf{X}}(u_0, v_0)](a, b).$$

Multiplying both sides by $[D_{\mathbf{X}}(u_0, v_0)]^T$, we obtain $[D_{\mathbf{X}}(u_0, v_0)]^T \mathbf{T} = [I]_{\mathbf{p}}(a, b)$. Therefore,

$$(a, b) = [I]_{\mathbf{p}}^{-1}([D_{\mathbf{X}}(u_0, v_0)]^T \mathbf{T}) \quad \text{whenever} \quad \mathbf{T} = a\mathbf{X}_u(u_0, v_0) + b\mathbf{X}_v(u_0, v_0). \quad (6.46)$$

Finally, if $\mathbf{T}_1 = a\mathbf{X}_u(u_0, v_0) + b\mathbf{X}_v(u_0, v_0)$ and $\mathbf{T}_2 = c\mathbf{X}_u(u_0, v_0) + d\mathbf{X}_v(u_0, v_0)$ are two tangent vectors at \mathbf{p} , we have

$$\mathbf{T}_1 \cdot \mathbf{T}_2 = (a, b) \cdot [I]_{\mathbf{p}}(c, d) \quad (6.47)$$

since

$$\mathbf{T}_1 \cdot \mathbf{T}_2 = ([D_{\mathbf{X}}(u_0, v_0)](a, b)) \cdot ([D_{\mathbf{X}}(u_0, v_0)](c, d)) = (a, b) \cdot [D_{\mathbf{X}}(u_0, v_0)]^T [D_{\mathbf{X}}(u_0, v_0)](c, d).$$

6.3.3 Curvature and the second fundamental matrix

Curvature has to do with how a parameterized surface \mathcal{S} “curves away” from its tangent plant at a point $\mathbf{p} = \mathbf{X}(u_0, v_0)$. To quantify this, recall that if \mathbf{N} is a unit vector in \mathbb{R}^3 , the plane through \mathbf{p} with normal vector \mathbf{N} has the equation $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N} = 0$, and moreover, $|(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}|$ is the distance from \mathbf{x} to the plane. It will be more informative to take of the absolute value: Then the quantity $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{N}$ is the *signed distance* between \mathbf{x} and the plane: It is positive if \mathbf{x} lies on the “positive” side of the plane, and negative if \mathbf{x} lies on the “negative” side of the plane. In either case, its absolute value is the distance between \mathbf{x} and the plane.

Now define a function f from U to \mathbb{R} by

$$f(u, v) = (\mathbf{X}(u, v) - \mathbf{X}(u_0, v_0)) \cdot \widehat{\mathbf{N}}(u_0, v_0). \quad (6.48)$$

Then $f(u, v)$ is the signed distance between $\mathbf{X}(u, v)$ and the tangent plane at $\mathbf{p} = \mathbf{X}(u_0, v_0)$. Notice that $f(u_0, v_0) = 0$. Next, differentiating in u and v , we find that

$$\nabla f(u_0, v_0) = (\mathbf{X}_u \cdot \hat{\mathbf{N}}(u_0, v_0), \mathbf{X}_v \cdot \hat{\mathbf{N}}(u_0, v_0)) = (0, 0)$$

since $\mathbf{X}_u(u_0, v_0)$ and $\mathbf{X}_v(u_0, v_0)$ are orthogonal to $\hat{\mathbf{N}}(u_0, v_0)$. Therefore, in the best quadratic approximation to f at (u_0, v_0) , the only term that can possibly be non-zero is the term involving the Hessian. We now compute the Hessian of f at (u_0, v_0) . For this purpose we introduce the following simple notation:

$$\mathbf{X}_{uu}(u, v) = \frac{\partial^2}{\partial u^2} \mathbf{X}(u, v), \quad \mathbf{X}_{uv}(u, v) = \frac{\partial^2}{\partial u \partial v} \mathbf{X}(u, v) \quad \text{and} \quad \mathbf{X}_{vv}(u, v) = \frac{\partial^2}{\partial v^2} \mathbf{X}(u, v). \quad (6.49)$$

$$\text{Hess}_f(u_0, v_0) = \begin{bmatrix} \hat{\mathbf{N}} \cdot \mathbf{X}_{uu}(u_0, v_0) & \hat{\mathbf{N}} \cdot \mathbf{X}_{uv}(u_0, v_0) \\ \hat{\mathbf{N}} \cdot \mathbf{X}_{vu}(u_0, v_0) & \hat{\mathbf{N}} \cdot \mathbf{X}_{vv}(u_0, v_0) \end{bmatrix}. \quad (6.50)$$

To write this matrix more conveniently, we define the functions L , M and N on U by

$$L(u, v) = \hat{\mathbf{N}} \cdot \mathbf{X}_{uu}(u, v), \quad M(u, v) = \hat{\mathbf{N}} \cdot \mathbf{X}_{uv}(u, v) \quad \text{and} \quad N(u, v) = \hat{\mathbf{N}} \cdot \mathbf{X}_{vv}(u, v). \quad (6.51)$$

Definition 78 (Second Fundamental Matrix). *Let $\mathbf{X}(u, v)$ be a twice continuously differentiable parameterization of a surface \mathcal{S} in \mathbb{R}^3 . Then the Second Fundamental Matrix, $[II]_{\mathbf{p}}$, of the parameterized surface as a function of u and v , $[II]_{(u,v)}$ is*

$$[II]_{(u,v)} = \begin{bmatrix} L(u, v) & M(u, v) \\ M(u, v) & N(u, v) \end{bmatrix} \quad (6.52)$$

where L , M and N are given by (6.51).

As you can see now, the I in $[I]_{\mathbf{p}}$ the II in $[II]_{\mathbf{p}}$ denote the Roman numeral 1 and 2. Using this notation, we can write the second order Taylor expansion of the signed distance function $f(u, v)$ at $\mathbf{p} = \mathbf{X}(u_0, v_0)$ as

$$f(u, v) \approx \frac{1}{2}(u - u_0, v - v_0) \cdot [II]_{\mathbf{p}}(u - u_0, v - v_0). \quad (6.53)$$

Whenever both eigenvalues of the Hessian of f at (u_0, v_0) are strictly positive, f is strictly positive for (u, v) sufficiently close to, but not equal to, (u_0, v_0) . Hence, locally at $\mathbf{p} = \mathbf{X}(u_0, v_0)$, the surface \mathcal{S} lies strictly on the positive side of the tangent plane to \mathcal{S} at \mathbf{p} . Likewise, whenever both eigenvalues of the Hessian of f at (u_0, v_0) are strictly negative, the surface \mathcal{S} lies strictly on the negative side of the tangent plane to \mathcal{S} at \mathbf{p} . Finally, if one eigenvalues is positive and one is negative, part of the surface lies on each side of the tangent plane. Hence the most basic question about the curvature of \mathcal{S} at a point \mathbf{p} – namely whether locally the surface lies to one side of the other of the tangent plane at \mathbf{p} , and if so, which one – can be answered by computing the eigenvalues of the second fundamental matrix in any set of coordinates.

Example 100 (Second fundamental matrix for the sphere). *Let $\mathbf{X}(u, v)$ be the “latitude and longitude” parameterization of the sphere S^2 that is given in (6.39). Then we compute*

$$\mathbf{X}_{uu}(u, v) = (-\sin u \cos v, -\sin u, \sin v, -\cos u)$$

$$\mathbf{X}_{uv}(u, v) = (-\cos u \sin v, \cos u, \cos v, 0)$$

$$\mathbf{X}_{vv}(u, v) = (-\sin u \cos v, \sin u, \sin v, 0)$$

Since $\mathbf{X}_{uu}(u, v) = -\mathbf{X}(u, v) = -\hat{\mathbf{N}}(u, v)$, we find that

$$\begin{bmatrix} L(u, v) & M(u, v) \\ M(u, v) & N(u, v) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -\sin^2 u \end{bmatrix}.$$

Evidently both eigenvalues are negative, so that in a neighborhood of any point $\mathbf{p} \in S^2$, the sphere lies entirely on the negative side of the tangent plane, when we orient the sphere so the $\hat{\mathbf{N}}(\mathbf{p})$ is the outward unit normal. Of course this is geometrically obvious: The entire sphere lies on the negative side of each of its tangent planes.

It is instructive to do the same computation using the stereographic parameterization (6.40). Differentiating and computing the dot products, this time we find:

$$\begin{bmatrix} L(u, v) & M(u, v) \\ M(u, v) & N(u, v) \end{bmatrix} = \frac{-4}{1+u^2+v^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Again, we see that both eigenvalues are negative, but this time they are equal to one another for each u and v , which was not the case when we used “latitude and longitude”.

Example 101 (Second Fundamental Matrix for a graph). Let $f(x, y)$ be a continuously differentiable function on \mathbb{R}^2 . Consider the parameterized surface S given by $\mathbf{X}(u, v) = (u, v, f(u, v))$. As we have seen above,

$$\begin{aligned} \mathbf{X}_{uu}(u, v) &= \left(0, 0, \frac{\partial^2 f(u, v)}{\partial u^2}\right) \\ \mathbf{X}_{uv}(u, v) &= \left(0, 0, \frac{\partial^2 f(u, v)}{\partial u \partial v}\right) \\ \mathbf{X}_{vv}(u, v) &= \left(0, 0, \frac{\partial^2 f(u, v)}{\partial v^2}\right). \end{aligned} \quad (6.54)$$

As we have seen in Example 96,

$$\hat{\mathbf{N}}(u, v) = \frac{1}{\sqrt{1 + \|\nabla f(u, v)\|^2}} \left(-\frac{\partial f(u, v)}{\partial u}, -\frac{\partial f(u, v)}{\partial v}, 1 \right).$$

Therefore,

$$\begin{bmatrix} L(u, v) & M(u, v) \\ M(u, v) & N(u, v) \end{bmatrix} = \frac{1}{\sqrt{1 + \|\nabla f(u, v)\|^2}} \begin{bmatrix} \frac{\partial^2 f(u, v)}{\partial u^2} & \frac{\partial^2 f(u, v)}{\partial u \partial v} \\ \frac{\partial^2 f(u, v)}{\partial u \partial v} & \frac{\partial^2 f(u, v)}{\partial v^2} \end{bmatrix}. \quad (6.55)$$

Note that if (u_0, v_0) is a critical point of f , so that $\nabla f(u_0, v_0) = \mathbf{0}$, then from (6.55), we have that

$$\begin{bmatrix} L(u_0, v_0) & M(u_0, v_0) \\ M(u_0, v_0) & N(u_0, v_0) \end{bmatrix} = [\text{Hess}_f(u_0, v_0)]. \quad (6.56)$$

For example, for $f(x, y) = x^2 + y^2$, the Second Fundamental Matrix for $\mathbf{X}(u, v) = (u, v, f(u, v))$ is $\frac{2}{\sqrt{1 + 4u^2 + 4v^2}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

To quantify curvature, we will use both the First and Second Fundamental Matrices. Consider any twice continuously differentiable curve $\mathbf{u}(t)$ in the surface that passes through (u_0, v_0) at $t = 0$, where $\mathbf{p} = \mathbf{X}(u_0, v_0)$. Let $\mathbf{x}(t) := \mathbf{X}(\mathbf{u}(t))$. Then from (6.48) and (6.53) we have that

$$\lim_{t \rightarrow 0} \frac{1}{t^2} (\mathbf{x}(t) - \mathbf{x}(0)) \cdot \hat{\mathbf{N}}(\mathbf{p}) = \lim_{t \rightarrow 0} \frac{f(\mathbf{u}(t)) - f(\mathbf{u}(0))}{t^2} = \frac{1}{2} \mathbf{u}'(t) \cdot [II]_{\mathbf{p}} \mathbf{u}'(t) .$$

Because $f(\mathbf{u}(t))$ and its first derivative in t are both zero at $t = 0$, this is the same as

$$\left. \frac{d^2}{dt^2} f(\mathbf{u}(t)) \right|_{t=0} = \mathbf{u}'(0) \cdot [II]_{\mathbf{p}} \mathbf{u}'(0) . \quad (6.57)$$

The tangent vector $\mathbf{x}'(0)$ corresponding to $\mathbf{u}'(0)$ is given by $\mathbf{x}'(0) = [D\mathbf{X}(\mathbf{p})]\mathbf{u}'(0)$. To define a geometrically meaningful notion of “directional curvature” of \mathcal{S} at \mathbf{p} , we consider the second derivative of the signed distance $f(\mathbf{u}(t))$ along curves $\mathbf{u}(t)$ such that $[D\mathbf{X}(\mathbf{p})]\mathbf{u}'(0)$ is a unit vector. Finally, note that the quantity in (6.57) depends on the particular curve $\mathbf{u}(t)$ only through its derivative at $t = 0$, $\mathbf{u}'(0)$. All curves have the same derivative at $t = 0$ are equivalent as far as the computation of the second derivative in (6.57) is concerned.

We finally define the *directional curvature* of \mathcal{S} at \mathbf{p} in the unit tangent direction \mathbf{T} at \mathbf{p} by

$$\kappa(\mathbf{p}, \mathbf{T}) = -(a, b) \cdot [II]_{\mathbf{p}}(a, b) \quad \text{where } \mathbf{T} = a\mathbf{X}_u(\mathbf{p}) + b\mathbf{X}_v(\mathbf{p}) . \quad (6.58)$$

Notice the minus sign in the definition; the reason for this will become clear after doing some computations. We now seek to maximize and minimize $\kappa_{\mathbf{p}}(\mathbf{T})$ over all unit tangent vectors \mathbf{T} at \mathbf{p} .

Definition 79 (Principal curvatures and Gaussian curvature). *The principal curvatures of \mathcal{S} at $\mathbf{p} = \mathbf{X}(u_0, v_0)$ are the quantities*

$$\kappa_1(\mathbf{p}) = \max \{ \kappa(\mathbf{p}, \mathbf{T}) : \mathbf{T} \text{ is a tangent vector at } \mathbf{p} \text{ and } \|\mathbf{T}\| = 1 \} ,$$

and

$$\kappa_2(\mathbf{p}) = \min \{ \kappa(\mathbf{p}, \mathbf{T}) : \mathbf{T} \text{ is a tangent vector at } \mathbf{p} \text{ and } \|\mathbf{T}\| = 1 \}$$

where $\kappa(\mathbf{p}, \mathbf{T})$ is defined in (6.58). The Gaussian curvature K of \mathcal{S} at \mathbf{x}_0 is the product $\kappa_1 \kappa_2$.

The formulas look much simpler if one suppresses the dependence on u_0 and v_0 , and multiplies out the products. For instance, we obtain:

$$\kappa_1 = \max \{ -La^2 - 2Mab - Nb^2 : E2a^2 + 2Fab + Gb^2 = 1 \} . \quad (6.59)$$

Notice that κ_1 and κ_2 will change sign if one changes the orientation; i.e., the sign of $\hat{\mathbf{N}}$. However, the Gaussian curvature does not: It is independent of the choice of an orientation, and is an intrinsic property of the surface itself.

To compute the principal curvatures, we must solve a constrained optimization problem: To obtain κ_1 from (6.59), one seeks to maximize

$$h(a, b) = -La^2 - 2Mab - Nb^2$$

subject to the constraint $g(a, b) = 1$ where $g(a, b) = E2a^2 + 2Fab + Gb^2$. Likewise, computing κ_2 is a constrained minimization problem. We know how to solve such problems using the method of Lagrange multipliers.

Theorem 81 (Curvature of Parameterized surfaces). *Let \mathcal{S} be a twice continuously differentiable parameterized surface. Then at any point $\mathbf{X}(u_0, v_0)$ in \mathcal{S} , the principal curvatures κ_1 and κ_2 are the two roots of the quadratic equation $p(t) = 0$ where*

$$p(t) = (L - tE)(N - tG) - (M - tF)^2 . \quad (6.60)$$

Moreover, the Gaussian curvature K is given by

$$K = \kappa_1 \kappa_2 = \frac{LN - M^2}{EG - F^2} . \quad (6.61)$$

All quantities are evaluated at (u_0, v_0) .

Remark 7. Note that K is the quotient of the determinants of the Second Fundamental Matrix and the First fundamental Matrix.

Example 102 (Gaussian curvature of a paraboloid). Let $f(x, y) = x^2 + y^2$. Consider the parameterized surface \mathcal{S} given by $\mathbf{X}(u, v) = (u, v, f(u, v)) = (u, v, u^2 + v^2)$. We have seen above that with $\mathbf{p} := \mathbf{X}(u, v)$, $[I]_{\mathbf{p}} = \begin{bmatrix} 1 + 4u^2 & 4uv \\ 4uv & 1 + 4v^2 \end{bmatrix}$ and $[II]_{\mathbf{p}} = \frac{2}{\sqrt{1 + 4u^2 + 4v^2}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Then by (6.61), the Gaussian curvature is given by $K(u, v) = \frac{(1 + 4u^2 + 4v^2)^2}{2}$.

Proof of Theorem 81. As we have explained above, computing the principal curvatures is a matter of maximizing and minimizing $h(a, b) = La^2 + 2Fab + Nb^2$ subject to the constraint $g(a, b) = 1$ where $g(a, b) = E2a^2 + 2Fab + Gb^2$. Note that the variables here are a and b . Lagrange's equation $\nabla h(a, b) = \lambda \nabla g(a, b)$ works out to

$$-\begin{bmatrix} L & M \\ M & L \end{bmatrix}(a, b) = \lambda \begin{bmatrix} E & F \\ F & G \end{bmatrix}(a, b) . \quad (6.62)$$

This means that

$$\begin{bmatrix} L + \lambda E & M + \lambda F \\ M + \lambda F & L + \lambda N \end{bmatrix}(a, b) = (0, 0) . \quad (6.63)$$

Since $E2a^2 + 2Fab + Gb^2 = 1$, (a, b) is not the zero vector, and so the matrix $\begin{bmatrix} L + \lambda E & M + \lambda F \\ M + \lambda F & L + \lambda N \end{bmatrix}$ is not invertible, and hence its determinant is zero. Therefore,

$$(L + \lambda E)(N + \lambda G) - (M + \lambda F)^2 = 0 ,$$

and hence the Lagrange multiplier λ must be one of the two roots of the quadratic equation

$$\det([II]_{\mathbf{p}} + t[I]_{\mathbf{p}}) = (L + tE)(N + tG) - (M + tF)^2 = 0 .$$

We now show that these roots are the principal curvatures.

Let (a_1, b_1) be a maximizer for our constrained optimization problem. Let λ_1 be the root of $(L + tE)(N + tG) = (M + tF)^2$ that is the Lagrange multiplier for which (6.62) is valid at the maximizer. Taking the dot product of both sides of

$$-\begin{bmatrix} L & M \\ M & L \end{bmatrix}(a_1, b_1) = \lambda_1 \begin{bmatrix} E & F \\ F & G \end{bmatrix}(a_1, b_1)$$

with (a_1, b_1) , and remembering the constraint $g(a_1, b_1) = 1$, we find

$$-La_1^2 - 2Ma_1b_1 - Nb_1^2 = \lambda_1 .$$

Since (a_1, b_1) is the maximizer, and κ_1 is the maximum value, we conclude that $\kappa_1 = \lambda_1$. In the exact same way, we conclude that $\kappa_2 = \lambda_2$.

At this point, we know that the two principal curvatures are the two roots of the quadratic polynomial

$$\begin{aligned} p(t) &= \det([II]_{\mathbf{P}} + t[I]_{\mathbf{P}}) \\ &= (L+tE)(N+tG) - (M+tF)^2 \\ &= (LN - M^2) + (EN + LG - 2MF)t + (EG - F^2)t^2 . \end{aligned} \quad (6.64)$$

Moreover, every quadratic polynomial with roots κ_1 and κ_2 can be written in the form

$$p(t) = C(t - \kappa_1)(t - \kappa_2) = C\kappa_1\kappa_2 + C(\kappa_1 + \kappa_2)t + Ct^2 . \quad (6.65)$$

Comparing the two coefficients of t^2 , we see that $C = EG - F^2$. Comparing the coefficient of $t = 0$ we then deduce the formula (6.61). \square

Example 103 (principal curvatures for a graph at a critical point). *Let $f(x, y)$ be a continuously differentiable function on \mathbb{R}^2 . Consider the parameterized surface \mathcal{S} given by*

$$\mathbf{X}(u, v) = (u, v, f(u, v)) .$$

As we have seen in (6.55) and (6.56), when (u_0, v_0) is a critical point of f , the First Fundamental Matrix at (u_0, v_0) is the identity matrix, and the Second Fundamental Matrix at (u_0, v_0) is $[\text{Hess}_f(u_0, v_0)]$. In this case, the polynomial $p(t)$ defined in (6.64) is simply the characteristic polynomial of $-[\text{Hess}_f(u_0, v_0)]$, and hence the principal curvatures are exactly the eigenvalues of $-[\text{Hess}_f(u_0, v_0)]$. In particular, if the principal curvatures κ_1 and κ_2 are both strictly positive, (u_0, v_0) is a local maximum of f .

The proof of Theorem 81 tells us something important about the first and second fundamental matrices. Let (a, b) be one of the unit vectors associated with either of the principal curvatures, as in (6.62) Then, multiplying both sides through by $\begin{bmatrix} E & F \\ F & G \end{bmatrix}^{-1}$, we see that

$$-\left[\begin{array}{cc} E & F \\ F & G \end{array} \right]^{-1} \left[\begin{array}{cc} L & M \\ M & N \end{array} \right] (a, b) = \lambda(a, b) . \quad (6.66)$$

In other words, the unit tangent vectors associated with the principal curvatures are eigenvectors of the matrix $[I]_{\mathbf{P}}^{-1}[II]_{\mathbf{P}}$ which is known as the matrix of the *shape operator*. We will study shape operator in the next section in connection with the *Gauss map*. For now, notice that the principal curvatures turn out to be the eigenvalues of the shape operator, and an alternate way to compute them is to compute the eigenvalues of $[I]_{\mathbf{P}}^{-1}[II]_{\mathbf{P}}$.

Example 104 (Gaussian curvature of the sphere). Let $\mathbf{X}(u, v)$ be the “latitude and longitude” parameterization of the sphere S^2 that is given in (6.39). Let $\mathbf{p} = \mathbf{X}(u, v)$. In Examples 98 and 100, we have computed that

$$[I]_{\mathbf{p}} = \begin{bmatrix} 1 & 0 \\ 0 & \sin^2 u \end{bmatrix} \quad \text{and} \quad [II]_{\mathbf{p}} = \begin{bmatrix} -1 & 0 \\ 0 & -\sin^2 u \end{bmatrix}.$$

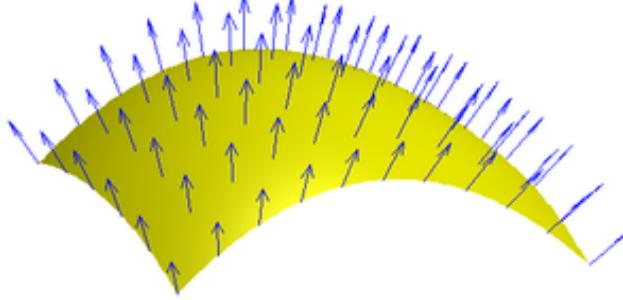
Then

$$K = \frac{\det([II]_{\mathbf{p}})}{\det([I]_{\mathbf{p}})} = \frac{\sin^2 u}{\sin^2 u} = 1.$$

6.3.4 The Gauss map

Given a parameterized surface \mathcal{S} , the *Gauss map* is the function from \mathcal{S} to the unit sphere, S^2 , that sends a point $\mathbf{p} \in \mathcal{S}$ to the unit normal vector $\hat{\mathbf{N}}(\mathbf{p})$ at \mathbf{p} . That is, if \mathbf{p} has coordinates (u, v) , so that $\mathbf{p} = \mathbf{X}(u, v)$, then the image of \mathbf{p} under the Gauss map is $\hat{\mathbf{N}}(u, v)$.

If \mathcal{S} is a plane, then $\hat{\mathbf{N}}(\mathbf{p})$ is constant, and of course there is no curvature. In this case, the image of all of \mathcal{S} under the Gauss map is a single point. However, if we consider the figure below,



we see that the image of the pictured patch of surface under the Gauss map covers a patch around the “North Pole” of the unit sphere S^2 . By studying how $\hat{\mathbf{N}}(u, v)$ changes as u and v vary, we can quantify the curvature of a parameterized surface, and this gives another way to think about the geometric meaning of the principal curvatures introduced in the previous section.

Let \mathcal{S} be a surface parameterized by $\mathbf{X}(u, v) = (x(u, v), y(u, v), z(u, v))$ for (u, v) in some open set $U \subset \mathbb{R}^2$. Suppose that the functions $x(u, v)$, $y(u, v)$, $z(u, v)$ are twice continuously differentiable. Let $(u_0, v_0) \in U$, and let $\mathbf{p} = \mathbf{X}(u_0, v_0) \in \mathcal{S}$.

We now give a formula for the Second Fundamental Matrix of \mathcal{S} in terms of the Jacobian matrix of the Gauss map. The 3×2 matrix $\left[\frac{\partial}{\partial u} \hat{\mathbf{N}}(u_0, v_0), \frac{\partial}{\partial v} \hat{\mathbf{N}}(u_0, v_0) \right]$ is just the Jacobian matrix of the function $\hat{\mathbf{N}}(u, v)$ at (u_0, v_0) , namely $[D_{\hat{\mathbf{N}}}(u_0, v_0)]$.

Lemma 21. Let \mathcal{S} be a parameterized surface given by a twice continuously differentiable function $\mathbf{X}(u, v)$ for $(u, v) \in U \subset \mathbb{R}^2$. Let $\mathbf{p} = \mathbf{X}(u_0, v_0)$, $(u_0, v_0) \in U$. Then

$$[II]_{\mathbf{p}} = -[D_{\mathbf{X}}(u_0, v_0)]^T [D_{\hat{\mathbf{N}}}(u_0, v_0)].$$

Proof. Note that

$$[D_{\mathbf{X}}(u_0, v_0)]^T [D_{\hat{\mathbf{N}}}(u_0, v_0)] = \begin{bmatrix} \mathbf{X}_u \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}} & \mathbf{X}_u \cdot \frac{\partial}{\partial v} \hat{\mathbf{N}} \\ \mathbf{X}_v \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}} & \mathbf{X}_v \cdot \frac{\partial}{\partial v} \hat{\mathbf{N}} \end{bmatrix} \quad (6.67)$$

where all derivatives on the right are evaluated at (u_0, v_0) .

Next, since $\mathbf{X}_u(u, v) \cdot \hat{\mathbf{N}}(u, v) = 0$ for all u and v , differentiating with respect to u yields

$$0 = \frac{\partial}{\partial u} \mathbf{X}_u(u, v) \cdot \hat{\mathbf{N}}(u, v) + \mathbf{X}_u(u, v) \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}}(u, v) .$$

Using the notation introduced in (6.49), $\cdot \mathbf{X}_u(u, v) \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}}(u, v) = -\hat{\mathbf{N}}(u, v) \cdot \mathbf{X}_{uu}(u, v)$. In the same way, we derive

$$\begin{aligned} \mathbf{X}_u(u, v) \cdot \frac{\partial}{\partial v} \hat{\mathbf{N}}(u, v) &= -\hat{\mathbf{N}}(u, v) \cdot \mathbf{X}_{vu}(u, v) \\ \mathbf{X}_v(u, v) \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}}(u, v) &= -\hat{\mathbf{N}}(u, v) \cdot \mathbf{X}_{vu}(u, v) \\ \mathbf{X}_v(u, v) \cdot \frac{\partial}{\partial v} \hat{\mathbf{N}}(u, v) &= -\hat{\mathbf{N}}(u, v) \cdot \mathbf{X}_{vv}(u, v) \end{aligned}$$

Hence by the definition (6.51) and (6.52) of $[II]$, and with all derivatives evaluated at (u_0, v_0) ,

$$\begin{bmatrix} \mathbf{X}_u \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}} & \mathbf{X}_u \cdot \frac{\partial}{\partial v} \hat{\mathbf{N}} \\ \mathbf{X}_v \cdot \frac{\partial}{\partial u} \hat{\mathbf{N}} & \mathbf{X}_v \cdot \frac{\partial}{\partial v} \hat{\mathbf{N}} \end{bmatrix} = - \begin{bmatrix} \hat{\mathbf{N}} \cdot \mathbf{X}_{uu} & \hat{\mathbf{N}} \cdot \mathbf{X}_{vu} \\ \hat{\mathbf{N}} \cdot \mathbf{X}_{vu} & \hat{\mathbf{N}} \cdot \mathbf{X}_{vv} \end{bmatrix} = -[II]_{\mathbf{P}} .$$

□

Now consider a curve $\mathbf{x}(t) = \mathbf{X}(\mathbf{u}(t))$. At each time t , we have the unit normal vector $\hat{\mathbf{N}}(\mathbf{x}(t))$ and the tangent vector $\mathbf{x}'(t) = [D_{\mathbf{X}}(\mathbf{u}(t))\mathbf{u}'(t)]$. Since $\hat{\mathbf{N}}(\mathbf{x}(t))$ is a unit vector,

$$0 = \frac{d}{dt} 1 = \frac{d}{dt} (\hat{\mathbf{N}}(\mathbf{x}(t)) \cdot \hat{\mathbf{N}}(\mathbf{x}(t))) = 2 \left(\frac{d}{dt} \hat{\mathbf{N}}(\mathbf{x}(t)) \right) \cdot \hat{\mathbf{N}}(\mathbf{x}(t)) .$$

Since for each t , $\left(\frac{d}{dt} \hat{\mathbf{N}}(\mathbf{x}(t)) \right)$ is orthogonal to $\hat{\mathbf{N}}(\mathbf{x}(t))$, it lies in the tangent plane to the surface at $\mathbf{x}(t)$.

Since the tangent plane is spanned by $\mathbf{X}_u(u_0, v_0)$ and $\mathbf{X}_v(u_0, v_0)$, there are numbers c and d such that

$$\left. \frac{d}{dt} \hat{\mathbf{N}}(\mathbf{x}(t)) \right|_{t=0} = c \mathbf{X}_u(u_0, v_0) + d \mathbf{X}_v(u_0, v_0) = [D_{\mathbf{X}}(u_0, v_0)](c, d) .$$

On the other hand, by the chain rule, for $\mathbf{x}(t) = (u(t), v(t))$ with $u'(0) = a$ and $v'(0) = b$,

$$\left. \frac{d}{dt} \hat{\mathbf{N}}(\mathbf{x}(t)) \right|_{t=0} = a \frac{\partial}{\partial u} \hat{\mathbf{N}}(u_0, v_0) + b \frac{\partial}{\partial v} \hat{\mathbf{N}}(u_0, v_0) = [D_{\hat{\mathbf{N}}}(u_0, v_0)](a, b) .$$

We conclude that $[D_{\mathbf{X}}(u_0, v_0)](c, d) = [D_{\hat{\mathbf{N}}}(u_0, v_0)](a, b)$. Multiplying on the left by $[D_{\mathbf{X}}(u_0, v_0)]^T$ and recalling that the First Fundamental Matrix is given by $[I]_{\mathbf{P}} = [D_{\mathbf{X}}(u_0, v_0)]^T [D_{\mathbf{X}}(u_0, v_0)]$, we have $[D_{\mathbf{X}}(u_0, v_0)]^T [D_{\mathbf{X}}(u_0, v_0)](c, d) = [D_{\mathbf{X}}(u_0, v_0)]^T [D_{\hat{\mathbf{N}}}(u_0, v_0)](a, b)$, which is the same as

$$[I]_{\mathbf{P}}(c, d) = -[II]_{\mathbf{P}}(a, b) . \quad (6.68)$$

We have proved:

Theorem 82. Let \mathcal{S} be a parameterized surface given by a twice continuously differentiable function $\mathbf{X}(u, v)$ for $(u, v) \in U \subset \mathbb{R}^2$. Let $\mathbf{x}(t)$ be any curve in \mathcal{S} passing through $\mathbf{x}(u_0, v_0)$ at $t = 0$ with

$$\mathbf{x}'(0) = a\mathbf{X}_u(u_0, v_0) + b\mathbf{X}_v(u_0, v_0) .$$

Then

$$\frac{d}{dt}\widehat{\mathbf{N}}(\mathbf{x}(t)) = c\mathbf{X}_u(u_0, v_0) + d\mathbf{X}_v(u_0, v_0) ,$$

where $(c, d) = -[I]_{\mathbf{p}}^{-1}[II]_{\mathbf{p}}(a, b)$.

Definition 80 (The shape operator). The shape operator S is the linear transformation on the tangent space to \mathcal{S} at a point $\mathbf{p} = \mathbf{X}(u_0, v_0)$ that sends the tangent vector \mathbf{T} to the tangent vector $-\frac{d}{dt}\widehat{\mathbf{N}}(\mathbf{x}(t))\Big|_{t=0}$ where $\mathbf{x}(t)$ is any continuously differentiable curve in \mathcal{S} with $\mathbf{x}'(0) = \mathbf{T}$.

By Theorem 82, if $\mathbf{T} = a\mathbf{X}_u(u_0, v_0) + b\mathbf{X}_v(u_0, v_0)$, then $S(\mathbf{T}) = c\mathbf{X}_u(u_0, v_0) + d\mathbf{X}_v(u_0, v_0)$ where $(c, d) = [I]_{\mathbf{p}}^{-1}[II]_{\mathbf{p}}(a, b)$.

For any two tangent vectors $\mathbf{T}_1 = a_1\mathbf{X}_u + b_1\mathbf{X}_v$ and $\mathbf{T}_2 = a_2\mathbf{X}_u + b_2\mathbf{X}_v$, consider the quantity

$$\mathbf{T}_1 \cdot S(\mathbf{T}_2) .$$

Then $S(\mathbf{T}_2) = c\mathbf{X}_u + d\mathbf{X}_v$ where $(c, d) = -[I]_{\mathbf{p}}^{-1}[II]_{\mathbf{p}}(a_2, b_2)$. Then by (6.47),

$$\mathbf{T}_1 \cdot S(\mathbf{T}_2) = (a_1, b_1) \cdot [I]_{\mathbf{p}}(c, d) = -(a_1, b_1) \cdot [I]_{\mathbf{p}}[I]_{\mathbf{p}}^{-1}[II]_{\mathbf{p}}(c, d)(a_2, b_2) = -(a_1, b_1) \cdot [II]_{\mathbf{p}}(a_2, b_2) .$$

Since the Second Fundamental Matrix is symmetric, this means that the shape operator is symmetric in the sense that for all tangent vectors \mathbf{T}_1 and \mathbf{T}_2 ,

$$\mathbf{T}_1 \cdot S(\mathbf{T}_2) = S(\mathbf{T}_1) \cdot \mathbf{T}_2 .$$

We have seen that the matrix $-[I]_{\mathbf{p}}^{-1}[II]_{\mathbf{p}}$ has eigenvalues κ_1 and κ_2 , the principal curvatures. Let (a_1, b_1) and (a_2, b_2) be the corresponding eigenvectors normalized by $(a_j, b_j) \cdot [I]_{\mathbf{p}}(a_j, b_j) = 1$ for $j = 1, 2$. Let $T_j = a_j\mathbf{X}_u(u_0, v_0) + b_j\mathbf{X}_v(u_0, v_0)$. Then by Theorem 82 and the definition of the shape operator

$$S(\mathbf{T}_1) = \kappa_1 \mathbf{T}_1 \quad \text{and} \quad S(\mathbf{T}_2) = \kappa_2 \mathbf{T}_2 .$$

When $\kappa_1 \neq \kappa_2$, we have

$$\kappa_1 \mathbf{T}_1 \cdot \mathbf{T}_2 = S(\mathbf{T}_1) \cdot \mathbf{T}_2 = \mathbf{T}_1 \cdot S(\mathbf{T}_2) = \kappa_2 \mathbf{T}_1 \cdot \mathbf{T}_2 ,$$

and consequently $\mathbf{T}_1 \cdot \mathbf{T}_2 = 0$. When $\kappa_1 = \kappa_2$, the shape operator is simply a multiple of the identity. In any case, we see that there is always an orthonormal basis of the tangent space consisting of eigenvectors of the shape operator.

6.4 Exercises

6.1: Let $f(x, y) = x^2y + xy^2 - xy$. Evaluate the Hessian matrix of f at $\mathbf{x}_0 := (1, 1)$ and $\mathbf{v} := (1, 2)$, and compute the second directional derivative

$$\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}_0)\Big|_{t=0} .$$

6.2: Let $f(x, y) = x^4 + y^4 - 4xy$. Evaluate the Hessian matrix of f at $\mathbf{x}_0 := (0, 1)$ and $\mathbf{v} := (1, -1)$, and compute the second directional derivative

$$\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}_0) \Big|_{t=0} .$$

6.3: Let $f(x, y, z) = x^2yz + xy^2 - xz$. Evaluate the Hessian matrix of f at $\mathbf{x}_0 := (1, 1, 1)$ and $\mathbf{v} := (1, 0, 1)$, and compute the second directional derivative

$$\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}_0) \Big|_{t=0} .$$

6.4: Let $f(x, y, z) = xyz - xy + xz$. Evaluate the Hessian matrix of f at $\mathbf{x}_0 := (1, 0, 1)$ and $\mathbf{v} := (1, 2, 1)$, and compute the second directional derivative

$$\frac{d^2}{dt^2} f(\mathbf{x}_0 + t\mathbf{v}_0) \Big|_{t=0} .$$

6.5: Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$. Find the eigenvalues of A and an orthonormal basis of \mathbb{R}^2 consisting of eigenvectors of A

6.6: Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

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

(a) Show that

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

and

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

and that both of these partial derivatives are continuous everywhere on \mathbb{R}^2 .

(b) Show that for $(x, y) \neq (0, 0)$,

$$\frac{\partial}{\partial x} \frac{\partial}{\partial y} f(x, y) = \frac{\partial}{\partial y} \frac{\partial}{\partial x} f(x, y) = \frac{x^6 + 9x^4y^2 - 9x^2y^4 - y^6}{(x^2 + y^2)^3} .$$

Show that the function on the right hand side does not have a limiting value at $(0, 0)$; i.e.,

$$\lim_{(x,y) \rightarrow (0,0)} \frac{x^6 + 9x^4y^2 - 9x^2y^4 - y^6}{(x^2 + y^2)^3}$$

does not exist.

(c) Show that $\frac{\partial}{\partial x} \frac{\partial}{\partial y} f(0, 0)$ and $\frac{\partial}{\partial y} \frac{\partial}{\partial x} f(0, 0)$ both exist, and compute the values. How does this result fit with Clairaut's Theorem?

6.7: Let $A = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$. Find the eigenvalues of A and an orthonormal basis of \mathbb{R}^2 consisting of eigenvectors of A

6.8: Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$. Find the eigenvalues of A and an orthonormal basis of \mathbb{R}^2 consisting of eigenvectors of A

6.9: Let $A = \begin{bmatrix} -1 & 2 \\ 2 & 4 \end{bmatrix}$. Find the eigenvalues of A and an orthonormal basis of \mathbb{R}^2 consisting of eigenvectors of A

6.10: Let $f(x, y) = x^2y + xy^2 - xy$.

(a) Find all of the critical points of f . Evaluate the Hessian matrix of f at each of these critical points, and determine where each is a local maximum, a local minimum, a saddle, or undecidable from the Hessian.

(b) There is one critical point in the interior of the upper right quadrant. Sketch a contour plot of f in the vicinity of this critical point. Show the computations that lead to the plot.

6.11: Let $f(x, y) = 3xy^3 + 2x + \frac{1}{4}x^4 + \frac{9}{2}y^2$.

(a) Find all of the critical points of f . Evaluate the Hessian matrix of f at each of these critical points, and determine where each is a local maximum, a local minimum, a saddle, or undecidable from the Hessian.

(b) Choose one of the critical points and sketch a contour plot of f in the vicinity of this critical point. Show the computations that lead to the plots.

6.12: Let $f(x, y) = 3x^3 + 5xy + 5x^2 - 5y^2$.

(a) Find all of the critical points of f . Evaluate the Hessian matrix of f at each of these critical points, and determine where each is a local maximum, a local minimum, a saddle, or undecidable from the Hessian.

(b) Sketch a contour plot of f in the vicinity of each of the critical points. Show the computations that lead to the plots.

6.13: Let $f(x, y) = x^2 + y^2 - 2yx^2$.

(a) Find all of the critical points of f . Evaluate the Hessian matrix of f at each of these critical points, and determine where each is a local maximum, a local minimum, a saddle, or undecidable from the Hessian.

(b) Sketch a contour plot of f in the vicinity of each of the critical points. Show the computations that lead to the plots.

6.14: Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = 4xy - x^4 - y^4$.

(a) Let $\mathbf{x}(t)$ be given by $\mathbf{x}(t) = (t + t^2, t^2 + t^3)$. Compute $\frac{d}{dt} f(\mathbf{x}(t)) \Big|_{t=1}$.

- (b) Find all of the critical points of f , and find the value of f at each of the critical points.
- (c) Compute the Hessian of f at each critical point and determine whether each critical point is a local minimum, a local maximum, a saddle point, or if it cannot be classified through a computation of the Hessian.
- (d) Does f have a maximum value? Explain why or why not. If it does, find all points at which the value of f is maximal; i.e., find all maximizers.
- (e) Does f have a minimum value? Explain why or why not. If it does, find all points at which the value of f is minimal; i.e., find all minimizers.
- (f) Sketch a contour plot of f near each critical point.

6.15: Let $f(x, y) = x^4 + y^4 - 2x^2y$. There is exactly one critical point (x_0, y_0) with $x_0 > 0$ and $y_0 > 0$.

- (a) Compute the Hessian of f at (x_0, y_0) , and determine whether it is a local minimum, a local maximum, a saddle point, or if it cannot be classified through a computation of the Hessian.
- (b) Let $\mathbf{u} = (u, v)$ be a unit vector, and consider the directional second derivative

$$\frac{d^2}{dt^2} f(x_0 + tu, y_0 + tv) \Big|_{t=0}.$$

Which choice of the unit vector (u, v) makes this as large as possible? What is the largest possible value? Also, which choice of the unit vector (u, v) makes this as small as possible, and what is the smallest possible value?

- (c) Sketch a contour plot of f near (x_0, y_0) .

6.16: Let $f(x, y) = x^4 + y^4 - 4xy$.

- (a) Find all of the critical points of f , and for each of them, determine whether it is a local minimum, a local maximum, a saddle point, or if it cannot be classified through a computation of the Hessian.
- (b) There is one critical point of f in the interior of the upper right quadrant. Let $\mathbf{x}_0 = (x_0, y_0)$ denote this critical point. Let $\mathbf{u} = (u, v)$ be a unit vector, and consider the directional second derivative

$$\frac{d^2}{dt^2} f(x_0 + tu, y_0 + tv) \Big|_{t=0}.$$

Which choices of the unit vector (u, v) make this as large as possible? What is the largest possible value? Also, which choices of the unit vector (u, v) make this as small as possible, and what is the smallest possible value?

- (c) Sketch a contour plot of f near (x_0, y_0) .

6.17: Let $f(x, y) = 4x^2 + y^2 + 4xy - (x - 1)^4 - (y - 1)^4 - 4x - 4y$.

- (a) The point $(0, 0)$ is a critical point of f . Compute the Hessian of f at this point, and determine whether it is a local minimum, a local maximum, a saddle point, or if it cannot be classified through a computation of the Hessian.

(b) Let $\mathbf{u} = (u, v)$ be a unit vector, and consider the directional second derivative

$$\frac{d^2}{dt^2} f(tu, tv) \Big|_{t=0} .$$

Which choice of the unit vector (u, v) makes this as large as possible? What is the largest possible value? Also, which choice of the unit vector (u, v) makes this as small as possible, and what is the smallest possible value?

(c) Sketch a contour plot of f near $(0, 0)$.

6.18 Consider the function $f(x, y, z)$ given by

$$f(x, y, z) = x^3yz^2 + 4xy - 3yz .$$

Determine whether all of the eigenvalues of the Hessian at $\mathbf{x}_0 = (1, 1, 1)$ are positive or if they are all negative, or neither.

6.19 Consider the function $f(x, y, z)$ given by

$$f(x, y, z) = xyz^2 + xy^2z + x^2yz .$$

Determine whether all of the eigenvalues of the Hessian at $\mathbf{x}_0 = (1, 1, 1)$ are positive or if they are all negative, or neither.

6.20 Compute κ_1 , κ_2 and K for all points on the sphere of radius R in \mathbb{R}^3 .

6.21 Compute κ_1 , κ_2 and K at all points on the cone $z = \sqrt{x^2 + y^2}$ at which $z > 0$. (The surface is not differentiable at the apex of the cone.)

6.22 Compute κ_1 , κ_2 and K at all points on the surface given by $z = xy$.

Chapter 7

INTEGRATION IN SEVERAL VARIABLES

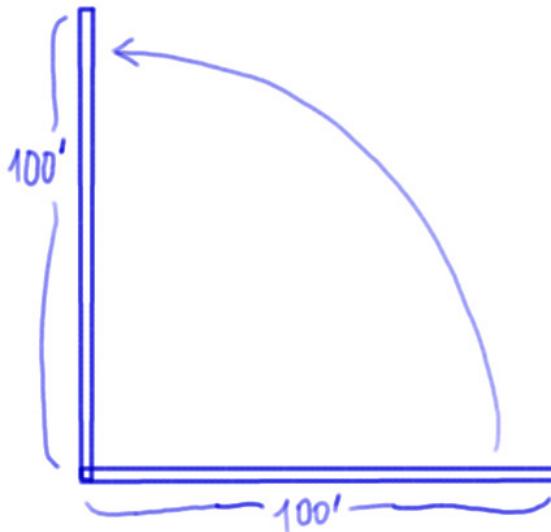
7.1 Integration and summation

7.1.1 A look back at integration in one variable

This chapter is focused on the problem of integrating functions of several variables. There are many ways to think about what it means to integrate a function of one variable, and not all of them are useful starting points for the transition to several variables. For example, integration is often thought of as the procedure that “undoes differentiation”. This is not unreasonable: in practice, one computes integrals by finding antiderivatives. But suppose we have a function $f(x, y)$ of two variables. What could it possibly mean to find an “antiderivative” of $f(x, y)$? We have come to understand the gradient as the derivative in two variables, but that is a vector quantity, and so it would make no sense to seek an “antigradient” of $f(x, y)$.

Therefore, we begin with a problem in one variable. Our aim is to explain *what integrals are* from a point of view that facilitates the transition to several variables. Consider the following problem:

- *How much work is required to raise a 100 foot flagpole that weighs one pound per foot from horizontal to vertical?*



Recall that *work* is product of force and the distance traveled in moving against the direction of that force. (Force is a vector quantity, so it has a direction). In the case at hand, the force is gravity. The direction is “straight down”, so we are only concerned with vertical displacement.

If we lift a one pound weight one foot, we do one foot–pound of work. If we lift a one pound weight ten feet, we do ten foot–pounds of work. This is all simple multiplication. The flagpole problem, however, requires calculus because different parts of the flagpole get raised different amounts. Near the base, there is not much raising going on at all, while the parts of the flagpole near the top are raised nearly 100 feet. We cannot simply use the formula

$$\text{work} = \text{weight} \times \text{height raised} \quad (7.1)$$

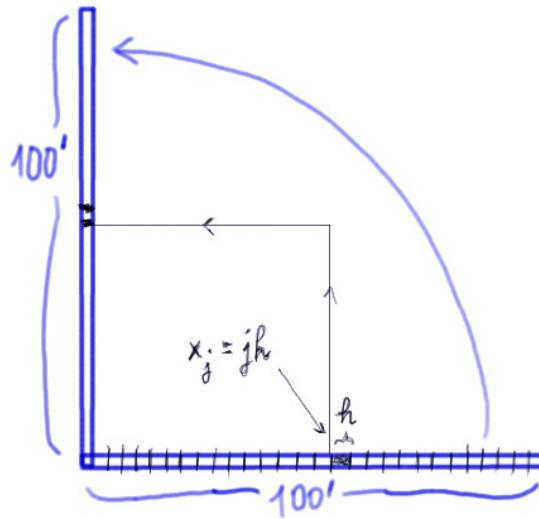
because there is no one value for “height lifted” that is valid for the whole flagpole.

The way to carry out the computation using calculus is to first “slice” the flagpole into small bits – in your mind only; do not ruin the flagpole! Pick a *small* “slice size” $h > 0$, and slice the flagpole perpendicular to its axis into *small* blocks that are h feet long. Now “raise” the flagpole by stacking the blocks in the right order. We are going to add up the amount of work we do lifting each block into place to get the total work done lifting the flagpole into place.

The j th block from the base will have to be raised to a height of $j \times h$ feet. Not everything in the block gets raised by *exactly* this amount, but if h is small compared with $j \times h$, the difference will be small *percentage-wise*.

“Percentage-wise” is the key word: The block itself is *small*, so an approximation of the amount of work it takes to lift it into place that is off by a factor of 2 would make a *small* error. But adding up all of these errors for each of the blocks – which will be many in number – we could be off by a factor of 2 for the flagpole as a whole. But if we are only off by a small percentage in each block, when we add up all of the errors, we will only be off by a small percentage for the flagpole as a whole. And if we can arrange that the percentage error goes to zero as the size of the small blocks goes to zero, we can get an *exact answer* by taking the limit in which the block size goes to zero. This is the fundamental idea on which the integral calculus is based.

Returning to our example, up to a small percentage-wise error, we can use the formula (7.1). Since the flagpole weighs 1 pound per foot, the weight of the block is h pounds.



Hence the work done in raising the j th block is

$$(h \text{ pounds}) \times (jh \text{ feet}) = (jh^2 \text{ footpounds}) .$$

Now the key point is that work is an *extensive*, or in other words, *additive*, quantity. Hence, letting N be the total number of blocks, which is $100/h$, we have that the total work is

$$\sum_{j=1}^N jh^2 .$$

Letting x_j denote jh , and letting Δx denote h , this becomes

$$\sum_{j=1}^N x_j \Delta x ,$$

and you recognize this as a Riemann sum for the integral

$$\int_0^{100} x dx = \frac{x^2}{2} \Big|_0^{100} = 5,000 .$$

This is what we get for the sum in the limit as $h \rightarrow 0$. Hence, the total work done is 5,000 foot-pounds.

Now let us consider what we have done, and identify the essential steps. Integration means “making whole”. This refers to the “adding up” procedure towards the end of the problem, and we used an antiderivative – namely $x^2/2$, which is an antiderivative of x – to do the sum in the limit $h \rightarrow 0$. (This is the passage from the Riemann sum to the integral, with which you are familiar from single variable calculus).

•However, before you can “make something whole”, you have to first “take it apart”, and higher dimensional integration problems generally begin as “disintegration problems”. Depending on how

you choose to “slice your problem into bits” at the beginning, you can be faced with integration problems of quite different degrees of difficulty.

So although we are studying integration in this part of the course, much of our effort will be focused on *disintegration* – we want to do this in a thoughtful, careful way that facilitates the integration steps at the end. We begin with some simple problems in which the most obvious sort of disintegration works well.

7.1.2 Integrals of functions on \mathbb{R}^2

Consider a region D in \mathbb{R}^2 . To be concrete, suppose that D is the closed unit disk in \mathbb{R}^2 . That is, D consists of all points (x, y) satisfying

$$x^2 + y^2 \leq 1 . \quad (7.2)$$

Suppose that we have a sheet of metal lying in this region, and it has a mass density of $f(x, y)$ mass units per area units. (Grams per square centimeter if you like). What is the total weight of the sheet of metal?

If the mass density function $f(x, y)$ were constant, we could use the formula

$$\text{mass} = \text{mass density} \times \text{area} . \quad (7.3)$$

If x and y are measured in centimeters, the area of D is π square centimeters, and so if the density were a uniform 1 gram per square centimeter, the total weight would be π grams.

But suppose that the disk of metal is thinner near the center, and has the mass density

$$f(x, y) = x^2 + y^2 .$$

What would be the total weight in this case? Less, clearly, but how much less?

The way forward is to *disintegrate the the disk* into small bits in which the mass density is effectively constant, and then to apply the formula (7.3) to each of these. This gives us the mass of each of the pieces. Since the mass of the whole is the sum of the mass of the parts, all we need do is to add up all of these masses, and make the disk whole again. This is the integration phase.

To disintegrate the disk, we chop it up on a rectangular grid. Let Δx be the spacing between the vertical grid lines and let Δy be the spacing between the horizontal grid lines. Most of the disk is covered by rectangular “tiles” of area $\Delta x \Delta y$. There are some tiles with more complicated shapes around the boundary, but these will account for a small percentage of the disk if both Δx and Δy are very small. Hence, let us ignore these for now, and focus on the rectangular tiles. In each of these, the mass density does not vary much – at least when both Δx and Δy are very small – so it makes sense to talk about the value of the mass density in the little tile. For each such tile, we have that the mass is

$$(\text{mass density in the tile}) \times \Delta x \Delta y .$$

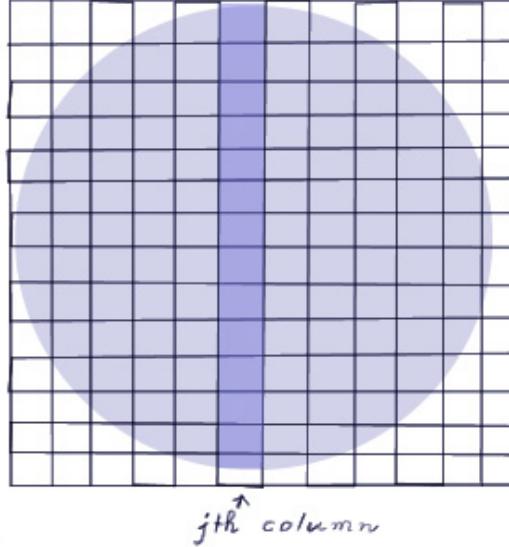
Hence the total mass is

$$\sum_{\text{little tiles}} (\text{mass density in the tile}) \times \Delta x \Delta y . \quad (7.4)$$

Now we are ready for the integration phase. We can add up the terms in the sum in any order we like – addition is commutative, and the sum is finite. There are two very natural ways to proceed:

- We and add up the contributions from the tiles in each column, and then we can add up the sums for each column, or we can add up the contributions from the tiles in each row, and then we can add up the sums for each row.

Lets first add up the contributions from the tiles in each column. Suppose that there are M columns, labeled by $j = 1, 2, \dots, M$.



Then we have

$$\begin{aligned}
 & \sum_{\text{little tiles}} (\text{mass density in the tile}) \times \Delta x \Delta y \\
 &= \sum_{i=1}^N \left(\sum_{\text{little tiles in column } j} (\text{mass density in the tile}) \times \Delta x \Delta y \right) \\
 &= \sum_{i=1}^N \left(\sum_{\text{little tiles in column } j} (\text{mass density in the tile}) \times \Delta y \right) \Delta x
 \end{aligned} \tag{7.5}$$

If x_j is the x coordinate of, say, the middle of the j th column, then the inner sum,

$$\sum_{\text{little tiles in column } j} (\text{mass density in the tile}) \times \Delta y$$

is the Riemann sum for the integral

$$\int_{a(x_j)}^{b(x_j)} f(x_j, y) dy ,$$

where $a(x_j)$ is the y coordinate at the bottom of the j th column and $b(x_j)$ is the y coordinate at the top of the j th column. In the case at hand, from the equation $x^2 + y^2 = 1$ at the boundary of the region, we have

$$a(x_j) = -\sqrt{1 - x_j^2} \quad \text{and} \quad b(x_j) = \sqrt{1 - x_j^2}$$

and so, in more concrete terms, our integral is

$$\int_{-\sqrt{1-x_j^2}}^{\sqrt{1-x_j^2}} f(x_j, y) dy .$$

For any fixed value of x_j , this is a garden variety definite integral in the single variable y . Doing it, we get

$$x_j^2 y + \frac{y^3}{3} \Big|_{-\sqrt{1-x_j^2}}^{\sqrt{1-x_j^2}} = 2x_j^2 \sqrt{1-x_j^2} + \frac{2}{3}(1-x_j^2)^{3/2} .$$

Going back to (7.5), we see that, upon replacing the inner sum by the integral to which it corresponds (when viewed as a Riemann sum), we have that the total mass is

$$\sum_{i=1}^N \left(2x_j^2 \sqrt{1-x_j^2} + \frac{2}{3}(1-x_j^2)^{3/2} \right) \Delta x .$$

Since the values of x in the disk range from -1 to 1 , this is the Reimann sum for

$$\int_{-1}^1 (2x^2(1-x^2)^{1/2} + \frac{2}{3}(1-x^2)^{3/2}) dx .$$

Using the trigonometric substitution $x = \sin \theta$, this is easily evaluated, and the answer is $\pi/2$.

In the limit as Δx and Δy both tend to zero, the approximations that we made in replacing sums by integrals, and choosing values in the small tiles, etc, all become increasingly negligible, and so this is the exact value for the total mass.

This problem makes for a good case study of the process of disintegration and integration. Here is the general process: Let D be some region given by inequalities of the form

$$a(x) \leq y \leq b(x) \quad c \leq x \leq d .$$

Let f be a continuous function defined on D . We summarize and generalize:

Definition 81 (Two dimensional area integral). *Given a continuous function f defined on a set $D \subset \mathbb{R}^2$ that is closed, bounded and has a non-empty interior whose boundary is a piecewise differentiable curve, we define the area integral $\int_D f(x, y) dA$ to be*

$$\int_D f(x, y) dA = \lim_{\substack{\max \text{ tile diameter} \rightarrow 0 \\ \text{little tiles}}} \left(\sum_{\text{little tiles}} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) \right) , \quad (7.6)$$

where the limit is taken along a sequence of disintegration of D into tiles in which the maximum tile diameter in the n th disintegration goes to zero as n goes to infinity. More precisely, we require that for some sequence $\{r_n\}$ of positive numbers with $\lim_{n \rightarrow \infty} r_n = 0$, each tile in the n th disintegration lies inside some disk of radius r_n .

Though we will not prove it here, it may be shown that under the conditions on f and D , the limit exists and is independent of the way one disintegrates D into finitely many tiles, as long as each of the tiles satisfies the same conditions we have imposed on D itself. We shall only make the following remarks about the conditions imposed on D and the disintegration.

(1) The point of having a piecewise differentiable boundary is that this condition ensures that when h is very small, an overwhelming percentage of the tiles lie in the interior of D .

(2) The point about the assumption on the diameters of the tiles is this: Suppose for simplicity that our integrand f is continuously differentiable on \mathbb{R}^2 ; (This will be the case in almost all of the problems we consider here.) Since D is closed and bounded, and since $\|\nabla f(\mathbf{x})\|$ is continuous, there is an $\mathbf{x}_0 \in D$ so that

$$M := \|\nabla f(\mathbf{x}_0)\| \geq \|\nabla f(\mathbf{x})\| \quad \text{for all } \mathbf{x} \in D.$$

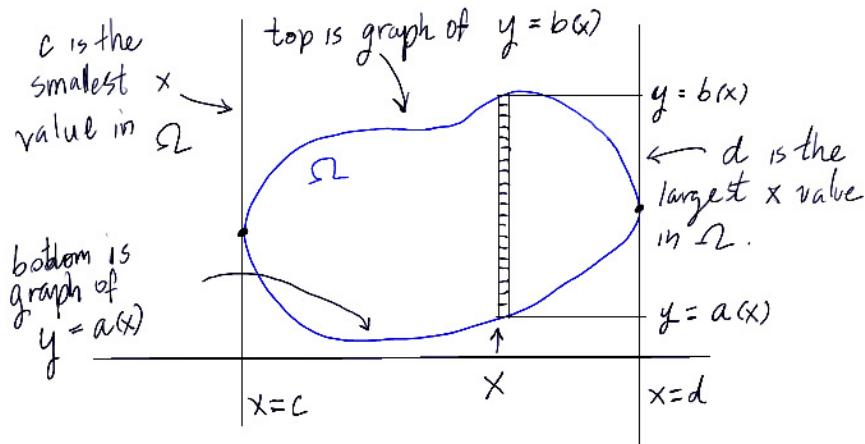
Then, if \mathbf{x} and \mathbf{y} are any two points in the same tile of diameter r , $\|\mathbf{x} - \mathbf{y}\| \leq r$, and so, by the Fundamental Theorem of Calculus, The Chain Rule, and the Cauchy-Schwarz inequality

$$|f(\mathbf{x}) - f(\mathbf{y})| = \left| \int_0^1 \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \cdot (\mathbf{y} - \mathbf{x}) dt \right| \leq \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\| \|\mathbf{y} - \mathbf{x}\| dt \leq Mr.$$

Thus, for any $\epsilon > 0$, there is an $r > 0$ so that if all tiles in a disintegration have a diameter no greater than r , then the difference between the values of f at any two points in any tile is no greater than ϵ . This is what allows us to regard f as constant on the tiles; in the limit this is exactly correct.

7.1.3 Computing area integrals

We now have a definition of area integrals. The definition does not restrict us to rectangular tiles, but as our first order of business, let us systematize our approach to using such tiles. Such an approach is often efficient when the region D is bounded above by a curve $y = b(x)$ and below by a curve $y = a(x)$, and lies between the vertical lines $x = c$ and $x = d$. The following diagram shows D , and the tiles in the column above x .



Suppose that, as in this diagram, every vertical line “slices” D in a single line segment (or misses it altogether). That is, the vertical line through x intersects D in an interval $[a(x), b(x)]$ of y values, or else is empty. If D were more complicated, the intersection could consist of several intervals, or

worse. But for now, let us consider this nice case. Then, using a rectangular disintegration, as above, and summing over columns first, we are led to the following formula for $\int_D f(x, y) dx dy$:

$$\int_D f(x, y) dx dy = \int_c^d \left(\int_{a(x)}^{b(x)} f(x, y) dy \right) dx . \quad (7.7)$$

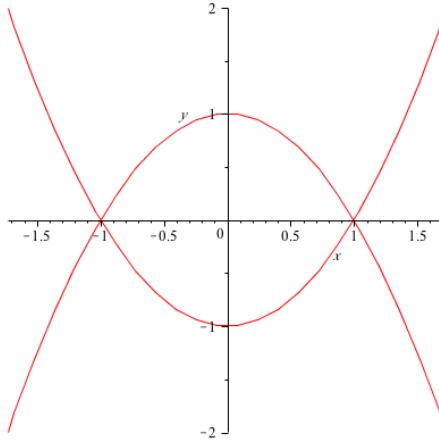
In the inner integral, x is just a parameter, not a variable, so that this integral is a garden variety integral in the single variable y . Once it is done, y is eliminated, and what remains is a garden variety integral in the single variable x . Do that, and you are done.

Example 105 (Computation of an area integral). *Let D be the region bounded above by the parabola $y = 1 - x^2$, and below by the parabola $y = x^2 - 1$. Let $f(x, y) = x^2 + 2xy$. Let us compute $\int_D f(x, y) dA$.*

The very first thing to do in such a problem is to make a sketch of D . Translating the verbal information above into a system of inequalities, we have

$$x^2 - 1 \leq y \leq 1 - x^2 .$$

Here is a plot of the bounding parabolas, which are $y = x^2 - 1$ and $y = 1 - x^2$:



The region D is the region between the two parabolas. Notice that every vertical line intersects D in a single segment or else the empty set, and so we can use (7.7) and the disintegration of D into little rectangular blocks. We only need to determine c and d , and $a(x)$ and $b(x)$.

Since the two parabolas meet at $x = \pm 1$, so $c = -1$ is the smallest x value in D , and $d = 1$ is the largest x value in D . The upper part of the boundary is $y = 1 - x^2$, so we take $a(x) = 1 - y^2$. The lower part of the boundary is $y = x^2 - 1$, so we take $b(x) = x^2 - 1$.

Hence, (7.7) becomes

$$\int_D f(x, y) dx dy = \int_{-1}^1 \left(\int_{x^2-1}^{1-x^2} (x^2 + 2xy) dy \right) dx . \quad (7.8)$$

In the inner integral, x is a parameter, and y is the variable of integration. So we treat x as a constant and have

$$\int_{x^2-1}^{1-x^2} (x^2 + 2xy) dy = (x^2 y + xy^2) \Big|_{x^2-1}^{1-x^2} = 2x^2(1 - x^2) .$$

Now (7.8) reduces to

$$\int_D f(x, y) dA = \int_{-1}^1 2x^2(1 - x^2) dx = \frac{8}{15} .$$

We get another integration formula by summing over rows first instead of columns. Doing the sum in (7.6) by summing over rows first amounts to interchanging the roles of x and y so that we have the alternate formula

$$\int_D f(x, y) dx dy = \int_c^d \left(\int_{a(y)}^{b(y)} f(x, y) dx \right) dy , \quad (7.9)$$

provided each horizontal line intersects D in a single line segment (or not at all). This time, c is the smallest y value in D , and d is the largest y value in D , and for values of y in between, the intersection of D with the horizontal line through y is the line segment corresponding to the interval $[a(y), b(y)]$ of x values.

In the inner integral, y is just a parameter, not a variable, so that this integral is a garden variety integral in the single variable x . Once it is done, x is eliminated, and what remains is a garden variety integral in the single variable y . Do that, and you are done.

Example 106 (Alternate computation of an area integral). *Let D be the region bounded above by the parabola $y = x^2 - 1$, and below by the parabola $y = x^2 - 1$. Let $f(x, y) = x^2 + 2xy$. Let's compute $\int_D f(x, y) dx dy$, but this time by integrating first in x . We can do this using (7.9) since every horizontal line intersects D in a single segment. We only need to determine c and d , and $a(y)$ and $b(y)$.*

For values of y with $0 \leq y \leq 1$, the interval is given by the equation for the upper parabola, and for values of y with $-1 \leq y \leq 0$, the interval is given by the equation for the lower parabola. Hence we break the region into two pieces, the upper region D_u and the lower region D_ℓ . It is clear from the definition that

$$\int_D f(x, y) dA = \int_{D_u} f(x, y) dA + \int_{D_\ell} f(x, y) dA ,$$

so we just need to compute these separately.

In D_ℓ , the endpoints of the segment obtained by slicing the region horizontally at height y are given by the equation $y = x^2 - 1$. Solving for x , we find $x = \pm\sqrt{1+y}$. Hence in D_ℓ we have

$$-\sqrt{1+y} \leq x \leq \sqrt{1+y} .$$

Therefore, we take $a(y) = -\sqrt{1+y}$ and $b(y) = \sqrt{1+y}$, and clearly $c = -1$ and $d = 0$. Then (7.9) gives us

$$\int_{D_u} f(x, y) dA = \int_{-1}^0 \left(\int_{-\sqrt{1+y}}^{\sqrt{1+y}} (x^2 + 2xy) dx \right) dy .$$

Doing the inner integral, treating y as constant,

$$\int_{-\sqrt{1+y}}^{\sqrt{1+y}} (x^2 + 2xy) dx = \left(\frac{x^3}{3} + x^2 y \right) \Big|_{-\sqrt{1+y}}^{\sqrt{1+y}} = \frac{2}{3}(1+y)^{3/2} .$$

Hence

$$\int_{D_u} f(x, y) dA = \int_{-1}^0 \frac{2}{3}(1+y)^{3/2} dy = \frac{4}{15} .$$

For the upper region, the endpoints of the segment obtained by slicing the region horizontally at height y are given by the equation $y = 1 - x^2$. solving for x , we find $x = \pm\sqrt{1-y}$. Hence in D_u we have

$$-\sqrt{1-y} \leq x \leq \sqrt{1-y} .$$

Hence we take $a(y) = -\sqrt{1-y}$ and $b(y) = \sqrt{1-y}$, and clearly $c = 0$ and $d = 1$. Then (7.9) gives us

$$\int_{D_u} f(x, y) dA = \int_0^1 \left(\int_{-\sqrt{1-y}}^{\sqrt{1-y}} (x^2 + 2xy) dx \right) dy .$$

Doing the inner integral, treating y as constant,

$$\int_{-\sqrt{1-y}}^{\sqrt{1-y}} (x^2 + 2xy) dx = \left(\frac{x^3}{3} + x^2 y \right) \Big|_{-\sqrt{1-y}}^{\sqrt{1-y}} = \frac{2}{3}(1-y)^{3/2} .$$

Hence

$$\int_{D_u} f(x, y) dA = \int_0^1 \frac{2}{3}(1-y)^{3/2} dy = \frac{4}{15} .$$

Finally, we have

$$\int_D f(x, y) dA = \frac{4}{15} + \frac{4}{15} = \frac{8}{15} ,$$

which is what we found before.

We get the same value both ways – of course – but notice that the first way was easier.

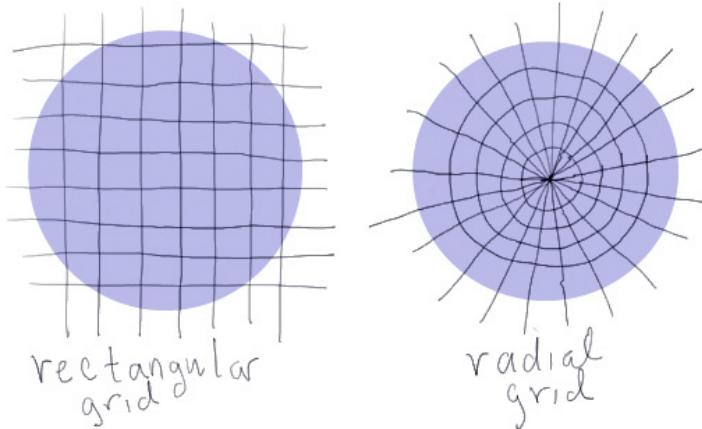
- How much calculation one has to in order to evaluate an integral depends very much on how one goes about the the disintegration and integration processes.

Both disintegration and integration involve choices – how do we slice? Do we add up columns first, or rows? So far we have only discussed slicing the region D into rectangles, but there are many other choices to consider. And as we have seen, the order in which we choose to integrate the variables will affect the amount of work we must do.

7.1.4 Polar coordinates

How would you cut a cake? That would probably depend on the shape of the cake. If the cake were rectangular, cutting it into square or rectangular slices would seem sensible. But if it were round, you would probably cut it into wedges. Making cuts along the radii, it is easy to divide a round cake into, say, a dozen equal pieces. This is not so easy if you only make cuts parallel to the lines in a rectangular grid.

When we are disintegrating a region D in \mathbb{R}^2 , it can be quite advantageous, for some of the same reasons, to slice using a grid of radii and concentric circles:

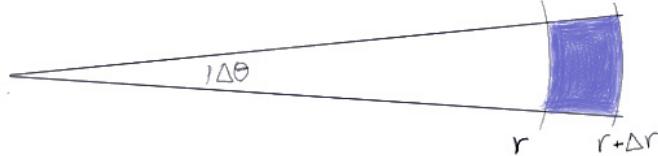


The basic formula that defines the integral is

$$\int_D f(x, y) dx dy = \lim_{\substack{\text{tile diameter} \rightarrow 0 \\ \text{little tiles}}} \left(\sum \text{(value of } f \text{ in the tile)} \times \text{(area of tile)} \right). \quad (7.10)$$

We can use this with tiles of *any shape that we find convenient*, provided the maximum tile diameter goes to zero in the limit. Of course, the *sine qua non* of convenience is that we have a simple formula for the area of the tiles. This is one of the things that is so attractive about rectangular tiles: The area of a rectangular tile of width Δx and height Δy is simply $\Delta x \Delta y$.

Now consider a “keystone” shaped tile that comes from a wedge of angle $\Delta\theta$, and lies between the radii r and $r + \Delta r$. What is its area?



The keystone shaped tile can be thought of as the part of the circular wedge with opening angle $\Delta\theta$ and radius $r + \Delta r$, that lies outside the circular wedge of the same angle and radius r . Subtracting the smaller wedge area from the larger, we are left with the area of the tile.

A circular wedge of opening angle θ and radius R is the fraction $\frac{\theta}{2\pi}$ of a disk of radius R . (That is how we measure angles – by the fraction of the circumference they subtend). The area of the disk is πR^2 , and hence the area of the wedge is

$$\frac{\theta}{2\pi} \pi R^2 = \frac{\theta R^2}{2}.$$

The area of our keystone is therefore the difference of the area of two wedges:

$$\frac{\Delta\theta(r + \Delta r)^2}{2} - \frac{\Delta\theta r^2}{2} = r\Delta r\Delta\theta + \frac{\Delta\theta(\Delta r)^2}{2}.$$

When both Δr and $\Delta\theta$ are very small, the second term on the right is negligible compared to the first, and hence:

$$\text{area of keystone tile} \approx r\Delta r\Delta\theta.$$

As Δr and $\Delta\theta$ diminish, the error in this approximation diminishes in the sense that it becomes a negligibly small percentage-wise compared to the main term, $r\Delta r\Delta\theta$.

Thus, the formula for the *area element* in polar coordinates is

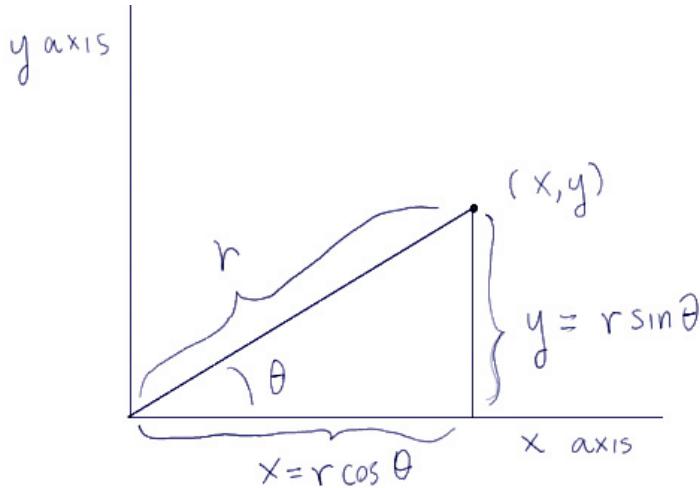
$$dA = r dr d\theta .$$

This is the area of an infinitesimal keystone tile.

The grid that we are using to cut the plane into keystone shaped tiles is based on the polar coordinate system, and we will need to be able to convert between polar coordinates – r and θ – and Cartesian coordinates – x and y – to use this slicing strategy. As you see, the keystone tiles are naturally indexed by r and θ . Therefore, it is natural to express the integrand $f(x, y)$ in these terms.

This is easy: If we measure θ counterclockwise from the positive x axis, and if r is the distance from the origin, then

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta . \quad (7.11)$$



In particular,

$$x^2 + y^2 = r^2 . \quad (7.12)$$

By definition, r is always positive. It has a geometric meaning – distance from the origin – and distances cannot be negative. (You may have worked with polar coordinates before using a different convention in which a negative value of r meant that the point would lie at distance $|r|$ from the origin in the opposite direction, namely, the one corresponding to $\theta + \pi$. There are some advantages to this convention. However, there are disadvantages as well, and these are more important here. In the examples that follow, we shall use the positivity of r several times.)

Think of (7.11) as a *dictionary* for translating Cartesian coordinates into polar coordinates. You might think we would be more interested in formulas for r and θ in terms of x and y . We do have a formula for r in terms of x and y , namely (7.12), and we could solve (7.11) for θ , but actually, what we really need is just (7.11) itself:

- To translate a function f from Cartesian into polar terms, define a new function $g(r, \theta)$ by

$$g(r, \theta) = f(r \cos \theta, r \sin \theta) .$$

The meaning of this is that if \mathbf{x} is any point in \mathbb{R}^2 , then we can evaluate f at \mathbf{x} by substituting the polar coordinates of \mathbf{x} into g .

Example 107 (Translating a function into polar terms). Let $f(x, y) = x^2y$. Then

$$g(r, \theta) = (r \cos \theta)^2 r \sin \theta = r^3 \cos^2 \theta \sin \theta.$$

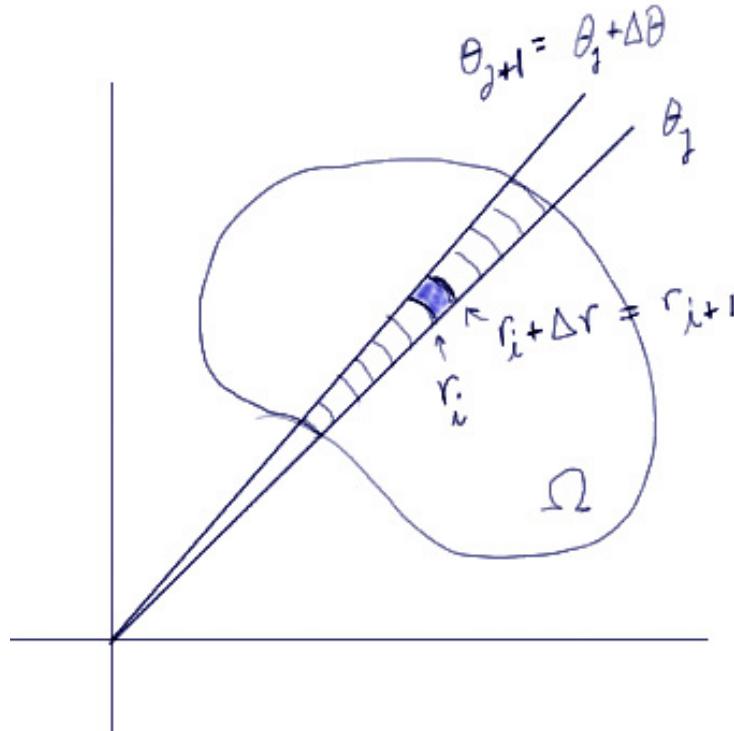
Let us apply our considerations to the problem of evaluating $\int_D f(x, y) dA$. If we cut D into keystone shaped tiles using polar coordinates, and then want to compute

$$\sum_{\text{little tiles}} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}),$$

we must choose an order in which to add up the contributions from each tile. The one that is most often convenient is to add up all of the contributions from each wedge, and then to add up the subtotals for each wedge.

The following diagram shows a region D , with the wedge cut through D by the radii at θ_j and θ_{j+1} , where some small $\Delta\theta$ has been fixed, and $\theta_j = j\Delta\theta$. For example, suppose we choose some large integer N , and let $\Delta\theta = 2\pi/N$, so that we divide \mathbb{R}^2 into N wedges with opening angle $\Delta\theta$, so that $\theta_j = j\theta/N$.

This wedge has been further broken up into keystone tiles by cutting along circular arc of radius r_i where some small value of Δr has been chosen and $r_i = i\Delta r$.



We now organize our summation as follows: For each j , we hold j fixed, and sum up the contributions from each of the tiles in the j th wedge. That is, we sum over i first, holding j fixed. Then we add up these subtotals into the grand total by summing on j :

$$\begin{aligned} \sum_{\text{little tiles}} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) &= \sum_{j=1}^N \left(\sum_{\text{little tiles in wedge } j} g(r_i, \theta_j) r_i \Delta r \Delta \theta \right) \\ &= \sum_{j=1}^N \left(\sum_{\text{little tiles in wedge } j} g(r_i, \theta_j) r_i \Delta r \right) \Delta \theta , \end{aligned}$$

where $r_i = i\Delta r$ is the i th value of r used in our grid.

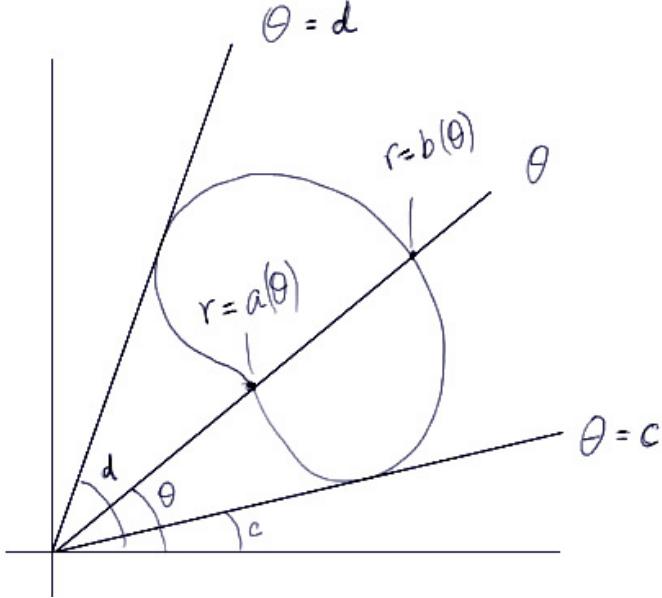
Notice that the inner sum,

$$\sum_{\text{little tiles in wedge } j} g(r_i, \theta_j) r_i \Delta r$$

is just the Riemann sum for an integral. If $a(\theta_j)$ is the smallest value of r in D that lies in the j th wedge, and if $b(\theta_j)$ is the largest value of r in D that lies in the j th wedge, then this is a Riemann sum for

$$\int_{a(\theta_j)}^{b(\theta_j)} g(r, \theta_j) r dr . \quad (7.13)$$

Here is a diagram showing $a(\theta)$ and $b(\theta)$.



The diagram also shows the smallest and largest values of θ for which the ray in direction θ intersects the region D . These are denoted c and d . Clearly $a(\theta)$ and $b(\theta)$ are only defined for $c \leq \theta \leq d$.

The value of the integral (7.13) depends on θ_j of course. For $c \leq \theta_j \leq d$, call it $G(\theta_j)$. There are no keystones to worry about for other values of θ_j , so our sum reduce to

$$\begin{aligned} \sum_{\text{little tiles}} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) &= \\ &\sum_{j \text{ such that } c \leq \theta_j \leq d} G(\theta_j) \Delta \theta , \quad (7.14) \end{aligned}$$

and this is a Riemann sum for $\int_c^d G(\theta) d\theta$. Altogether, we have the formula

$$\int_D f(x, y) dA = \int_c^d \left(\int_{a(\theta)}^{b(\theta)} g(r, \theta) r dr \right) d\theta .$$

Example 108 (An integral in polar coordinates). Let $f(x, y) = x^2$, and let D be the region bounded by the circle

$$(x - 1)^2 + y^2 = 1 . \quad (7.15)$$

Compute $\int_D f(x, y) dA$.

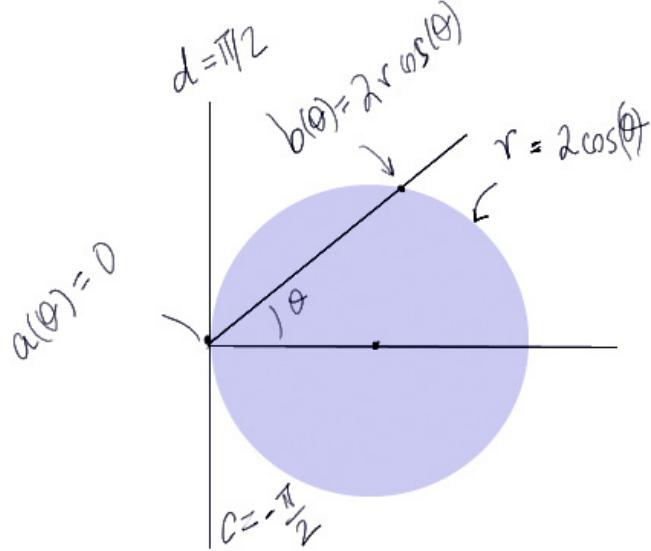
The region is a circle, and though it is not centered, we might expect it to have a nice description in polar coordinates. Let us see. Simplifying, the equation reduces to

$$x^2 + y^2 = 2x .$$

Using (7.11), this translates into $r^2 = 2r \cos \theta$. Since r is strictly positive except at the origin, (7.15) reduces to

$$r = 2 \cos \theta \quad (7.16)$$

This equation is very simple, and will enable us to find simple expressions for $a(\theta)$ and $b(\theta)$, and also c and d . To do this, draw a diagram, and label the boundary of D with the equation that specifies it:



Notice that the formula (7.16) would give a negative value for r in the second and third quadrants, but has a positive value in the first and fourth quadrants. This tells us that the region D “lives” in the first and fourth quadrants, and $c = -\pi/2$ and $d = \pi/2$. You see this also in the picture, but drawing a picture is not always so easy. Hence it is important to see how the values of c and d can be read off of (7.16).

As for $a(\theta)$ and $b(\theta)$, draw in a ray at angle θ , as in the diagram. It enters D at $r = 0$, and leave through the boundary with the equation $r = 2 \cos(\theta)$. Hence $a(\theta) = 0$, and $b(\theta) = 2 \cos(\theta)$. That takes care of the limits. The rest is easy.

Translating the integrand using (7.11),

$$g(r\theta) = (r \cos \theta)^2 = r^2 \cos^2 \theta .$$

Therefore,

$$\begin{aligned} \int_D f(x, y) dA &= \int_{-\pi/2}^{\pi/2} \left(\int_0^{2 \cos(\theta)} r^2 \cos^2 \theta r dr \right) d\theta \\ &= \int_{-\pi/2}^{\pi/2} \left(\int_0^{2 \cos \theta} r^3 dr \right) \cos^2 \theta d\theta \\ &= \int_{-\pi/2}^{\pi/2} \left(\frac{2^4 \cos^4 \theta}{4} \right) \cos^2 \theta d\theta \\ &= 4 \int_{-\pi/2}^{\pi/2} \cos^6 \theta d\theta . \end{aligned}$$

The problem is now reduced to a single variable integral with explicit limits, and for our purposes, the problem is solved. We will regard all such integrals as “trivial” for the purposes of this course. It is a non-trivial matter to make the reduction to such a single variable integral, and once you have done this, computer programs can do the rest, and give you the numerical value, which in this case is $\frac{5\pi}{4}$. But computer programs cannot make the reduction to the single variable problem.

Example 109 (Area enclosed by the Bernoulli lemniscate). The Bernoulli lemniscate is the “infinity symbol” curve given by

$$(x^2 + y^2)^2 = 2(x^2 - y^2) . \quad (7.17)$$

Let us compute the enclosed area, which is

$$\int_D 1 dA .$$

That is, to get an area, the integrand should just be 1. (Reflect on the definition to make sure this is clear).

Since the integrand features $x^2 + y^2$, which will reduce to r^2 in polar coordinates, we will translate (7.17) into polar terms, hoping for something nice. As it stands, (7.17) is pretty awful. It is a quartic equation, and solving to find either x as a function of y , or y as a function of x , is a daunting proposition: It can be done, but is big mess. We do not like messes, big or small. Let us try something else: perhaps polar coordinates avoid a mess; let us try.

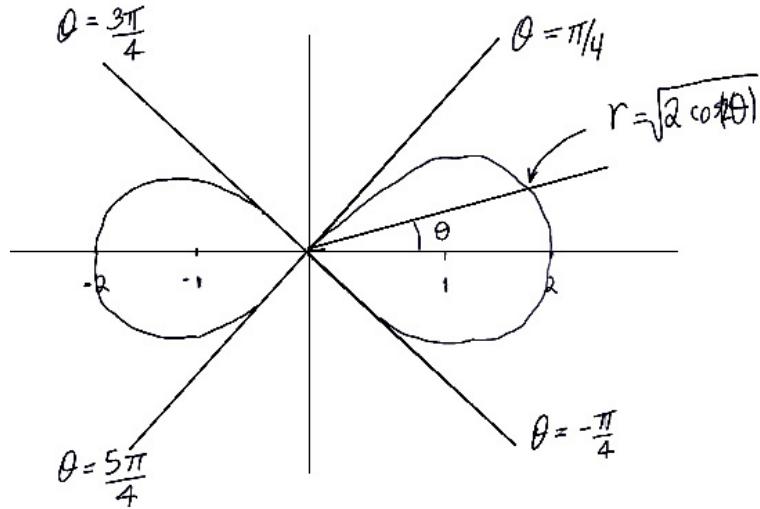
Using (7.11), (7.17) becomes $r^4 = 2r^2(\cos^2 \theta - \sin^2 \theta)$. Then using the double angle formulas, and dividing through by r^2 , (7.17) reduces to

$$r^2 = 2 \cos(2\theta) . \quad (7.18)$$

This is progress: The variables are separated, and you can also see from (7.18) that the right hand side is negative unless

$$-\pi/4 \leq \theta \leq \pi/4 \quad \text{or} \quad 3\pi/4 \leq \theta \leq 5\pi/4 .$$

Hence the curve described by (7.18), or equivalently (7.17), “lives” in these two angular sectors. Here is a rough sketch:



You could produce such a sketch by evaluating $r = \sqrt{2 \cos(2\theta)}$ for a few values of θ in the range $-\pi/4 \leq \theta \leq \pi/4$, drawing those points in, and connecting the dots. You do not have to know in advance that our equation describes the infinity symbol.

Notice that the equation (7.17) only involves x^2 and y^2 , so if (x, y) satisfies the equation, so do the mirror image points

$$(-x, y) \quad (x, -y) \quad (-x, -y).$$

That is, we can see from the equation that the region is symmetric under reflection about the x -axis and about the y -axis. This is not so evident from the rough sketch, but that is O.K.; the equations make it clear.

Because of the symmetry, the area in the first quadrant is exactly one fourth of the total. Hence we can take $c = 0$ and $d = \pi/4$, and remember to multiply by 4 when we have finished integrating.

From the diagram, you see that $a(\theta) = 0$ and $b(\theta) = \sqrt{2 \cos(2\theta)}$, so the integral we need to do is

$$\int_1^{\pi/4} \left(\int_0^{\sqrt{2 \cos(2\theta)}} 1 r dr \right) d\theta.$$

The inner integral is trivial, and we are left with $\int_1^{\pi/4} \cos(2\theta) d\theta = 1/2$. Multiplying by 4, the area is 2.

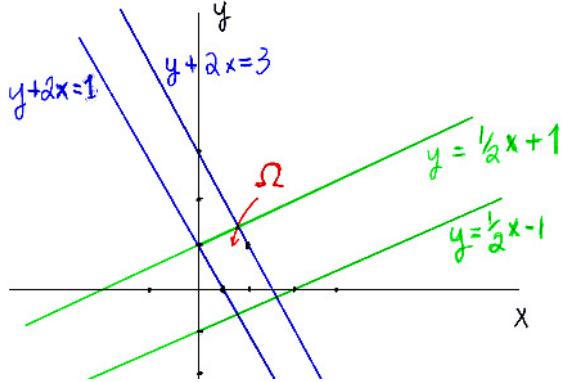
7.2 Jacobians and changing variables of integration in \mathbb{R}^2

7.2.1 Letting the boundary of D determine the disintegration strategy

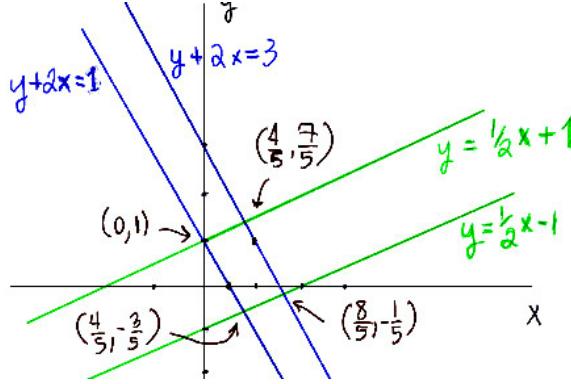
Consider the problem of computing $\int_D f(x, y) dA$ where $f(x, y) = y$, and where D is the open rectangle bounded by the 4 lines

$$y + 2x = 1 \quad y + 2x = 3 \quad 2y - x = -2 \quad 2y - x = 2. \quad (7.19)$$

Here is a picture of the region:



To find the limits of integration, we next work out the coordinates of the vertices by solving the systems of equations for each pair of crossing lines:



If we integrated in x first we would need to break D in the three separate subregions for

$$-\frac{3}{5} \leq y \leq -\frac{1}{5} \quad -\frac{1}{5} \leq y \leq 1 \quad 1 \leq y \leq \frac{7}{5}$$

since in each of these regions we need a different formula for $a(y)$ or $b(y)$ – horizontal segments at height y begin and end on the same bounding line in only when y stays in one of these ranges.

If we integrated in y first, we could do better: We would only need to break D into the two separate subregions for

$$0 \leq x \leq \frac{4}{5} \quad \frac{4}{5} \leq x \leq \frac{8}{5}$$

since in each of these regions we need a different formula for $a(x)$ or $b(x)$ – vertical segments at x begin and end on the same bounding line only when x stays in one of these ranges.

So, if these were our only choices, certainly we would integrate in y first. However, there is something better we can do. Instead of disintegrating D using a grid composed of lines parallel to the axes, let's disintegrate D using a grid of lines parallel to the bounding lines.

To do this, define new variables

$$u = y + 2x \quad v = 2y - x .$$

In terms of these variables, the equation for the our lines bounding D reduce to

$$u = 1 \quad u = 3 \quad v = -2 \quad v = 2 . \quad (7.20)$$

In the u, v plane, the lines bound an open rectangle with sides parallel to the axes, and we can easily divide it up along a rectangular grid. Let us call the rectangle \hat{D} .

It will be useful to think of these new coordinates as defining a *coordinate transformation*: Define

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

where

$$u(x, y) = y + 2x \quad v(x, y) = 2y - x . \quad (7.21)$$

This coordinate transformation is *linear*: Let J denote the 2×2 matrix

$$J := \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} .$$

Then

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

Since $\det(J) = 5$, the matrix J is invertible, and hence this coordinate transformation has the inverse transformation

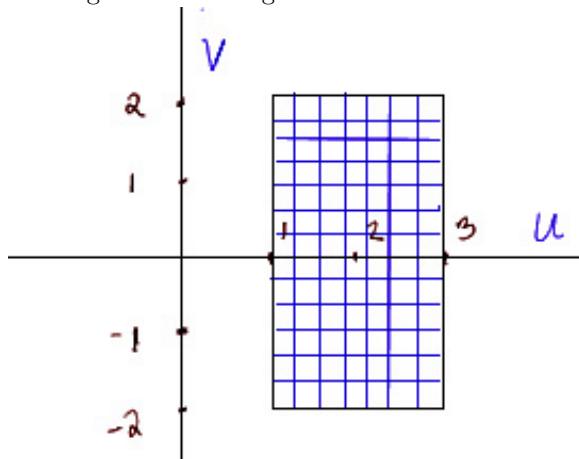
$$\mathbf{x}(u, v) = J^{-1}(u, v) = \frac{1}{5} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} (u, v) = \frac{1}{5}(2u - v, u + 2v) . \quad (7.22)$$

That is, $\mathbf{x}(u, v) = (x(u, v), y(u, v))$ where

$$x(u, v) = \frac{2u - v}{5} \quad y(u, v) = \frac{u + 2v}{5} . \quad (7.23)$$

- We will now use this coordinate transformation to “transplant” a simple and convenient disintegration of \hat{D} onto D , which is possible because the coordinate transformation has been set up as a one-to-one map from D onto \hat{D} .

Recall that \hat{D} is simply a rectangle with sides parallel to the coordinate axes, so that we may conveniently disintegrate it using a coordinate grid:



The j th vertical line in this grid is the line

$$u = 1 + j\Delta u \quad (7.24)$$

where Δu is the horizontal spacing in the grid, and the i th horizontal line in the grid is

$$v = -2 + i\Delta v \quad (7.25)$$

where Δv is the vertical spacing in the grid. (We are ordering the lines left to right and bottom to top respectively).

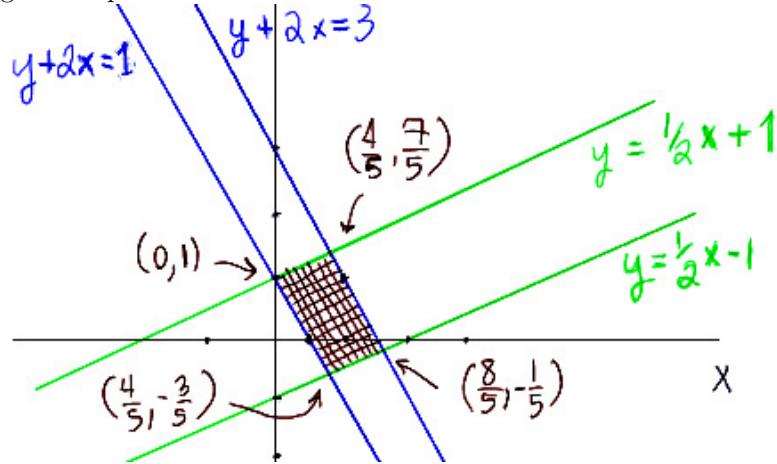
Using (7.2.1) to express (7.24) and (7.25) in terms of x and y instead of u and v we get

$$y + 2x = 1 + j\Delta u \quad (7.26)$$

and

$$2y - x = -2 + i\Delta v \quad (7.27)$$

This gives us two sets of parallel, evenly spaced lines in the x, y plane that divide D up into similar parallelogram shaped tiles.



This grid is our “induced” disintegration of D – it is induced by the simple rectangular disintegration of \hat{D} , and the coordinate transformation relating D and \hat{D} .

Using the tiles of this induced disintegration, we will compute

$$\int_D f(x, y) dA = \lim_{\text{tile diameter} \rightarrow 0} \left(\sum_{\text{little tiles}} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) \right). \quad (7.28)$$

Each of these tiles in D corresponds to a tile in the u, v plane, and so we can enumerate the tiles in our induced disintegration of D using an enumeration of the corresponding tiles in our disintegration of the rectangle $1 \leq u \leq 3, -2 \leq v \leq 2$ is the u, v plane. To compute the integral using these coordinates, we need to answer two questions:

- Given a tile with $u_j \leq u \leq u_j + \Delta u$ and $v_i \leq v \leq v_i + \Delta v$, what is the value of $f(x, y)$ at some point (x, y) in the corresponding tile in the x, y plane?
- Given a tile with $u_j \leq u \leq u_j + \Delta u$ and $v_i \leq v \leq v_i + \Delta v$, what is the area of the corresponding tile in the x, y plane?

It is easy to answer the first question, using the inverse coordinate transformation $\mathbf{x}(u, v) = (x(u, v), y(u, v))$ given in (7.23). Since (u_j, v_i) is in the j, i th tile in the disintegration of \hat{D} , $\mathbf{x}(u_j, v_i) =$

$(x(u_j, v_i), y(u_j, v_i))$ is in the j, i th tile in the induced disintegration of D . Therefore,

$$f(x(u_j, v_i), y(u_j, v_i))$$

is a representative value for f in the j, i th tile of the induced disintegration of D . (And since f is nearly constant on this small tile it is a good a representative as any other.)

We turn to the second question, concerning the area of the tiles. The tiles in the x, y plane are the images of tiles in the u, v plane under the linear transformation in (7.22). The j, i th tile is the parallelogram in the x, y plane with vertices

$$\begin{aligned} \mathbf{x}(u_j, v_i) &= \mathbf{x}(u_j, v_i) \\ \mathbf{x}(u_j + \Delta u, v_i) &= \mathbf{x}(u_j, v_i) + \Delta u J^{-1} \mathbf{e}_1 \\ \mathbf{x}(u_j, v_i + \Delta v) &= \mathbf{x}(u_j, v_i) + \Delta v J^{-1} \mathbf{e}_2 \\ \mathbf{x}(u_j + \Delta u, v_i + \Delta v) &= \mathbf{x}(u_j, v_i) + \Delta u J^{-1} \mathbf{e}_1 + \Delta v J^{-1} \mathbf{e}_2. \end{aligned} \quad (7.29)$$

We know that for any vectors \mathbf{x}_0 , \mathbf{a} and \mathbf{b} in \mathbb{R}^2 , the area of the parallelogram with vertices at \mathbf{x}_0 , $\mathbf{x}_0 + \mathbf{a}$, $\mathbf{x}_0 + \mathbf{b}$ and $\mathbf{x}_0 + \mathbf{a} + \mathbf{b}$ is given by $\det([\mathbf{a}, \mathbf{b}])$. Therefore, the area of the tile with the vertices given in (7.29) is

$$\Delta u \Delta v \det([J^{-1} \mathbf{e}_1, J^{-1} \mathbf{e}_2]) = \Delta u \Delta v \det(J^{-1}) = \frac{\Delta u \Delta v}{5}.$$

Notice that this area is the same for all tiles, independent if j and i .

With our two questions answered, going back to (7.28), we now have

$$\begin{aligned} \int_D f(x, y) dA &= \lim_{\Delta u, \Delta v \rightarrow 0} \left(\sum_{i,j} (g(u_j, v_i) \times \left(\frac{1}{5} \Delta u \Delta v \right)) \right) \\ &= \lim_{\Delta u, \Delta v \rightarrow 0} \left(\sum_i \left(\sum_j \frac{1}{5} (g(u_j, v_i) \Delta u) \right) \Delta v \right). \end{aligned} \quad (7.30)$$

You recognize the Riemann sums for

$$\int_{-2}^2 \left(\int_1^3 \frac{1}{5} g(u, v) du \right) dv.$$

In the case at hand, $g(u, v) = (2v + u)/5$, and so

$$\begin{aligned} \int_D f(x, y) dA &= \frac{1}{25} \int_{-2}^2 \left(\int_1^3 (2v + u) du \right) dv \\ &= \frac{1}{25} \int_{-2}^2 \left(2vu + \frac{u^2}{2} \Big|_{u=1}^{u=3} \right) dv \\ &= \frac{1}{25} \int_{-2}^2 (4v + 4) dv = \frac{16}{25}. \end{aligned}$$

What is the lesson to be drawn from this example? It is that:

- By using a disintegration scheme that “respects” the equations defining the boundaries of D , we were able to avoid breaking up D into subregions that would have to be handled separately, and we got very simple limits of integration – constants in this case.

We have encountered a very useful formula in the last example: The formula for the “*magnification factor*” of a linear transformation.

Theorem 83 (The determinant as a magnification factor). *Let A be any 2×2 matrix. Consider the corresponding linear transformation from \mathbb{R}^2 to \mathbb{R}^2 as a linear transformation from the u, v plane to the x, y plane. Let \widehat{D} be any closed, bounded set in \mathbb{R}^2 whose boundary is a piecewise differentiable curve, and let D be its image under the linear transformation. That is, let D be the set of points in the x, y plane of the form $A(u, v)$ with $(u, v) \in \widehat{D}$. Then D is also a closed and bounded set in \mathbb{R}^2 whose boundary is a piecewise differentiable curve, and*

$$\text{area}(D) = |\det(A)|\text{area}(\widehat{D}) .$$

In short, the linear transformation corresponding to A magnifies the area of sets nice enough to have a well-defined area by a factor of $|\det(A)|$. (Note that it may be the case that $|\det(A)| < 1$, in which case the “magnification” amounts to “shrinking”).

The proof of Theorem 83 is essentially a recapitulation of a calculation we have made in the last example.

Proof of Theorem 83: Let us disintegrate the region \widehat{D} into little square tiles with sides parallel to the u, v coordinate axes, and with side length h . Then the tile with the lower left hand corner at \mathbf{u} has vertices

$$\mathbf{u}, \quad bu_h\mathbf{e}_1, \quad \mathbf{u} + h\mathbf{e}_2 \quad \text{and} \quad \mathbf{u} + h\mathbf{e}_1 + h\mathbf{e}_2 .$$

The area of this triangle is of course h^2 , and the points in the tile are exactly the points of the form

$$\mathbf{u} + s\mathbf{e}_1 + t\mathbf{e}_2 \quad 0 \leq s, t \leq h .$$

The image of this tile under the linear transformation given by A is the set of points of the form

$$A\mathbf{u} + sA\mathbf{e}_1 + tA\mathbf{e}_2 \quad 0 \leq s, t \leq h ,$$

since $A(\mathbf{u} + s\mathbf{e}_1 + t\mathbf{e}_2) = A\mathbf{u} + sA\mathbf{e}_1 + tA\mathbf{e}_2$. Let us write $A = [\mathbf{v}_1, \mathbf{v}_2]$ so that $A\mathbf{e}_1 = \mathbf{v}_1$ and $A\mathbf{e}_2 = \mathbf{v}_2$. Then the transformed tile is the parallelogram with vertices

$$A\mathbf{u}, \quad A\mathbf{u} + h\mathbf{v}_1, \quad A\mathbf{u} + h\mathbf{v}_2 \quad \text{and} \quad A\mathbf{u} + h\mathbf{v}_1 + h\mathbf{v}_2 .$$

The area of this parallelogram is

$$|\det([h\mathbf{v}_1, h\mathbf{v}_2])| = |h^2 \det([\mathbf{v}_1, \mathbf{v}_2])| = h^2 |\det(A)| .$$

Thus for each square tile in our disintegration of \widehat{D} , the corresponding tile in the induced disin-

tegration of D has an area that is exactly $|\det(A)|$ times as large. Thus

$$\begin{aligned}\text{area}(D) &= \int_D 1 \, dA \\ &= \lim_{\text{tile diameter} \rightarrow 0} \left(\sum_{\text{little tiles}} (\text{area of tile in } D) \right) \\ &= \lim_{\text{tile diameter} \rightarrow 0} \left(\sum_{\text{little tiles}} |\det(A)|(\text{area of tile in } \hat{D}) \right) \\ &= |\det(A)| \left(\lim_{\text{tile diameter} \rightarrow 0} \left(\sum_{\text{little tiles}} |\det(A)|(\text{area of tile in } \hat{D}) \right) \right) \\ &= |\det(A)| \int_{\hat{D}} 1 \, dA = |\det(A)| \text{area}(\hat{D}) .\end{aligned}$$

7.2.2 The change of variables formula for integrals in \mathbb{R}^2

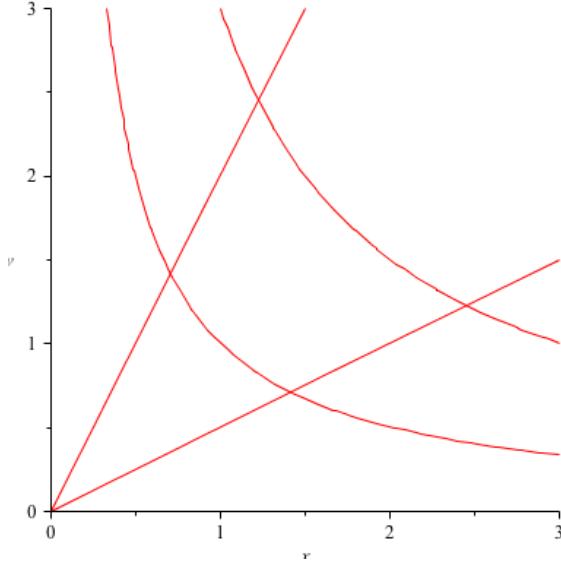
The strategy developed in the last subsection can be applied even when the boundary of D is not given by straight lines. There is very little adaptation required if we remember the main idea of the Differential Calculus: *Up close enough, all nice functions are linear for all practical purposes.* Let us consider an example.

Example 110 (The boundary determines the coordinates). *Consider the open set D in the upper right quadrant bounded by*

$$xy = 1 \quad xy = 3 \quad 2x = y \quad x = 2y .$$

Let us compute the area of D .

The first step is to sketch a plot of the bounding curves:



Two of the bounding curves are arcs of hyperbolae, and the other two are lines. However, notice that if we introduce

$$u = xy \quad \text{and} \quad v = y/x , \tag{7.31}$$

we can write the equations for the boundary as

$$u = 1 \quad u = 3 \quad v = 2 \quad v = 1/2 .$$

Again, these four lines bound a coordinate rectangle rectangle in the u, v plane.

To proceed, we use use (7.31) to define a nonlinear coordinate transformation $\mathbf{u}(x, y)$ by

$$\mathbf{u}(x, y) = (u(x, y), v(x, y)) = (xy, y/x) , \quad (7.32)$$

which is defined on $\{(x, y) : x \neq 0\} \subset \mathbb{R}^2$, which includes the set D .

Think of (7.32) as defining a transformation from the x, y plane to the u, v plane. Let \widehat{D} denote the open set in the u, v plane given by

$$1 < u < 3 \quad \text{and} \quad 1/2 < v < 2 \quad (7.33)$$

Then $\mathbf{u}(x, y)$ transforms D onto \widehat{D} in a one-to-one manner. As in the last subsection, we shall need the inverse transformation $\mathbf{x}(u, v)$. The inverse transformation transfers the rectangular grid disintegration of \widehat{D} into a convenient disintegration of D .

To obtain the inverse transformation $\mathbf{x}(u, v)$, all we have to do is to solve (7.31) for x and y as functions of u and v . From (7.23), we see that $uv = y^2$. Since D is in the upper right quadrant, $y > 0$, and so $y = \sqrt{uv}$. Next, $x^2 = u/v$, and since $x > 0$, $x = \sqrt{u/v}$. This gives us

$$x = \sqrt{u/v} \quad \text{and} \quad y = \sqrt{uv} \quad (7.34)$$

Thus we obtain

$$\mathbf{x}(u, v) = (\sqrt{u/v}, \sqrt{uv}) .$$

Now consider a small tile with $u_j \leq u \leq u_j + \Delta u$ and $v_i \leq v \leq v_i + \Delta v$ in the u, v plane. The image of this tile under the transformation $\mathbf{x}(u, v)$ is a slightly distorted parallelogram in the x, y plane with vertices at

$$\mathbf{x}(u_j, v_i) \quad \mathbf{x}(u_j + \Delta u, v_i) \quad \mathbf{x}(u_j, v_i + \Delta v) \quad \mathbf{x}(u_j + \Delta u, v_i + \Delta v) .$$

The distortion will be slight to the extent that Δu and Δv are small – everything nice looks linear up close enough.

To compute the area of this parallelogram, we first apply the approximation

$$\mathbf{x}(\mathbf{u}) \approx \mathbf{x}(\mathbf{u}_0) + [D\mathbf{x}(\mathbf{u}_0)](\mathbf{u} - \mathbf{u}_0)$$

with the basepoint $\mathbf{u}_0 = (u_j, v_i)$, which is the lower left vertex of the tile in the u, v plane. We have:

$$\begin{aligned} \mathbf{x}(u_j, v_i) &= \mathbf{x}(\mathbf{u}_0) \\ \mathbf{x}(u_j + \Delta u, v_i) &\approx \mathbf{x}(\mathbf{u}_0) + [D\mathbf{x}(\mathbf{u}_0)](\Delta u, 0) \\ \mathbf{x}(u_j, v_i + \Delta v) &\approx \mathbf{x}(\mathbf{u}_0) + [D\mathbf{x}(\mathbf{u}_0)](0, \Delta v) \\ \mathbf{x}(u_j + \Delta u, v_i + \Delta v) &\approx \mathbf{x}(\mathbf{u}_0) + [D\mathbf{x}(\mathbf{u}_0)](\Delta u, \Delta v) \end{aligned}$$

In this approximation, the parallelogram is the image of the rectangle with vertices

$$(0, 0) \quad (\Delta u, 0) \quad (0, \Delta v) \quad (\Delta u, \Delta v)$$

under the linear transformation induced by $[D\mathbf{x}(\mathbf{u}_0)]$, and then translated by $\mathbf{x}(\mathbf{u}_0)$.

Translation has no affect on area, and by Theorem 83, the linear transformation multiplies the area of the original rectangle, namely $\Delta u \Delta v$, by the factor $|\det[D\mathbf{x}(\mathbf{u}_0)]|$. Therefore, using the notation introduced above:

- The image under \mathbf{x} of the tile with $u_j \leq u \leq u_j + \Delta u$ and $v_i \leq v \leq v_i + \Delta v$ is a tile in the x, y plane whose area is

$$|\det[D\mathbf{x}(\mathbf{u}_0)]| \Delta u \Delta v$$

up to an error that is negligibly small percentage-wise as Δu and Δv both go to zero.

Everything is pretty much as it was in our last example, except that now $|\det[D\mathbf{x}(\mathbf{u})]|$ is not a constant. Computing, we find

$$[D\mathbf{x}(\mathbf{u})] = \frac{1}{2} \begin{bmatrix} u^{-1/2}v^{-1/2} & u^{-1/2}v^{-3/2} \\ u^{-1/2}v^{1/2} & u^{1/2}v^{-1/2} \end{bmatrix}$$

Therefore,

$$|\det[D\mathbf{x}(\mathbf{u})]| = \frac{1}{2uv} .$$

This gives us a formula for the area of the image of a small tile at u, v , namely

$$\frac{1}{2uv} \Delta u \Delta v .$$

This is often referred to as the formula for the area element.

In an area computation, our integrand is 1, which requires no translation. However, we can go ahead and say what we would do if the integrand were some function $f(x, y)$. We would define $g(u, v)$ by $g(\mathbf{u}) = f(\mathbf{x}(\mathbf{u}))$. The definition is such that if (x, y) corresponds to (u, v) under the transformation \mathbf{x} , then $f(x, y) = g(u, v)$.

Going back to the basic formula (7.28), we have

$$\text{area of } D = \int_D 1 dA = \lim_{\substack{\text{tile diameter} \rightarrow 0 \\ \text{little tiles}}} \left(\sum_{\text{little tiles}} 1 \times (\text{area of tile}) \right) . \quad (7.35)$$

Using the tiles induced by the transformation \mathbf{f} through the regular rectangular grid on the rectangle (7.33), we get the Riemann sums for

$$\int_{1/2}^2 \left(\int_1^3 \frac{1}{2uv} du \right) dv .$$

The two integrals are now easily worked out with the result that the area is $\ln(6)$.

What we have just worked out is a *substitution*, or *change of variables* formula for integrals in two variables.

The general picture is this: Suppose that \mathbf{x} is an invertible transformation from some open subset \widehat{D} of \mathbb{R}^2 to another open subset D of \mathbb{R}^2 . Think of \mathbf{x} as transforming (at least part of) the u, v plane to the x, y plane so that

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

Since the transformation \mathbf{x} is invertible, it sets up a one-to-one correspondence between points in \widehat{D} and points D so that any disintegration of \widehat{D} induces a disintegration of D .

Consider the image of a rectangular tile of width Δu and height Δv sitting in \widehat{D} with its lower left corner at (u, v) . As we have explained above, the area of the corresponding tile in the induced disintegration of D is well approximated by

$$|\det[D\mathbf{x}(u, v)]| \Delta u \Delta v .$$

Now let f be a continuous function on D . Then, for the tiles in the induced disintegration of D , we have

$$(\text{value of } f \text{ in the tile}) \times (\text{area of the tile}) \approx f(\mathbf{x}(u, v)) |\det[D\mathbf{x}(u, v)]| \Delta u \Delta v ,$$

where (u, v) is some point in the chosen tile in the disintegration of \widehat{D} . Therefore,

$$\int_D f(x, y) dA = \int_{\widehat{D}} f(\mathbf{x}(u, v)) |\det(D\mathbf{x}(u, v))| dA . \quad (7.36)$$

On the left, dA is the area element in the x, y plane, and on the right dA is the area element in the u, v plane. To avoid confusion, it is sometimes helpful to use another common notation, and to write $d^2\mathbf{x}$ to denote the area element in the x, y plane, and $d^2\mathbf{u}$ to denote the area element in the u, v plane.

We can use this notation to write

$$\int_D f(\mathbf{x}) d^2\mathbf{x} = \int_{\widehat{D}} f(\mathbf{x}(\mathbf{u})) |\det(D\mathbf{x}(\mathbf{u}))| d^2\mathbf{u} . \quad (7.37)$$

This may be compared to the formula for substitution, or change of variables, in one dimension. Suppose $x(u)$ is a differentiable function on \mathbb{R} . Then if f is any continuous function of one variable, we have

$$\int_a^b f(x) dx = \int_c^d f(x(u)) x'(u) du \quad (7.38)$$

$$a = x(c) \text{ and } b = x(d).$$

Notice that the determinant of the Jacobian of $\mathbf{x}(u)$ is the higher dimensional replacement for the derivative $x'(u)$ in the one dimensional formula. However, in the one dimensional formula, there is no absolute value sign. Why is this?

Suppose that $c < d$ as usual, and suppose that $x(u)$ is invertible. Even though $x(u)$ is invertible, it might be decreasing, so that $a = x(c) > x(d) = b$. In this case x' is negative, but we can cancel this minus sign with the minus sign that comes from swapping the limits on the left. In other words, if we define

$$\tilde{a} = \min\{x(c), x(d)\} \quad \text{and} \quad \tilde{b} = \max\{x(c), x(d)\}$$

so that $\tilde{a} < \tilde{b}$ and $[\tilde{a}, \tilde{b}]$ defined an interval, we could rewrite (7.38) as

$$\int_{[\tilde{a}, \tilde{b}]} f(x) dx = \int_{[c, d]} f(x(u)) |x'(u)| du \quad (7.39)$$

and now we get a formula that looks even more like (7.48).

In writing the simpler formula (7.38), we are taking advantage of the fact that the real numbers are ordered. There is no natural ordering of the points in a region of \mathbb{R}^2 , and so there is no natural analog of “switching the limits of integration”.

It is important to stress that the formula (7.38) is valid even if x is not a one-to-one function, but not so (7.39), and not so its higher dimensional analog (7.48). For example, if as u sweeps through $[c, d]$, the interval $\mathbf{x}(u)$ sweeps through the interval $[\tilde{a}, \tilde{b}]$ three times, then you would need a factor of 3 on the left in (7.39) for it to be valid. Similar rules counting the number of times the image of \tilde{D} covers D under \mathbf{x} would allow us to consider transformations that are not invertible. Here, we will only work with invertible transformations; this suffices for the solution of many practical problems.

Now that we have the change of variables formula (7.48), we can put it to work directly, without explicitly going through considerations of “little tiles”. That is not to say that the “little tiles” way of thinking is expendable in any way. Among other things, it is essential for setting up integrals that arise in word problems – the only way they arise in real life.

Let us close this subsection with some examples of (7.48) in action. We will focus on how one finds $\mathbf{x}(u)$ and hence \tilde{D} .

Actually, in practice, one is led first to a formula for $\mathbf{u}(x)$, giving u and v as functions of x and y . Usually, staring at the definition of D , we come up with some definitions of $u(x, y)$ and $v(x, y)$ in terms of which one can give a simple characterization of the set D . The first order of business then is to solve this system of equations to find x and y as functions of u and v , or, in other words, to find $\mathbf{x}(u)$.

Example 111 (Using the change of variables formula in \mathbb{R}^2). *Let D be the open set in the upper right quadrant between the curves*

$$x = \frac{1}{y^2} \quad \text{and} \quad x = \frac{4}{y^2}$$

and between the curves

$$y = x^2 \quad \text{and} \quad y = 4x^2 .$$

Lets compute $\int_D x^2 d^2\mathbf{x}$. If we define

$$u = xy^2 \quad \text{and} \quad v = y/x^2 , \quad (7.40)$$

the transformation $\mathbf{u}(x, y) = (u(x, y), v(x, y))$ transforms the the region D into the region \tilde{D} described by

$$1 \leq u \leq 4 \quad \text{and} \quad 1 \leq v \leq 4 . \quad (7.41)$$

To find the inverse transformation $\mathbf{x}(u, v)$, we solve (7.40) for x and y in terms of u and v . We eliminate x by forming $u^2v = y^5$, so $y = u^{2/5}v^{1/5}$. Next, we eliminate y by forming $uv^{-2} = x^5$, so that $x = u^{1/5}v^{-2/5}$. This gives us

$$\mathbf{x}(u, v) = (u^{1/5}v^{-2/5}, u^{2/5}v^{1/5}) .$$

With this definition of \mathbf{x} , \hat{D} is the rectangle (7.41).

Next, we compute

$$D\mathbf{x} = \frac{1}{5} \begin{bmatrix} u^{-4/3}v^{-2/5} & -2u^{1/5}v^{-7/5} \\ 2u^{-3/5}v^{1/5} & u^{2/5}v^{-4/5} \end{bmatrix}.$$

Therefore, $\det(D\mathbf{x}(\mathbf{u})) = \frac{1}{5}u^{-2/5}v^{-6/5}$. Next, with $f(x, y) = x^2$, $f(\mathbf{x}(u, v)) = u^{2/5}v^{-4/5}$. Hence, from (7.48), we have

$$\int_D f(\mathbf{x}) d^2\mathbf{x} = \int_{\hat{D}} (u^{2/5}v^{-4/5}) \frac{1}{5}u^{-2/5}v^{-6/5} d^2\mathbf{u} = \frac{1}{5} \int_{\hat{D}} v^{-2} d^2\mathbf{u}$$

and since \hat{D} is just the rectangle (7.41), this becomes

$$\frac{1}{5} \int_1^4 \left(\int_1^4 v^{-2} du \right) dv = \frac{3}{5} \int_1^4 v^{-2} du = \frac{9}{20}.$$

7.2.3 An alternative computational method

In many integration problems, we are led first to define the functions $u(x, y)$ and $v(x, y)$ which the gives us the coordinate transformation $\mathbf{u}(x, y) = (u(x, y), v(x, y))$. The next step in the procedure described above is to invert this transformation to find $\mathbf{x}(u, v) = (x(u, v), y(u, v))$, and then to compute $\det[D\mathbf{x}(u, v)]$.

It is not always possible to compute useful formulas for the inverse transformation, even when the Inverse Function Theorem guarantees us that it exists. However, it is not always necessary to do so.

What we can compute directly from a continuously differentiable transformation $\mathbf{u}(x, y) = (u(x, y), v(x, y))$ is $\det[D\mathbf{u}(x, y)]$. Wherever this is non-zero, the Inverse Function Theorem guarantees us that the inverse $\mathbf{x}(u, v)$ exists and is differentiable, and moreover,

$$[D\mathbf{x}(u, v)] = [D\mathbf{u}(\mathbf{x}(u, v))]^{-1}.$$

As we shall see in the next chapter, whenever an $n \times n$ matrix A is invertible, $\det(A^{-1}) = (\det(A))^{-1}$. This is easily checked for 2×2 matrices using the explicit formula for the inverse. Hence

$$\det[D\mathbf{x}(u, v)] = (\det[D\mathbf{u}(\mathbf{x}(u, v))])^{-1},$$

and for any $f(x, y)$,

$$f(\mathbf{x}(u, v)) \det[D\mathbf{x}(u, v)] = f(\mathbf{x}(u, v)) (\det[D\mathbf{u}(\mathbf{x}(u, v))])^{-1}. \quad (7.42)$$

That is, we do not really need to find formulas expressing x and y in terms of u and v , we only need to express the function $f(x, y)(\mathbf{x}(u, v))(\det[D\mathbf{u}(x, y)])^{-1}$ in terms of u and v . Of course, if we have formulas that express x and y in terms of u and v , we can write *any* function of x and y in terms of u and v . But we only need to write one special function in terms of u and v and this can be easier, as we see in the next examples.

Example 112. As in Example 111, let us define

$$\mathbf{u}(x, y) = (u(x, y), v(x, y)) = (xy^2, y/x^2).$$

We compute the Jacobian matrix of this transformation:

$$[D\mathbf{u}(x, y)] = \begin{bmatrix} y^2 & 2xy \\ -2y/x^3 & 1/x^2 \end{bmatrix}.$$

Then $\det[D\mathbf{u}(x, y)] = 5y^2/x^2$, and so with $f(x, y) = x^2$ as in Example 111, we have

$$f(x, y)(\det[D\mathbf{u}(x, y)])^{-1} = \frac{1}{5}x^4y^{-2}.$$

We recognize the right hand side as x^4y^{-2} as v^{-2} , and so

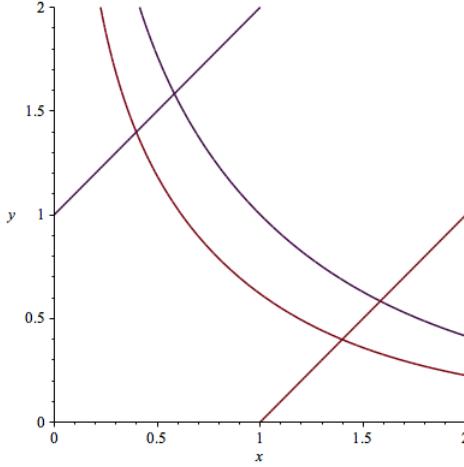
$$\int_{\tilde{D}} f(\mathbf{x}(u, v))(\det[D\mathbf{u}(\mathbf{x}(u, v))])^{-1} d^2\mathbf{u} = \frac{1}{5} \int_{\tilde{D}} v^{-2} d^2\mathbf{u} = \frac{9}{20}$$

as before.

Example 113. Let D be the region in the upper right quadrant such that

$$1 \leq x^2y + y^2x \leq 2 \quad \text{and} \quad -1 \leq y - x \leq 1.$$

here is a diagram showing the bounding curves:



We will now compute $\int_D (4xy + x^2 + y^2) d^2\mathbf{x}$. To do so, we introduce the coordinate transformation suggested by the form of the boundary:

$$\mathbf{u}(x, y) = (u(x, y), v(x, y)) = (x^2y + y^2x, y - x).$$

It is not at all easy to invert this. But we do not need to. We compute

$$[D\mathbf{u}(x, y)] = \begin{bmatrix} 2xy + y^2 & 2xy + x^2 \\ -1 & 1 \end{bmatrix}.$$

Therefore $(\det[D\mathbf{u}(x, y)])^{-1} = \frac{1}{4xy + x^2 + y^2}$ which is exactly the inverse of our integrand, so that

$$f(x, y)(\det[D\mathbf{u}(x, y)])^{-1} = 1.$$

This is trivial to express in terms of u and v : constants are constants. Hence the integral becomes

$$\int_{\tilde{D}} 1 d^2\mathbf{u} = \left(\int_{-1}^1 dv \right) \left(\int_1^2 du \right) = 2.$$

7.3 Integration in \mathbb{R}^3

7.3.1 Reduction to iterated integrals in lower dimension

What we have studied so far concerning integrals of functions on \mathbb{R}^2 readily extends to integrals of functions on \mathbb{R}^3 .

The basic formula defining the integral of a continuous real valued function over a region $D \subset \mathbb{R}^3$ is

$$\int_D f(x, y, z) d^3x = \lim_{\text{box diameter} \rightarrow 0} \left(\sum_{\text{little boxes}} (\text{value of } f \text{ in the box}) \times (\text{volume of box}) \right) \quad (7.43)$$

where the sum is over the little “boxes” in a “disintegration” of D into a disjoint union of sets, here called “boxes” whose volume we know how to measure, or at least accurately estimate. As before, we require that the maximum diameter of the boxes tends to zero in the limit. This is the obvious generalization of (7.10).

Then, as in two dimensions, if f is any continuous real valued function on \mathbb{R}^3 , and D is any regular bounded subset of \mathbb{R}^3 , the limit in (7.43) exists no matter what kind of decomposition we use and no matter where in the box we evaluate f (as long as we can compute the volume of the boxes, and as long as the maximum diameter of the boxes goes to zero).

The simplest example of such a disintegration is to use the coordinate planes to “slice” the region into cubes of side length $h > 0$. This gives us a particularly easy formula for the volume of the boxes - each one has volume h^3 .

For integers j, k, ℓ and $h > 0$, define

$$x_j := jh \quad y_k := kh \quad \text{and} \quad z_\ell = \ell h .$$

Then the planes

$$x = x_j , \quad y = y_k , \quad z := z_\ell , \quad -\infty \leq j, k, \ell < \infty$$

slice up \mathbb{R}^3 into a regular grid of cubes of side length h . Of course, they also slice up any subset $D \subset \mathbb{R}^3$ into sets we shall call boxes, each of which is contained in one of the cubes. If the set D has a “nice” boundary and h is small, so that “most” of D is at least a distance \sqrt{nh} from the boundary, most of the boxes will be entire cubes. (Note that \sqrt{nh} is the diameter of a cube of side length h .)

When we do the sum over all of the little boxes in (7.43), we can do the sum in any order we like. Here is one good order of summation: For each ℓ , let Dz_ℓ be the intersection of D with the slab

$$\{ (x, y, z) : z_\ell \leq z < z_{\ell+1} \} .$$

We then have

$$\begin{aligned} \sum_{\text{little boxes in } D} (\text{value of } f \text{ in the box}) \times (\text{volume of box}) &= \\ \sum_{\ell} \left(\sum_{\text{little boxes in } Dz_\ell} (\text{value of } f \text{ in the box}) \times (\text{volume of box}) \right) . \end{aligned} \quad (7.44)$$

Let the “tile” associated to each box be the bottom of the box; i.e., the intersection of the box with the plane $z = z_\ell$. Thus there is one tile to each box. Let us evaluate f somewhere in the tile – since we are free to evaluate it anywhere in the box. Then since the area of the tile is h^2 and the volume of the box is h^3 ,

$$\left(\sum_{\text{little boxes in } Dz_\ell} (\text{value of } f \text{ in the box}) \times (\text{volume of box}) \right) = h \left(\sum_{\text{little tiles in } Dz_\ell} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) \right) \quad (7.45)$$

It is now easy to recognize

$$\lim_{h \rightarrow 0} \left(\sum_{\text{little tiles in } Dz_\ell} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) \right)$$

as an area integral in \mathbb{R}^2 . Specifically, for each $z \in \mathbb{R}$, define

$$\hat{D}z = \{ (x, y) : (x, y, z) \in D \} .$$

Note that for each z_0 , $\hat{D}z_0$ is the “slice” of D by the plane $z = z_0$, projected onto the x, y plane. Then we have

$$\begin{aligned} \lim_{h \rightarrow 0} \left(\sum_{\text{little tiles in } Dz_\ell} (\text{value of } f \text{ in the tile}) \times (\text{area of tile}) \right) &= \\ &\int_{\hat{D}z_\ell} f(x, y, z_\ell) dx dy . \end{aligned} \quad (7.46)$$

Now combining (7.44), (7.45) and (7.46), we obtain (formally interchanging limits and sums without justification, but in a way that can be justified), we obtain

$$\begin{aligned} \lim_{h \rightarrow 0} \left(\sum_{\text{little boxes in } D} (\text{value of } f \text{ in the box}) \times (\text{volume of box}) \right) &= \\ &\lim_{h \rightarrow 0} \left(\sum_\ell \left(\int_{\hat{D}z_\ell} f(x, y, z_\ell) dx dy \right) h \right) . \end{aligned}$$

Now let a be the greatest lower bound on the set of all z such that \hat{D}_z is non-empty, and let b be the least upper bound on the set of all z such that \hat{D}_z is non-empty. Then the right hand side is the limit of Riemann sums for the integral

$$\int_a^b \left(\int_{\hat{D}_z} f(x, y, z) dx dy \right) dz .$$

Thus we finally have:

$$\int_D f(x, y, z) d^3 \mathbf{x} = \int_a^b \left(\int_{\hat{D}_z} f(x, y, z) dx dy \right) dz \quad (7.47)$$

with a, b and \hat{D}_z defined as above.

- This formula reduces the integration of functions over the three dimension set D to an integration over its two dimensional slices, followed by a one dimensional integral to put everything together.

Since we learned in the previous section how to do the integrals over the two dimensional slices, and since we know how to do one dimensional integrals, we know how to compute integrals of functions over three dimension sets D . The derivation of this formula can be completely justified under the assumptions that f is continuous, and D is bounded has a “reasonably smooth” boundary.

We shall return later to a more careful justification of this formula. In the rest of this section, we focus on explaining how it is used.

First, it is not necessary to integrate in z last: We can do the sums in any order we like, so we can integrate in any order we like. And in general, we are likely to like some orders better than others: How easy or complicated the limits of integration are will depend strongly on the chosen order of integration.

7.3.2 The change of variables formula for integrals in \mathbb{R}^3

The same reasoning that led to the change of variables formula in \mathbb{R}^2 yields a change of variables formula in \mathbb{R}^3 : Let \mathbf{x} be a differentiable and invertible transformation from $\hat{D} \subset \mathbb{R}^3$ onto $D \subset \mathbb{R}^3$. Then, for any continuous function f on D ,

$$\int_D f(\mathbf{x}) d^3\mathbf{x} = \int_{\hat{D}} f(\mathbf{x}(\mathbf{u})) |\det(D\mathbf{x}(\mathbf{u}))| d^3\mathbf{u}. \quad (7.48)$$

The first examples to consider are the mappings \mathbf{x} associated to standard coordinate systems on \mathbb{R}^3 . For example, consider

$$\begin{aligned} \mathbf{x}(r, \theta, \varphi) &= (x(r, \theta, \varphi), y(r, \theta, \varphi), z(r, \theta, \varphi)) \\ &= (r \sin \varphi \cos \theta, r \sin \varphi \sin \theta, r \cos \varphi), \end{aligned} \quad (7.49)$$

where

$$\begin{aligned} 0 &\leq r \\ 0 &\leq \theta < 2\pi \\ 0 &\leq \varphi \leq \pi. \end{aligned}$$

This is the *spherical coordinate system* for \mathbb{R}^3 .

An easy computation shows that $\|\mathbf{x}(r, \theta, \varphi)\| = r^2$, so as r is held fixed, and θ and φ vary, $\mathbf{x}(r, \theta, \varphi)$ ranges over the sphere of radius r in \mathbb{R}^3 .

We now compute

$$[D\mathbf{x}(r, \theta, \varphi)] = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \varphi} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} \sin \varphi \cos \theta & -r \sin \varphi \sin \theta & r \cos \varphi \cos \theta \\ \sin \varphi \sin \theta & r \sin \varphi \cos \theta & r \cos \varphi \sin \theta \\ \cos \varphi & 0 & -r \sin \varphi \end{bmatrix}.$$

Taking the determinant of the right hand side, we find

$$\begin{aligned}\det([D\mathbf{x}(r, \theta, \varphi)]) &= -r^2 \sin \varphi [(\sin^2 \varphi + \cos^2 \varphi)(\sin^2 \theta + \cos^2 \theta)] \\ &= -r^2 \sin \varphi .\end{aligned}\tag{7.50}$$

Since $\sin \varphi$ is non-negative for $0 \leq \varphi \leq \pi$, $|\det([D\mathbf{x}(r, \theta, \varphi)])|d\varphi d\theta = r^2 \sin \varphi$, and hence the volume element dV for spherical coordinates is given by

$$dV = r^2 \sin \varphi dr d\theta d\varphi .\tag{7.51}$$

Example 114 (The average height in the upper hemisphere). Let \mathcal{V} be the upper hemisphere of radius R . That is, \mathcal{V} consists of the points (x, y, z) satisfying

$$\begin{aligned}x^2 + y^2 + z^2 &\leq R^2 \\ z &\geq 0\end{aligned}$$

The average height in \mathcal{V} is

$$\frac{\int_{\mathcal{V}} z dV}{\int_{\mathcal{V}} 1 dV} .$$

The denominator is the total volume of \mathcal{V} . Thus, from (7.43), this ratio is

$$\lim_{\text{box diameter} \rightarrow 0} \left(\sum_{\text{little boxes}} (\text{value of } z \text{ in the box}) \times \left(\frac{\text{volume of box}}{\text{volume of } \mathcal{V}} \right) \right) .$$

That is, the ratio of the integrals is the weighted average of the heights of all the little boxes in a disintegration of \mathcal{V} into infinitesimal boxes, where the weighting is by the fractional volume of the box.

We already know that volume of the sphere of radius R is $\frac{4}{3}\pi R^3$, so we only need to compute the integral in the numerator (though it would be a good exercise to go back and compute the integral in the denominator afterwards, using the same method).

Translating the description of \mathcal{V} into spherical coordinates using (7.49) we find $r^2 \leq R^2$. since both r and R are non-negative, this means $r \leq R$. Likewise, $z \geq 0$ becomes $r \cos \varphi \geq 0$, which, since $r > 0$ except at the origin, means $\varphi \leq \pi/2$. Then going back to (7.51) we have, for this problem,

$$\begin{aligned}0 &\leq r \leq R \\ 0 &\leq \theta < 2\pi \\ 0 &\leq \varphi \leq \pi/2 .\end{aligned}$$

As r , θ and φ range over the sets described by (7.52), $\mathbf{x}(r, \theta, \varphi)$ ranges over \mathcal{V} , the upper hemisphere of radius R .

To compute $\int_{\mathcal{V}} z dV$ we first translate z into spherical terms. By (7.49), $z = r \cos \varphi$, and then from (7.51) and (7.52)

$$\begin{aligned}\int_{\mathcal{V}} z dV &= \int_0^{2\pi} \left(\int_0^{\pi/2} \left(\int_0^R r^3 \cos \varphi \sin \varphi dr \right) d\varphi \right) d\theta \\ &= \left(\int_0^{2\pi} 1 d\theta \right) \left(\int_0^{\pi/2} \cos \varphi \sin \varphi d\varphi \right) \left(\int_0^R r^3 dr \right) . \\ &= (2\pi)(1/2)(R^4/4) = \frac{\pi}{4} R^4 .\end{aligned}\tag{7.52}$$

Thus, the average height, weighted by volume, in the upper hemisphere of radius R is

$$\left(\frac{2}{3}\pi R^3\right)^{-1} \frac{\pi}{4}R^4 = \frac{3}{8}R.$$

This makes sense: There is more volume in the lower part of V than the upper part, so lower values of z will get a higher weight than higher values of z . Therefore, we certainly expect an answer that is less than $R/2$. If anything, the surprise is that the actual value is as close as it is to $R/2$.

The next example to consider is the change of variables associated to cylindrical coordinates:
Let consider

$$\begin{aligned}\mathbf{x}(r, \theta, z) &= (x(r, \theta, z), y(r, \theta, z), z(r, \theta, z)) \\ &= (r \cos \theta, r \sin \theta, z),\end{aligned}\tag{7.53}$$

where

$$\begin{aligned}0 &\leq r \\ 0 &\leq \theta < 2\pi \\ -\infty &< z < \infty.\end{aligned}$$

An easy computation shows that $\|\mathbf{x}(r, \theta, \varphi)\| = r^2 + z^2$, so as r is held fixed, and θ and z vary, $\mathbf{x}(r, \theta, z)$ ranges over the cylinder of radius r in centered on the z -axis in \mathbb{R}^3 .

We now compute

$$[D\mathbf{x}(r, \theta, z)] = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial z} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial z} \end{bmatrix} = \begin{bmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Taking the determinant of the right hand side, we find

$$\det([D\mathbf{x}(r, \theta, z)]) = r(\sin^2 \theta + \cos^2 \theta) = r.$$

Hence the volume element dV for cylindrical coordinates is given by

$$dV = r dr d\theta dz.\tag{7.54}$$

Example 115 (Total mass of an object). Consider an object that has a mass density of $\varrho(x, y, z) = x^2 + y^2$ grams per cubic centimeter at (x, y, z) , and that occupies the region \mathcal{V} given by

$$\begin{aligned}(y-1)^2 + x^2 &\leq 1 \\ z^2 &\leq x^2 + y^2\end{aligned}$$

Let us compute the total mass of the object. This is given by the integral $\int_{\mathcal{V}} \varrho(\mathbf{x}) dV$.

We will compute this in cylindrical coordinates. Let us first translate the description of \mathcal{V} into cylindrical terms. We find $(y-1)^2 + x^2 - 1 = x^2 + y^2 - 2y = r^2 - 2r \sin \theta$, so from $(y-1)^2 + x^2 - 1 \leq 0$, we obtain $r^2 - 2r \sin \theta \leq 0$. Away from the z -axis (which has zero volume), $r > 0$, and so we conclude that $0 \leq r \leq 2 \sin \theta$. Since $r \geq 0$, $\sin \theta \geq 0$, and so $0 \leq \theta \leq \pi$.

Next, $z^2 \leq x^2 + y^2$ becomes $|z| \leq r$, and so, altogether, we have

$$\begin{aligned} 0 &\leq r \leq 2 \sin \theta \\ 0 &\leq \theta \leq \pi \\ -r &\leq z \leq r . \end{aligned}$$

Next, translating $\varrho(\mathbf{x})$ into cylindrical terms, we find $\varrho(\mathbf{x}) = r^2$. Thus,

$$\begin{aligned} \int_{\mathcal{V}} \varrho(\mathbf{x}) dV &= \int_0^{2\pi} \left(\int_0^{2 \sin \theta} \left(\int_{-r}^r r^2 dz \right) dr \right) d\theta \\ &= \int_0^{2\pi} \left(\int_0^{2 \sin \theta} 2r^3 dr \right) d\theta \\ &= \int_0^{2\pi} 8 \sin^4 \theta d\theta = \frac{3}{2}\pi . \end{aligned}$$

7.4 Integration on parameterized surfaces

7.4.1 Parameterized surfaces

Recall that a parameterized surface in \mathbb{R}^3 is a continuously differentiable function $\mathbf{X}(u, v)$ defined on an subset U of \mathbb{R}^2 with values in \mathbb{R}^3 such that $[D\mathbf{X}(u, v)]$ has linearly independent columns at each $(u, v) \in U$. For example consider

$$U = \{ (u, v) : 0 < u < \pi, 0 < v < 2\pi \} , \quad (7.55)$$

and

$$\mathbf{X}(u, v) = (\sin u \cos v, \sin u \sin v, \cos u) . \quad (7.56)$$

Computing $\|\mathbf{X}(u, v)\|$ we find $\|\mathbf{X}(u, v)\| = \sqrt{\sin^2 u(\cos^2 v + \sin^2 v) + \cos^2 u} = 1$. Thus, for each $(u, v) \in U$, $\mathbf{X}(u, v)$ lies in the unit sphere, and $\mathbf{X}(u, v)$ is a parameterization of the unit sphere in \mathbb{R}^3 , take away the semicircle running between the North and South Poles along $(\sin u, 0, \cos u)$ for $0 \leq u \leq \pi$. We exclude this semicircle to keep U open, and to keep $\mathbf{X}(u, v)$ one-to-one. The price we pay is that this is not a parameterization of the *entire* unit sphere. However, what we have left out accounts for *none* of the surface area of the unit sphere. For the purposes of this section, the fact that we have parameterized a part of the sphere that accounts for all of its surface area, and have done so in a one-to-one and continuously differentiable way, will be what matters.

Other examples of parameterized surfaces are given by the graphs of real valued functions f defined on \mathbb{R}^2 . Given such a function, define

$$\mathbf{X}(u, v) = (u, v, f(u, v)) .$$

A surface in \mathbb{R}^3 can also be given in implicit form. For example, the equation

$$x^2 + y^2 + z^2 = 1 \quad (7.57)$$

has the unit sphere in \mathbb{R}^3 as its solution set. Finding a parameterization of the sphere means to find an explicit description of the solution set of this equation. The parameterization given above is one way to do this, and is not only a parameterization, but a continuously differentiable one with a Jacobian of rank 2 for each choice of the parameters.

Finding such parameterizations for implicitly defined surfaces is a key step in the solution of many problems involving such surfaces. In the next example, we find such a parameterization.

Example 116 (Parameterizing a surface). *Consider the equation*

$$(x^2 + y^2 + z^2)^2 = 2(z^2 - x^2 - y^2) .$$

To parameterize the corresponding surface, we must solve this equation. To do this, we rewrite the equation in terms of spherical coordinates:

$$x = r \sin u \cos v , \quad y = r \sin u \sin v , \quad z = r \cos u ,$$

where $(u, v) \in U$ where U is given by (7.55). In these variables, the equation becomes

$$r^4 = r^2(\cos^2 u - \sin^2 u) = r^2 \cos(2u) .$$

The parameter v has dropped out of the equation, which facilitates its solution.

Since $r = 0$ corresponds to the point $(0, 0, 0)$, which is one solution, let us assume $r \neq 0$, and find all other solutions. Dividing by r^2 and taking a square root, we find:

$$r = \sqrt{\cos(2u)} .$$

For values of u such that $\cos(2u) < 0$, we have no solution. In fact for $r > 0$, we must have $\cos(2u) > 0$. The values of u corresponding to solutions of our equation, other than $(0, 0, 0)$, are $0 < u < \pi/4$ and $3\pi/4 < u < \pi$.

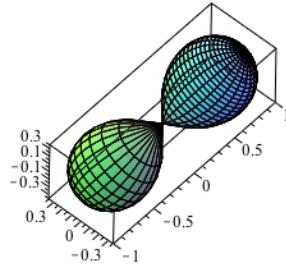
Now, we are ready to do the parameterization: First write out the general point in \mathbb{R}^3 using our chosen system of coordinates.

$$\mathbf{X}(r, u, v) = (r \sin u \cos v, r \sin u \sin v, r \cos u) .$$

Now use the equation $r = \sqrt{\cos(2u)}$ to eliminate the variable r . The other two variables become the parameters. We obtain:

$$\mathbf{X}(u, v) = \sqrt{\cos(2u)} (\sin u \cos v, \sin u \sin v, \cos u) .$$

This gives us our parameterization. Here is a plot of the surface:

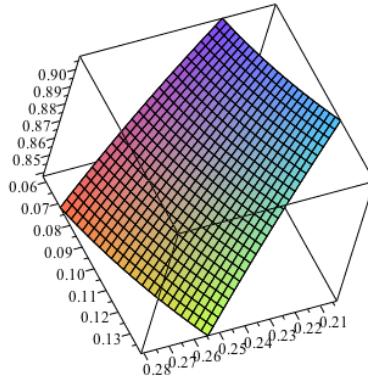


This is an example of a surface of revolution; it is the result of rotating the Bernoulli Lemiscate in the x, z plane about the z -axis.

Whenever we need to parameterize a surface, the procedure will always be the same: We find some system of coordinates in which we can use the equation defining the surface to express one of the variables in terms of the other two. (In the previous example, we used the equation to express r in terms of u and v , though the dependence on v turned out to be trivial.) We then use this expression to eliminate this variable from the expression for a general point in \mathbb{R}^3 in the chosen coordinate system. The two remaining variables become the two parameters.

7.4.2 The surface area of a parameterized surface

Let $\mathbf{X}(u, v)$ be a differentiable parameterized surface defined on an open set $U \subset \mathbb{R}^2$. The mapping $(u, v) \mapsto \mathbf{X}(u, v)$ transfers the coordinate grid in the u, v plane onto the parameterized surface, producing a coordinate grid there. For instance, if $\mathbf{X}(u, v) := \sqrt{\cos(2u)}(\sin u \cos v, \sin u \sin v, \cos u)$, here is a plot of $\mathbf{X}(u, v)$ showing the coordinate grid for $(u, v) \in [1/4, 13] \times [1/4, 1/2]$



The grid you see on this piece of the surface consists of lines of constant u and constant v . They carve the surface up into little “tiles”. Each of these has vertices of the form

$$\mathbf{X}(u_0, v_0) \quad \mathbf{X}(u_0 + h, v_0) \quad \mathbf{X}(u_0, v_0 + h) \quad \text{and} \quad \mathbf{X}(u_0 + h, v_0 + h),$$

for some small value of $h > 0$.

To the extent that the tangent plane approximation is valid, the tile is the parallelogram in \mathbb{R}^3 with a corner at $\mathbf{X}(\mathbf{u}_0)$ and sides $h\mathbf{X}_u(\mathbf{u}_0)$ and $h\mathbf{X}_v(\mathbf{u}_0)$ issuing from this corner. The area of the parallelogram is given by the magnitude of the cross product of the vectors giving the sides:

$$\text{surface area of parallelogram} = h^2 \|\mathbf{X}_u(\mathbf{u}_0) \times \mathbf{X}_v(\mathbf{u}_0)\|$$

Now if we add up the surface areas of the parallelograms, we get a good approximation of the area of this piece of the parameterized surface. Taking the limit $h \rightarrow 0$, the approximation becomes exact, and so the area of this piece of the parameterized surface is given by

$$\lim_{h \rightarrow 0} \left(\sum_{\text{littletiles}} h^2 \times (\text{the value of } \|\mathbf{X}_u \times \mathbf{X}_v\| \text{ in the tile}) \right).$$

This of course is nothing other than the limit of Riemann sums for the integral

$$\int_{1/4}^{1/2} \left(\int_{1/4}^{1/3} \|\mathbf{X}_u \times \mathbf{X}_v\| du \right) dv.$$

Integrating over any open U set on which $\mathbf{X}(\mathbf{u})$ is defined and differentiable, we obtain in the same way that the surface area of the part of the surface with parameters in U is given by

$$\text{surface area} = \int_U \|\mathbf{X}_u \times \mathbf{X}_v\| d^2\mathbf{u}. \quad (7.58)$$

Just as we did with arc length, we define the *surface area element* dS by

$$dS = \|\mathbf{X}_u \times \mathbf{X}_v\| d^2\mathbf{u}.$$

We may then express the integral formula for the surface area of a parameterized surface \mathcal{S} as

$$\text{surface area} = \int_{\mathcal{S}} 1 dS. \quad (7.59)$$

Notice the difference between (7.58) and (7.59): In (7.58), we consider the domain of integration to be U , the set of parameter values. Indeed, when it comes to actual computation, this is what we will be working with. In (7.59), we consider the domain of integration to be the surface \mathcal{S} itself, and we do not refer to any particular set of coordinates. This puts the emphasis on what the integral actually represents: The sum of all of the contributions of the various area elements to the total area. We will use either sort of notation, depending on what is best suited to the matter at hand.

Example 117 (Computing surface area). *Let \mathcal{S} be the part of the paraboloid $z = 1 - x^2 - y^2$ that lies above the plane $x + z = 1$. Let us compute the surface area of \mathcal{S} .*

The key to doing this is coming up with a good parameterization. To find the intersection of the plane and paraboloid, we equate their z values and find

$$1 - x^2 - y^2 = 1 - x$$

which is the same as $x^2 + y^2 = x$ or $(x - 1/2)^2 + y^2 = 1/4$. This is the circle bounding the disk in the x, y plane centered on $(1/2, 0)$ with radius $1/2$. This is what we would see in a top view diagram. Our surface \mathcal{S} is the part of the paraboloid that lies above this disk. Let us use cylindrical coordinates:

$$(x, y, z) = (r \cos \theta, r \sin \theta, z).$$

The equation for the paraboloid is $x = 1 - r^2$, and so we have our parameterization

$$\mathbf{X}(r, \theta) = (r \cos \theta, r \sin \theta, 1 - r^2) .$$

The equation $x^2 + y^2 = x$ translates to $r^2 = r \cos \theta$, so the limits on our parameters are

$$0 \leq r \leq \cos \theta \quad \text{and} \quad -\pi/2 \leq \theta \leq \pi/2 .$$

We next compute $\mathbf{X}_r \times \mathbf{X}_\theta = (2r^2 \cos \theta, 2r^2 \sin \theta, r)$, and hence the surface area element dS is given by

$$dS = r \sqrt{4r^2 + 1} dr d\theta .$$

Hence the surface area is given by $\int_S 1 dS = \int_{-\pi/2}^{\pi/2} \left(\int_0^{\cos \theta} r \sqrt{4r^2 + 1} dr \right) d\theta$. The inner integral is easily done by substitution:

$$\int_0^{\cos \theta} r \sqrt{4r^2 + 1} dr = \frac{1}{8} \frac{2}{3} u^{3/2} \Big|_1^{4 \cos^2 \theta + 1} = \frac{1}{12} ((4 \cos^2 \theta + 1)^{3/2} - 1) .$$

We finally have

$$\int_S f(x, y, z) dS = \frac{1}{12} \int_{-\pi/2}^{\pi/2} ((4 \cos^2 \theta + 1)^{3/2} - 1) d\theta .$$

This may be evaluated in terms of special functions (incomplete elliptic integrals of the first kind), and perhaps more meaningfully, one finds the numerical value

$$1.21458608\dots .$$

Example 118 (Surface area again). Let \mathcal{V} be the region in \mathbb{R}^3 that lies inside the sphere $x^2 + y^2 + z^2 = 4$, and above the graph of $z = 1/\sqrt{x^2 + y^2}$. Compute the total surface area of its boundary \mathcal{S} . (There are two pieces to the boundary.)

First of all, we must parameterize each of the two pieces of the boundary \mathcal{S} . The boundary equations are simple in cylindrical coordinates. We write

$$\mathbf{X}(r, \theta) = (r \cos \theta, r \sin \theta, z) . \tag{7.60}$$

We then use $z = \sqrt{4 - r^2}$ to eliminate z in (7.60), obtaining

$$\mathbf{X}_1(r, \theta) = (r \cos \theta, r \sin \theta, \sqrt{4 - r^2}) . \tag{7.61}$$

as the parameterization of the upper part of the boundary; let us call this \mathcal{S}_1 .

We next use $z = 1/r$ to eliminate z in (7.60), obtaining

$$\mathbf{X}_2(r, \theta) = (r \cos \theta, r \sin \theta, 1/r) . \tag{7.62}$$

as the parameterization of the upper part of the boundary; let us call this \mathcal{S}_2 .

We now determine the range of the parameters r and θ in our parameterization.

From $r^2 + z^2 = 4$ and $z = 1/r$, we deduce

$$(r^2)^2 - 4(r^2) = -1 .$$

This quadratic equation for r^2 has the roots $r^2 = 2 \pm \sqrt{3}$. Hence the two bounding surfaces intersect at

$$r = \sqrt{2 - \sqrt{3}} \quad \text{and} \quad r = \sqrt{2 + \sqrt{3}}.$$

Hence U is the set of vectors (r, θ) with

$$\begin{aligned} \sqrt{2 - \sqrt{3}} &\leq r \leq \sqrt{2 + \sqrt{3}} \\ 0 &\leq \theta \leq 2\pi. \end{aligned}$$

Next, we work out the area element dS_1 for the upper surface \mathcal{S}_1 . We compute

$$\mathbf{X}_r \times \mathbf{X}_\theta(r, \theta) = \left(\frac{r^2 \cos \theta}{\sqrt{4 - r^2}}, \frac{r^2 \sin \theta}{\sqrt{4 - r^2}}, r \right).$$

We then compute

$$\left\| \left(\frac{r^2 \cos \theta}{\sqrt{4 - r^2}}, \frac{r^2 \sin \theta}{\sqrt{4 - r^2}}, r \right) \right\|^2 = \frac{r^4}{4 - r^2} + \frac{r^2(4 - r^2)}{4 - r^2} = \frac{4r^2}{4 - r^2}.$$

Thus,

$$dS_1 = \frac{2r}{\sqrt{4 - r^2}} dr d\theta.$$

We then have

$$\text{surface area of } \mathcal{S}_1 = \int_0^{2\pi} \left(\int_{\sqrt{2-\sqrt{3}}}^{\sqrt{2+\sqrt{3}}} \frac{2r}{\sqrt{4 - r^2}} dr \right) d\theta = \int_0^{2\pi} 2\sqrt{2} d\theta = 4\sqrt{2}\pi.$$

For the other part of the boundary, we work out the area element dS_2 for the upper surface \mathcal{S}_2 .

We compute

$$\mathbf{X}_r \times \mathbf{X}_\theta(r, \theta) = \left(\frac{\cos \theta}{r}, \frac{\sin \theta}{r}, r \right).$$

We then compute

$$\left\| \left(\frac{\cos \theta}{r}, \frac{\sin \theta}{r}, r \right) \right\|^2 = \frac{1}{r^2} + r^2.$$

Thus,

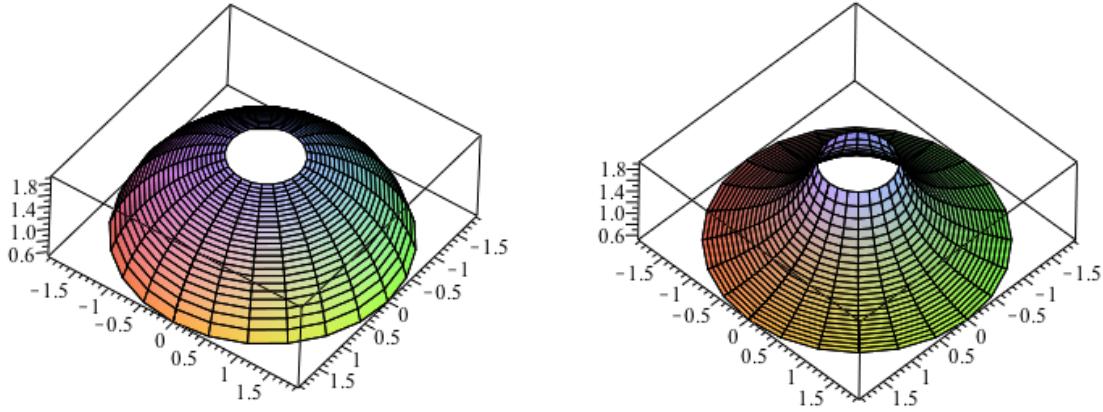
$$dS_2 = \sqrt{r^{-2} + r^2} dr d\theta.$$

We then have

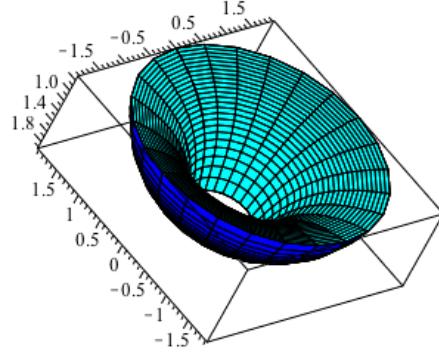
$$\begin{aligned} \text{surface area of } \mathcal{S}_2 &= \int_0^{2\pi} \left(\int_{\sqrt{2-\sqrt{3}}}^{\sqrt{2+\sqrt{3}}} \sqrt{r^{-2} + r^2} dr \right) d\theta \\ &= 2\pi \left[\sqrt{2} + \frac{1}{2} \operatorname{arctanh} \left(\frac{1}{2} \frac{\sqrt{2}}{\sqrt{3}-1} \right) - \frac{1}{2} \operatorname{arctanh} \left(\frac{1}{2} \frac{\sqrt{2}}{\sqrt{3}+1} \right) \right]. \end{aligned}$$

Combining the two computations, we get the total area.

Here are plots of \mathcal{S}_1 , \mathcal{S}_2 :



Here is a plot of $\mathcal{S}_1 \cup \mathcal{S}_2$:



We may also integrate functions defined on parameterized surface \mathcal{S} . For example, let \mathcal{S} be the centered sphere of radius $R > 0$. Suppose it has a mass density of ϱ grams per centimeter squares. If it is rotating about the z -axis with angular velocity ω , the total kinetic energy is

$$\int_{\mathcal{S}} \frac{\varrho}{2} (x^2 + y^2) \omega^2 dS ,$$

and by definition, the *moment of inertia* of the spherical shell is the quantity I such that the total kinetic energy is given by

$$\frac{1}{2} I \omega^2 .$$

Since the total mass of the shell is $M = 4\pi R^2 \varrho$, we have

$$I = \varrho \int_{\mathcal{S}} (x^2 + y^2) dS = M \frac{1}{4\pi R^2} \int_{\mathcal{S}} (x^2 + y^2) dS . \quad (7.63)$$

In the next example, we compute the integral on the right.

Example 119 (The moment of inertia of a spherical shell). *Let \mathcal{S} be the centered sphere of radius $R > 0$. Let $f(x, y, z) = (x^2 + y^2)/2$. Let us compute*

$$\int_{\mathcal{S}} f dS .$$

First, we need to parameterize the surface. We could use spherical coordinates, but we can express the integrand more simply in cylindrical coordinates, and the parameterization is almost as simple in cylindrical coordinates.

The equation for the sphere in cylindrical coordinates is $r^2 + z^2 = R^2$. Therefore, we eliminate $z = \pm\sqrt{R^2 - r^2}$, and obtain

$$\mathbf{x}(r, \theta) = (r \cos \theta, r \sin \theta, \pm\sqrt{R^2 - r^2}).$$

There are two pieces of the surface, say \mathcal{S}_+ and \mathcal{S}_- corresponding to the two choices for the sign. By symmetry, we can integrate over \mathcal{S}_+ , and double our answer.

The domain U of integration is

$$0 < r < R \quad \text{and} \quad 0 < \theta < 2\pi.$$

We then work out $\mathbf{X}_r \times \mathbf{X}_\theta(r, \theta) = \left(\frac{r^2 \cos \theta}{\sqrt{R^2 - r^2}}, \frac{r^2 \sin \theta}{\sqrt{R^2 - r^2}}, r \right)$ and so

$$dS = \frac{Rr}{\sqrt{R^2 - r^2}} dr d\theta.$$

Since $f(r \cos \theta, r \sin \theta, z) = r^2$, we have

$$\int_{\mathcal{S}_+} f dS = \int_0^{2\pi} \left(\int_0^1 r^2 \frac{Rr}{\sqrt{R^2 - r^2}} dr \right) d\theta = 2\pi \left(\frac{2R^4}{3} \right) = \frac{4\pi R^4}{3}.$$

Remembering to double our answer, we have $\int_{\mathcal{S}} f dS = \frac{8\pi R^4}{3}$. The total mass M of the spherical shell is its area times the density ϱ .

Thus, going back to the remarks made right before the example, and (7.63) in particular, the computation shows that the moment of inertial I of a spherical shell of radius R and mass M is given by

$$I = \frac{2}{3}MR^2.$$

7.5 Exercises

7.1 Let $f(x, y) = x^3y$, and let D be the region that lies to the right of the parabola $x = y^2$, and below the line $2y = -x$.

(a) Write down $\int_D f(x, y) dA$ in terms of integrated integrals, integrating in x first, then y .

(b) Write down $\int_D f(x, y) dA$ as an integrated integral integrating in y first, then x .

(c) Evaluate one of the integrals.

7.2 Let $f(x, y) = x^2y^2$, and let D be the region that lies inside both of the circles $(x - 1)^2 + y^2 = 4$ and $(x + 1)^2 + y^2 = 4$.

(a) Write down $\int_D f(x, y) dx dy$ as an integrated integral, integrating in x first, then y .

(b) Write down $\int_D f(x, y) dx dy$ as an integrated integral, integrating in y first, then x .

(c) Evaluate one of the integrals.

7.3 Let $f(x, y) = x^2y^2$, and let D be the region that lies below the parabola $y = 4 - (x - 2)^2$ and above the x axis.

(a) Write down $\int_D f(x, y)dxdy$ as an integrated integral, integrating in x first, then y .

(b) Write down $\int_D f(x, y)dxdy$ as an integrated integral, integrating in y first, then x .

(c) Evaluate one of the integrals.

7.4 Let $f(x, y) = xy$, and let D be the region bounded by the lines $y = x$, $y = 3x$, and $y = 5x - 6$.

(a) Write down $\int_D f(x, y)dxdy$ in terms of integrated integrals, integrating in x first, then y .

(b) Write down $\int_D f(x, y)dxdy$ in terms of integrated integrals, integrating in y first, then x .

(c) Evaluate one of the integrals.

7.5 Let $f(x, y) = x^2 + y^2$, and let D be the region bounded by the lines $y = -x$, $y = x$ and $y = 5 - 2x$.

(a) Write down $\int_D f(x, y)dxdy$ in terms of integrated integrals, integrating in x first, then y .

(b) Write down $\int_D f(x, y)dxdy$ in terms of integrated integrals, integrating in y first, then x .

(c) Evaluate one of the integrals.

7.6 Let $f(x, y) = y$, and let D be the region bounded by $x + y = 2$, $x + y = 4$, $xy = 1$ and $xy = 2$.

Compute $\int_D f(x, y)dxdy$.

7.7 Let D be the region bounded by $x^4 + y^4 = 1$. Compute its area. (Use symmetry to conclude that the area is 4 times the area of the piece in the upper right quadrant, and set up an integral to compute that). Leave your answer in the form of an explicit integral over one variable. If you do this in the way that is intended, you will be left with what is known as an *elliptic integral*. These are well studied and Maple, for example, is programmed to deal with them.

7.8 Let $f(x, y) = xy$, and let D be the region in the positive quadrant of \mathbb{R}^2 bounded by $xy = 1$, $xy = 2$, $y/x = 1$ and $y/x = 2$. Compute $\int_D f(x, y)dA$.

7.9 Let $f(x, y) = y$, and let D be the region inside both of the circles

$$(x - 1)^2 + (y - 1)^2 = 2 \quad \text{and} \quad (x + 1)^2 + (y - 1)^2 = 2 .$$

Compute $\int_D f(x, y)dA$.

7.10 Consider the region enclosed by the curve

$$x^2 + y^2 = (x^2 + y^2 - x)^2 .$$

Show that in polar coordinates, this curve is given by

$$r = 1 + \cos \theta .$$

Sketch the curve, and compute the area it encloses.

7.11 Consider the closed curve given in polar terms by $r = \sin^3 \theta$. Sketch this curve, and compute the area enclosed.

7.12 Consider the closed curve given in polar terms by $r = 1 + \sin(2\theta)$. Sketch this curve, and compute the area enclosed.

7.13 Let D be the region in \mathbb{R}^2 that is bounded by the lines

$$y - x = 0 \quad y - x = 3 \quad x + 2y = 2 \quad \text{and} \quad x + 2y = 4 .$$

Compute

$$\int_D \frac{x - y}{x^2 + 4xy + 4y^2} dA .$$

7.14: Let $D \subset \mathbb{R}^2$ be the region bounded by $xy = 1$, $xy = 2$, $x^2y = 3$ and $x^2y = 4$. Compute $\int_D xy dA$.

7.15: Let D be the set in \mathbb{R}^2 that is given by

$$1 \leq \frac{y}{x^2} \leq 2 \quad \text{and} \quad 1 \leq \frac{x}{y^2} \leq 2 .$$

Let $f(x, y) = \frac{1}{x^2y^2}$. Compute $\int_D f(x, y) dA$.

7.16: Let D be the set in \mathbb{R}^2 that is given by

$$0 \leq x^2 - y^2 \leq 4 \quad \text{and} \quad 1 \leq xy \leq 2 .$$

Let $f(x, y) = x^2 + y^2$. Compute $\int_D f(x, y) dA$.

7.17 Let \mathcal{V} be the region in \mathbb{R}^3 that is bounded above by the sphere $x^2 + y^2 + z^2 = 4$, and below by the cone $4z = 4 - \sqrt{x^2 + y^2}$. Let $f(x, y, z) = 1/(x^2 + y^2 + z^2)^2$. Compute $\int_{\mathcal{V}} f(x, y, z) dV$.

7.18: Let \mathcal{V} be the region in \mathbb{R}^3 that is the intersection of the three cylinders of unit radius along the three coordinate axes. That is, D is the set of points (x, y, z) satisfying

$$y^2 + z^2 \leq 1 \quad x^2 + z^2 \leq 1 \quad \text{and} \quad x^2 + y^2 \leq 1 .$$

Compute the volume of \mathcal{V} .

7.19: Let \mathcal{V} be the region in \mathbb{R}^3 that is above the sphere $x^2 + y^2 + z^2 = 6$ and below the paraboloid $z = 4 - x^2 - y^2$. Compute the volume of this region.

7.20: Let \mathcal{V} be the region in \mathbb{R}^3 that is bounded by the surfaces

$$\begin{aligned} \sqrt{x^2 + y^2} &= z^2 \\ \sqrt{x^2 + y^2} &= 8 - z^2 \end{aligned}$$

Compute the volume of \mathcal{V} and the total surface area of its boundary. (There are two pieces to the boundary.)

(a) Draw a plot of the intersection of the bounding surfaces with the x, z plane.

(b) Compute the volume of \mathcal{V} .

(c) Compute the total surface area of the boundary of \mathcal{V} .

7.21: Let \mathcal{V} be the region in \mathbb{R}^3 consisting of points (x, y, z) satisfying

$$0 \leq z \leq y^2 \sin x$$

for $0 \leq x \leq \pi$, and $0 \leq y \leq 1$. Find the average height in \mathcal{V} .

7.22: Let \mathcal{V} be the region in \mathbb{R}^3 bounded by $y + z = 2$, $2x = y$, $x = 0$ and $z = 0$. Let $f(x, y, z) = xe^z$. Compute $\int_{\mathcal{V}} f(\mathbf{x}) dV$.

7.23: Let \mathcal{V} be the region in the positive octant of \mathbb{R}^3 bounded by the elliptic cylinder $9x^2 + 4y^2 = 36$ and the sphere $x^2 + y^2 + z^2 = 16$. Let $f(x, y, z) = z$. Compute $\int_{\mathcal{V}} f(\mathbf{x}) dV$.

7.24: Let \mathcal{V} be the region in \mathbb{R}^3 contained in the cylinder $x^2 + y^2 = 9$, and lying below the plane $y = z$, and above the plane $z = 0$. Compute the volume of \mathcal{V} .

7.25: Let \mathcal{V} be the region in \mathbb{R}^3 lying between the paraboloids $z = x^2 + y^2$ and $z = 8 - x^2 - y^2$. Compute the volume of \mathcal{V} .

7.26: Let \mathcal{S} be the triangle in \mathbb{R}^3 with vertices $(0, 4, 1)$, $(1, 0, 2)$ and $(0, 0, 3)$.

(a) Find a parameterization of the surface \mathcal{S} . (Hint: Find an equation for the plane containing \mathcal{S} .)

(b) Let $f(x, y, z) = xyz$. Compute $\int_{\mathcal{S}} f dS$.

7.27: Let \mathcal{S} be the surface in \mathbb{R}^3 given by $z = xy$ with $x^2 + y^2 \leq 1$.

(a) Find a parameterization of the surface \mathcal{S} .

(b) Let $f(x, y, z) = z^2$. Compute $\int_{\mathcal{S}} f dS$.

7.28: Let \mathcal{S} be the torus in \mathbb{R}^3 obtained by rotating the circle of radius a on the x, z plane centered at $(b, 0, 0)$, where $b > a$, about the z -axis.

(a) Find a parameterization of the surface \mathcal{S} .

(b) Compute the surface area of the torus.

7.29: Let $\mathbf{x}(r, \theta) := (r \cos \theta, r \sin \theta, g(r))$ for $0 \leq \theta \leq 2\pi$ and, for given positive numbers a and b , $a \leq r \leq b$. Show that the area of the corresponding parameterized surface \mathcal{S} is

$$2\pi \int_a^b \sqrt{1 + (g'(r))^2} r dr .$$

Chapter 8

DETERMINANTS

8.1 Permuations

8.1.1 The permutation group

The concept of a *transformation group* is fundamentally important to modern mathematics, and to geometry in particular. We now introduce a basic example: *the Permutation Group*, which plays an central role in the computation of area, volume and their higher dimensional generalizations.

Definition 82 (Permutation). *A permutation of $\{1, 2, \dots, n\}$ is a function σ from this set onto itself.*

Recall that “onto” means that for every j in $\{1, 2, \dots, n\}$, there is an i with $\sigma(i) = j$. Permutations are also automatically one-to-one and invertible. To see this note that if $\sigma(i) = \sigma(j)$ for some $i \neq j$, then σ can take on at most $n - 1$ values, since two inputs have been spent covering one value. Hence, were σ not one-to-one, it would not be onto either. Hence σ is invertible. The inverse σ^{-1} is itself invertible (its inverse is σ), and so it too is a permuation. (It is just the original function “in reverse”).

We can specify a permutation σ of $\{1, 2, \dots, n\}$ by listing the assignments it makes:

$$\begin{array}{ccccc} 1 & 2 & 3 & \cdots & n \\ \downarrow & \downarrow & \downarrow & \cdots & \downarrow \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots & \sigma(n) \end{array}$$

$$\begin{array}{ccc} 1 & 2 & 3 \end{array}$$

For example, if $n = 3$, and $\sigma(1) = 2$, $\sigma(2) = 3$ and $\sigma(3) = 1$, $\sigma = \begin{array}{ccc} & \downarrow & \downarrow & \downarrow \\ & 2 & 3 & 1 \end{array}$. The arrows do not

really tell us much; we can remember that the top row is inputs, and the bottom row is outputs, and shorten the notation to

$$\sigma = \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \end{array}.$$

The generalization of this way of writing permutations to higher values of n is plain, and we use it freely.

There are exactly $n!$ permutations of $\{1, 2, \dots, n\}$: Consider any permutation σ of $\{1, 2, \dots, n\}$. There are n choices for the value of $\sigma(1)$. Make this choice, and then, $\sigma(1)$ being taken, there are $n - 1$ choices remaining for value of $\sigma(2)$. Next, there are $n - 2$ choices for $\sigma(3)$, the value to be assigned to 3. Continuing in this way, there are $n(n - 1)(n - 2) \cdots 1 = n!$ choices to make when specifying a permutations σ .

Example 120 (Permutations of $\{1, 2, 3\}$). *There are six permutations of $\{1, 2, 3\}$:*

$$\begin{array}{rcl} \sigma_a & = & \begin{matrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{matrix} & \sigma_b = & \begin{matrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{matrix} & \sigma_c = & \begin{matrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{matrix} \\ \\ \sigma_d & = & \begin{matrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{matrix} & \sigma_e = & \begin{matrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{matrix} & \sigma_f = & \begin{matrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{matrix} \end{array} \quad (8.1)$$

Since permutations of $\{1, 2, \dots, n\}$ are functions from this set into itself, we can compose them: If σ_1 and σ_2 are two permutations of $\{1, 2, \dots, n\}$, then $\sigma_2 \circ \sigma_1$ is defined by

$$\sigma_2 \circ \sigma_1(i) = \sigma_2(\sigma_1(i)) \quad , \quad \text{for each } i = 1, \dots, n . \quad (8.2)$$

Example 121 (Composing permutations). *Let us compute $\sigma_d \circ \sigma_b$ where σ_d and σ_b are the permutations given in (8.1). From (8.1) we see that*

$$\begin{aligned} \sigma_d \circ \sigma_b(1) &= \sigma_d(\sigma_b(1)) = \sigma_d(2) = 3 \\ \sigma_d \circ \sigma_b(2) &= \sigma_d(\sigma_b(2)) = \sigma_d(1) = 2 \\ \sigma_d \circ \sigma_b(3) &= \sigma_d(\sigma_b(3)) = \sigma_d(3) = 1 \end{aligned}$$

$$\text{Thus, } \sigma_d \circ \sigma_b = \begin{matrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{matrix} = \sigma_f .$$

The composition product of two permutations is always another permutation, since the composition of invertible functions is invertible (and hence onto).

The permutation σ_a at the upper left of the list in (8.1) is called the *identity* permutation since it sends each element of $\{1, 2, 3\}$ to itself. This has an obvious generalization to other values of n .

Definition 83 (Permutation group). *Let \mathcal{S}_n denote the set of all $n!$ permutations of $\{1, \dots, n\}$, equipped with the composition product $\sigma_1 \circ \sigma_2$. This is the permutation group on $\{1, \dots, n\}$.*

The term “group” has a precise technical meaning in mathematics; It is a generalization of the more concrete notion of a “transformation group” which is what the permutations group is: A *transformation group on a set X* is a set \mathcal{G} of invertible functions from X to X such that whenever

$g \in \mathcal{G}$, then $g^{-1} \in \mathcal{G}$, and such that whenever $g_1, g_2 \in \mathcal{G}$, then $g_1 \circ g_2 \in \mathcal{G}$. Note that, as a consequence of the definition, \mathcal{G} contains the identity transformation $i(x) = x$ for all $x \in X$. Since \mathcal{S}_n contains all invertible transformations from $\{1, \dots, n\}$ into itself, it is the largest transformation group on $\{1, \dots, n\}$.

8.1.2 The character of a permutation

Some permutations “mix things up” more than others – for example, the identity permutations does not mix things up at all. There is a useful way to measure the *degree of mixing* of a permutation in terms of how many pair it “puts out of order”.

In this subsection, we define a function χ on \mathcal{S}_n with values in $\{-1, 1\}$, called the *character*, that is essential to the theory of determinants. The definition of χ depends on another function which measures the “degree of mixing” of a permutation σ , or in other words, “how far σ is from the identity permutation”.

Let P be the set $P := \{(i, j) : 1 \leq i, j \leq n\}$ of all distinct ordered pairs chosen from $\{1, \dots, n\}$, which is a set of $n(n - 1)$ elements. Define the disjoint sets

$$P_{\text{up}} = \{(i, j) : 1 \leq i < j \leq n\} \quad \text{and} \quad P_{\text{down}} = \{(i, j) : 1 \leq j < i \leq n\}.$$

P_{up} is the set of all “increasing” pairs and P_{down} is the set of all “decreasing” pairs. Note that both of these sets consist of $n(n - 1)/2$ ordered pairs, and $P = P_{\text{up}} \cup P_{\text{down}}$.

For any $\sigma \in \mathcal{S}_n$, the function f_σ from P into itself defined by

$$f_\sigma(i, j) = (\sigma(i), \sigma(j))$$

is invertible. In fact, it is a permutation of the elements of P .

Since f_σ is one-to-one and onto, each pair that f_σ moves out of P_{up} into P_{down} must be replaced by a pair that f_σ moves out of P_{down} into P_{up} so that *the number of pairs that f_σ moves out of P_{up} into P_{down} coincides with the number of pairs it moves out of P_{down} into P_{up}* : It is simply the number of pairs that f_σ “swaps” between P_{down} and P_{up} .

Definition 84 (Degree of mixing). *The degree of mixing of a permutation σ of $\{1, 2, \dots, n\}$ is the number of pairs of integers (i, j) in $\{1, 2, \dots, n\}$ with*

$$i < j \quad \text{and} \quad \sigma(i) > \sigma(j). \tag{8.3}$$

This number is denoted $D(\sigma)$. In terms of the notation introduced in the preceding paragraph, $D(\sigma)$ is the number of pairs that f_σ swaps between P_{down} and P_{up} . The more “reversed” pairs, the more mixing there is.

Example 122 (Computing the degree of mixing). *Let us compute $D(\sigma)$ for each of the six permutations of $\{1, 2, 3\}$. There are exactly three pairs (i, j) with $i < j$, namely $(1, 2)$, $(1, 3)$ and $(2, 3)$. To compute the degree of mixing of σ , we look at*

$$(\sigma(1), \sigma(2)) \quad (\sigma(1), \sigma(3)) \quad (\sigma(2), \sigma(3)),$$

and count the number of these pairs that are “out of order”. You can easily check that

$$D(\sigma_a) = 0 \quad D(\sigma_b) = 1 \quad D(\sigma_c) = 1 \quad D(\sigma_d) = 2 \quad D(\sigma_e) = 2 \quad D(\sigma_f) = 3 .$$

Thus, with this definition of the degree of mixing, the order reversing permutation $\begin{matrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{matrix}$ has the highest degree of mixing among all permutations of $\{1, 2, 3\}$.

Lemma 22 (Degree of mixing and inverses). *For each $\sigma \in \mathcal{S}_n$,*

$$D(\sigma^{-1}) = D(\sigma) .$$

Proof. Let f_σ be the invertible function on pairs induced by σ , as explained above. Then evidently $(f_\sigma)^{-1} = f_{\sigma^{-1}}$. However many pairs f_σ swaps between P_{up} and P_{down} , $f_{\sigma^{-1}}$ swaps the same same number back again in undoing the effects of f_σ . \square

The definition of $D(\sigma)$ is useful because of the way it interacts with the composition product: Consider

Lemma 23 (Degree of mixing and composition). *For any $\sigma_1, \sigma_2 \in \mathcal{S}_n$,*

$$D(\sigma_2 \circ \sigma_1) = D(\sigma_2) + D(\sigma_1) - 2c \tag{8.4}$$

where c is a non-negative integer.

Proof. Suppose that when σ_2 is applied, c pairs that had been put out of order by σ_1 are “reordered” when we apply σ_2 . Then,

it (1) Of the $D(\sigma_1)$ pairs reversed by σ_1 , exactly $D(\sigma_1) - c$ are still reversed after applying σ_2 .

(2) Of the $D(\sigma_2)$ pair reversals created by σ_2 , c are “used up” undoing reversals created by σ_1 , and so exactly $D(\sigma_2) - c$ new reversals are created.

Adding things up, $D(\sigma_2 \circ \sigma_1) = (D(\sigma_1) - c) + (D(\sigma_2) - c) = D(\sigma_1) + D(\sigma_2) - 2c$. \square

We now come to our first application of Lemma 23. Note that that whatever c is in (8.4), $2c$ is always an even integer, and so $(-1)^{2c} = 1$, and

$$(-1)^{D(\sigma_2 \circ \sigma_1)} = (-1)^{D(\sigma_1)} (-1)^{D(\sigma_2)} . \tag{8.5}$$

Definition 85 (Character of a Permutation). *The character $\chi(\sigma)$ of a permutation σ is defined by*

$$\chi(\sigma) = (-1)^{D(\sigma)} . \tag{8.6}$$

where $D(\sigma)$ is given by (1.7). A permutation σ is called an even permutation if $\chi(\sigma) = 1$, and an odd permutation if $\chi(\sigma) = -1$.

The key property of the character function is that $\chi(\sigma_2 \circ \sigma_1) = \chi(\sigma_2)\chi(\sigma_1)$. That is, *the character of a product equals the product of the characters*. This follows directly from (8.5). In Example 122 σ_a , σ_d and σ_e are even permutations whereas σ_b , σ_c and σ_f are odd permutations.

To determine $\chi(\sigma)$ for a given permutation σ , it is not necessary to compute $D(\sigma)$ first, and then apply the definition (8.6). There are some general rules for particular kinds of permutations.

Definition 86 (Pair Permutations). *For each $i < j$ in $\{1, 2, \dots, n\}$ the pair permutation $\sigma_{i,j}$ is defined by*

$$\sigma_{i,j}(i) = j, \quad \sigma_{i,j}(j) = i \quad \text{and} \quad \sigma_{i,j}(k) = k \quad \text{for } k \neq i, j. \quad (8.7)$$

It is called an adjacent pair permutation in case $j = i + 1$ for $i < n$, or if $(i, j) = (n, 1)$; i.e., if j follows i in the cyclic order on $\{1, \dots, n\}$.

Example 123. For $n = 4$, $\sigma_{2,4} = \begin{array}{cccc} 1 & 2 & 3 & 4 \\ & 1 & 4 & 3 & 2 \end{array}$.

Notice that each pair permutation is its own inverse – applying it twice swaps the reversed pair back into place. Next notice that for each adjacent pair permutation $\sigma_{i,i+1}$, $D(\sigma_{i,i+1}) = 1$, and hence $\chi(\sigma_{i,i+1}) = -1$.

Lemma 24. *For any $i < j$, $\sigma_{i,j}$ can be written as the product of $2k - 1$ adjacent pair permutations where $k = j - i$. In particular, for each pair permutation $\sigma_{i,j}$, $\chi(\sigma_{i,j}) = -1$.*

Proof. Write $j = i + k$. Then one can “move” i to the right of j using k adjacent pair permutations. One can then move j back to the i th spot with $k - 1$ pair permutations. Only $k - 1$ are required, because the last pair permutation used to move i into the j th place already put j one place to the left of i .

Finally, since the character of $\sigma_{i,j}$ is the product of the characters of $2k - 1$ adjacent pair permutations, $\chi(\sigma_{i,j}) = (-1)^{2k-1} = -1$. \square

We summarize the discussion in the following theorem:

Theorem 84 (Properties of the character). *For any two permutations σ_1 and σ_2 of $\{1, 2, \dots, n\}$,*

$$\chi(\sigma_2 \circ \sigma_1) = \chi(\sigma_2)\chi(\sigma_1). \quad (8.8)$$

Moreover, for any pair permutation $\sigma_{i,j}$,

$$\chi(\sigma_{i,j}) = -1. \quad (8.9)$$

The theorem gives us a convenient way to compute $\chi(\sigma)$: Bring the sequence $(1, 2, \dots, n)$ into the order $(\sigma(1), \sigma(2), \dots, \sigma(n))$ by swapping pairs; that is, by pair permutations. Then $\chi(\sigma)$ is the product of the characters of these pairs permutations, so it is $(-1)^\ell$, where ℓ is the number of pair permutations you used.

Example 124 (Computing $\chi(\sigma)$ counting pair permutations). Consider $\sigma = \begin{array}{cccc} 1 & 2 & 3 & 4 \\ & 4 & 1 & 3 & 2 \end{array}$. We can

transform $(1, 2, 3, 4)$ to $(4, 1, 3, 2)$ using pair permutations as follows:

$$(1, 2, 3, 4) \rightarrow (4, 2, 3, 1) \rightarrow (4, 1, 3, 2)$$

or as well by

$$(1, 2, 3, 4) \rightarrow (1, 2, 4, 3) \rightarrow (1, 4, 2, 3) \rightarrow (4, 1, 2, 3) \rightarrow (4, 1, 3, 2)$$

In the first case we used 2 pair permutations, and in the second case we used 4. Either way, we see $\chi(\sigma) = (-1)^2 = (-1)^4 = 1$, so σ is even.

You might wonder why we did not define $\chi(\sigma)$ to be $(-1)^\ell$ where ℓ is the number of “pair swaps” required to produce σ . The analysis we have done shows that it is impossible to write any permutation σ as a product of both an even and an odd number of pair permutations. However, this is not obvious. If it were not true, the parity of a permutation σ – even or odd – would not be well-defined, and we could not use it to define a function on the permutation group.

At this point we have covered as much of the theory of the permutation group as we shall use in explaining the theory of determinants. However, the permutation group is such a fundamental example of a transformation group, and the notion of a transformation group is so essential to modern analysis and geometry, that it is worthwhile to go somewhat further with the theory of permutations, and to study \mathcal{S}_n as a metric space. We do this in the next subsection.

8.1.3 The permutation group as a metric space

Definition 87 (Distance in \mathcal{S}_n). Let ϱ be the function on $\mathcal{S}_n \times \mathcal{S}_n$ given by

$$\varrho(\sigma_1, \sigma_2) = D(\sigma_1^{-1} \circ \sigma_2) .$$

This function is called the length function or distance function on \mathcal{S}_n .

It is not hard to see that the length function we have just defined is a metric on \mathcal{S}_n . That is, it satisfies the three requirements of a metric:

- (1) For all $\sigma_1, \sigma_2 \in \mathcal{S}_n$, $\varrho(\sigma_1, \sigma_2) \geq 0$, and $\varrho(\sigma_1, \sigma_2) = 0 \iff \sigma_1 = \sigma_2$.
- (2) For all $\sigma_1, \sigma_2 \in \mathcal{S}_n$, $\varrho(\sigma_1, \sigma_2) = \varrho(\sigma_2, \sigma_1)$.
- (3) For all $\sigma_1, \sigma_2, \sigma_3 \in \mathcal{S}_n$, $\varrho(\sigma_1, \sigma_3) \leq \varrho(\sigma_1, \sigma_2) + \varrho(\sigma_2, \sigma_3)$.

To see that this is the case, note for (1) that ϱ is defined to be a non-negative integer, and $D(\sigma_1^{-1} \circ \sigma_2) = 0$ if and only if there are “no crossings” in $\sigma_1^{-1} \circ \sigma_2$, which is the case if and only if $\sigma_1^{-1} \circ \sigma_2$ is the identity, which is the case if and only if $\sigma_1 = \sigma_2$. For (2), Note that

$$(\sigma_1^{-1} \circ \sigma_2)^{-1} = \sigma_2^{-1} \circ \sigma_1$$

and since, by Lemma 22, D is unaffected by taking inverses,

$$\varrho(\sigma_1, \sigma_2) = D(\sigma_1^{-1} \circ \sigma_2) = D((\sigma_1^{-1} \circ \sigma_2)^{-1}) = D(\sigma_2^{-1} \circ \sigma_1) = \varrho(\sigma_2, \sigma_1) .$$

Finally, for (3) we use (8.4) and the fact that, due to the associative nature of composition,

$$\sigma_1^{-1} \circ \sigma_3 = (\sigma_1^{-1} \circ \sigma_2) \circ (\sigma_2^{-1} \circ \sigma_3) .$$

Thus, by (8.4), since $2c \geq 0$ for all non-negative integers c ,

$$\varrho(\sigma_1, \sigma_3) = D(\sigma_1^{-1} \circ \sigma_3) = D((\sigma_1^{-1} \circ \sigma_2) \circ (\sigma_2^{-1} \circ \sigma_3)) \leq D(\sigma_1^{-1} \circ \sigma_2) + D(\sigma_2^{-1} \circ \sigma_3) = \varrho(\sigma_1, \sigma_2) + \varrho(\sigma_2, \sigma_3) .$$

We now explain how one can think of $\varrho(\sigma_1, \sigma_2)$ as the *length of the shortest path in \mathcal{S}_n from σ_1 to σ_2* . Given any $\sigma \in \mathcal{S}_n$, consider the set of permutations

$$\{\sigma \circ \tau : \tau \text{ is an adjacent pair permutation}\}$$

We call this set the set of the *nearest neighbors* of σ in \mathcal{S}_n . Now think of “moving” from σ to $\sigma \circ \tau$, where τ is an adjacent pair transposition, as a “step” from σ to one of its nearest neighbors. By a *path in \mathcal{S}_n from σ_1 to σ_2* , we mean a sequence of such steps starting at σ_1 and ending at σ_2 .

Definition 88 (Paths in \mathcal{S}_n). *For any σ_1 and σ_2 in \mathcal{S}_n , a path from σ_1 to σ_2 is a sequence $\{\tau_1, \dots, \tau_\ell\}$ of adjacent pair permutations such that*

$$\sigma_2 = \sigma_1 \circ \tau_1 \cdots \circ \tau_\ell.$$

For example, if $\{\tau_1, \tau_2, \tau_3\}$ is a path from σ_1 to σ_2 , then the sequences of steps

$$\sigma_1 \longrightarrow \sigma_1 \circ \tau_1 \longrightarrow \sigma_1 \circ \tau_1 \circ \tau_2 \longrightarrow \sigma_1 \circ \tau_1 \circ \tau_2 \circ \tau_3 = \sigma_2$$

is a sequence of “one step moves between nearest neighbors” that starts at σ_1 and ends at σ_2 .

Theorem 85 (The metric in \mathcal{S}_n as a minimal path length). *For each $\sigma_1, \sigma_2 \in \mathcal{S}_n$, there is a path from σ_1 to σ_2 , and*

$$\varrho(\sigma_1, \sigma_2) = \min\{\ell : \text{there exists a path } \{\tau_1, \dots, \tau_\ell\} \text{ from } \sigma_1 \text{ to } \sigma_2\}.$$

Theorem 85 says that for each $\sigma_1, \sigma_2 \in \mathcal{S}_n$, there is a way to get from σ_1 to σ_2 by making a finite number of steps from one nearest neighbor to another, and that $\varrho(\sigma_1, \sigma_2)$ is the *least* number of such steps in which this can be done. The following lemma is the key to the proof.

Lemma 25 (Reduction lemma). *For all $\sigma \in \mathcal{S}_n$ except the identity, there is some k with $1 \leq k \leq n-1$ such that $\sigma(k) > \sigma(k+1)$. For any such k , let τ be the adjacent pair permutation $\tau = \sigma_{k,k+1}$. Then*

$$D(\sigma \circ \tau) = D(\sigma) - 1.$$

Proof. Suppose for each $i = 1, \dots, n-1$, $\sigma(i+1) > \sigma(i)$. Then

$$\sigma(1) < \sigma(2) < \cdots < \sigma(n).$$

The only order preserving permutation is the identity, and since σ is not the identity, there is some $k \in \{1, \dots, n-1\}$ such that $\sigma(k) > \sigma(k+1)$. Let τ denote any adjacent pair permutation $\sigma_{k,k+1}$ for some such value of k .

Define the following sets of ordered pair (i, j) :

$$\begin{aligned} A &:= \{(i, j) : i < k, j > k+1\} \\ B &:= \{(i, j) : j = k \text{ or } k+1, j > k+1\} \\ C &:= \{(i, j) : i < k, j = k \text{ or } k+1\}. \end{aligned}$$

The sets A, B, C are disjoint from each other and from $\{(k, k+1)\}$, and $A \cup B \cup C \cup \{(k, k+1)\}$ is the set of all ordered pairs (i, j) with $i < j$.

Note that for $(i, j) \in A$, $(\sigma(i), \sigma(j)) = (\sigma \circ \tau(i), \sigma \circ \tau(j))$. Hence the image of A under f_σ is the same as the image of A under $f_{\sigma \circ \tau}$, and so σ and $\sigma \circ \tau$ reverse the same number of pairs in A .

Note that for $(i, j) \in B$,

$$(\sigma(i), \sigma(k)) = (\sigma \circ \tau(i), \sigma \circ \tau(k+1)) \quad \text{and} \quad (\sigma(i), \sigma(k+1)) = (\sigma \circ \tau(i), \sigma \circ \tau(k)) .$$

Hence the image of B under f_σ is the same as the image of B under $f_{\sigma \circ \tau}$, and so σ and $\sigma \circ \tau$ reverse the same number of pairs in B .

Note that for $(i, j) \in C$,

$$(\sigma(k), \sigma(j)) = (\sigma \circ \tau(k+1), \sigma \circ \tau(j)) \quad \text{and} \quad (\sigma(k+1), \sigma(j)) = (\sigma \circ \tau(k), \sigma \circ \tau(j)) .$$

Hence the image of C under f_σ is the same as the image of C under $f_{\sigma \circ \tau}$, and so σ and $\sigma \circ \tau$ reverse the same number of pairs in C .

Finally, by the choice of k , σ reverses $(k, k+1)$, but then by the definition of τ , $\sigma \circ \tau$ does not. Hence $\sigma \circ \tau$ reverses exactly one fewer pair than does σ . \square

Proof of Theorem 85. First, suppose that $\{\tau_1, \dots, \tau_\ell\}$ is a path of length ℓ from σ_1 to σ_2 . Then $\sigma_2 = \sigma_1 \circ \tau_1 \circ \dots \circ \tau_\ell$. Therefore, $\sigma_1^{-1} \sigma_2 = \tau_1 \circ \dots \circ \tau_\ell$ and so $D(\sigma_1^{-1} \sigma_2) = D(\tau_1 \circ \dots \circ \tau_\ell)$. Then by Lemma 23

$$\begin{aligned} D(\tau_1 \circ \dots \circ \tau_\ell) &\leq D(\tau_1) + D(\tau_2 \circ \dots \circ \tau_\ell) \\ &= 1 + D(\tau_2 \circ \dots \circ \tau_\ell) \end{aligned}$$

since $D(\tau) = 1$ for any adjacent pair permutation. Proceeding inductively, we find

$$D(\tau_1 \circ \dots \circ \tau_\ell) \leq \ell .$$

Hence, any path from σ_1 to σ_2 takes at least $D(\sigma_1^{-1} \sigma_2)$ steps.

On the other hand, by Lemma 25, as long as $\sigma_1 \neq \sigma_2$, or what is the same, $D(\sigma_2^{-1} \circ \sigma_1) \neq 0$, there exists an adjacent pair permutation τ_1 such that

$$D(\sigma_2^{-1} \circ \sigma_1 \circ \tau_1) = D(\sigma_2^{-1} \circ \sigma_1) - 1 .$$

Next as long as $D(\sigma_2^{-1} \circ \sigma_1 \circ \tau_1) \neq 0$, there exists an adjacent pair permutation τ_2 such that

$$\begin{aligned} D(\sigma_2^{-1} \circ \sigma_1 \circ \tau_1 \circ \tau_2) &= D(\sigma_2^{-1} \circ \sigma_1 \circ \tau_1) - 1 \\ &= D(\sigma_2^{-1} \circ \sigma_1) - 2 . \end{aligned}$$

Continuing this way, we find a sequence $\{\tau_1, \dots, \tau_{D(\sigma_2^{-1} \circ \sigma_1)}\}$ adjacent pair permutations such that

$$D(\sigma_2^{-1} \circ \sigma_1 \circ \tau_1 \circ \dots \circ \tau_{D(\sigma_2^{-1} \circ \sigma_1)}) = 0 .$$

But this means that $\sigma_2^{-1} \circ \sigma_1 \circ \tau_1 \circ \dots \circ \tau_{D(\sigma_2^{-1} \circ \sigma_1)}$ is the identity, and therefore,

$$\sigma_2 = \sigma_1 \circ \tau_1 \circ \dots \circ \tau_{D(\sigma_2^{-1} \circ \sigma_1)} .$$

Hence, there exists a path from σ_1 to σ_2 of length $D(\sigma_2^{-1} \circ \sigma_1)$. Note that $(\sigma_2^{-1} \circ \sigma_1)^{-1} = \sigma_1^{-1} \circ \sigma_2$, and then by Lemma 22, $D(\sigma_2^{-1} \circ \sigma_1) = D(\sigma_1^{-1} \circ \sigma_2)$. Therefore, there exists a path from σ_1 to σ_2 consisting of $D(\sigma_1^{-1} \circ \sigma_2)$ steps. By what we have proved above, this is the least number of steps taken in any path from σ_1 to σ_2 . \square

8.2 Algebraic properties of the determinant

8.2.1 The determinant formula

We are going to break down the formula for the determinant into “building blocks”. The building blocks will be two simple functions that we will combine to form the determinant function. The first one is the character function on the permutations. Here is the second one:

Definition 89 (The function $A \mapsto \sigma(A)$). *For any $n \times n$ matrix A , and any permutation σ on $\{1, \dots, n\}$, define the number $\sigma(A)$ by*

$$\sigma(A) := A_{\sigma(1),1}A_{\sigma(2),2} \cdots A_{\sigma(n),n} = \prod_{j=1}^n A_{\sigma(j),j}. \quad (8.10)$$

Definition 90 (The determinant function). *The determinant function $\det(A)$ on the set of $n \times n$ matrices is defined by*

$$\det(A) = \sum_{\sigma \in S_n} \chi(\sigma)\sigma(A). \quad (8.11)$$

Let us first check that this definition gives us what we expect for $n = 2$ and $n = 3$.

Example 125 (2×2 determinants). *Consider the general 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. There are only two permutations of $\{1, 2\}$ to consider, namely*

$$\sigma_1 = \begin{array}{cc} 1 & 2 \\ 1 & 2 \end{array} \quad \text{and} \quad \sigma_2 = \begin{array}{cc} 1 & 2 \\ 2 & 1 \end{array}.$$

Clearly $\chi(\sigma_1) = 1$ and $\chi(\sigma_2) = -1$. Hence $\det(A) = A_{1,1}A_{2,2} - A_{2,1}A_{1,2} = ad - bc$, which is the usual formula.

Example 126 (3×3 determinants). *Consider a general 3×3 matrix A . We have already worked out a list of the six permutations of $\{1, 2, 3\}$ in (8.1) of the previous section, and computed the characters of each of them. In the 3×3 case then, the definition (8.11) leads to*

$$\begin{aligned} \det(A) &= A_{1,1}A_{2,2}A_{3,3} + A_{2,1}A_{3,2}A_{1,3} + A_{3,1}A_{1,2}A_{2,3} \\ &\quad - A_{2,1}A_{1,2}A_{3,3} - A_{1,1}A_{3,2}A_{2,3} - A_{3,1}A_{2,2}A_{1,3}. \end{aligned}$$

This too is reassuring – the formula (8.11) leads us to the usual formula for 3×3 determinants.

Since there are $n!$ permutations, direct application of (8.11) requires a large computational effort for large n . However, there is one important case in which one *one single* permutation contributes to the sum, and then (8.11) is easy to use:

Theorem 86. Let A be an upper-triangular $n \times n$ matrix. That is $A_{i,j} = 0$ whenever $i > j$. Then

$$\det(A) = \prod_{j=1}^n A_{j,j} . \quad (8.12)$$

In particular, the determinant of the $n \times n$ identity matrix, $I_{n \times n}$ is 1:

$$\det(I_{n \times n}) = 1 . \quad (8.13)$$

Proof. Note that the product $\sigma(A) = A_{\sigma(1),1} \dots A_{\sigma(n),n}$ has a zero factor in case $\sigma(j) > j$ for any j . But the only permutation σ for which $\sigma(j) \leq j$ for all j is the identity permutation. Hence only this single term contributes to the sum in (8.11), and this proves (8.12). Then (8.13) follows, as a special case, from (8.12). \square

Before stating the next theorem, it will be useful to shift our way of thinking about determinants: Since an $n \times n$ matrix is an ordered list of n vectors in \mathbb{R}^n , we can think of the determinant function as a function of an ordered list of n vectors in \mathbb{R}^n . Slightly abusing notation, we may write

$$\det(\mathbf{v}_1, \dots, \mathbf{v}_n) = \det([\mathbf{v}_1, \dots, \mathbf{v}_n])$$

where, as usual, $[\mathbf{v}_1, \dots, \mathbf{v}_n]$ is the $n \times n$ matrix whose j th column is \mathbf{v}_j . There is another way to do this using the rows, but we will see below that

$$\det \left(\begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_n \end{bmatrix} \right) = \det([\mathbf{v}_1, \dots, \mathbf{v}_n]) , \quad (8.14)$$

so that for this purpose, it does not matter if you treat the vectors as columns or rows.

The determinant function, thought of as a function on ordered lists of n vectors in \mathbb{R}^n , has an important property: It is *linear* in each of the entries. That is,

$$\det(s\mathbf{x} + t\mathbf{y}, \mathbf{v}_2, \mathbf{v}_n) = s \det(\mathbf{x}, \mathbf{v}_2, \dots, \mathbf{v}_n) + t \det(\mathbf{x}, \mathbf{v}_2, \dots, \mathbf{v}_n) , \quad (8.15)$$

with an analogous formula for any other j in case $\mathbf{v}_j = s\mathbf{x} + t\mathbf{y}$. (See 8.18) below).

Theorem 87 (Properties of the determinant). Let \det be the numerically valued function on the $n \times n$ matrices defined by (8.11). Then:

- (1) The determinant is invariant under transposition; i.e., $\det(A^T) = \det(A)$.
- (2) $\det(A)$ changes sign when any two rows of A are interchanged, and when any two columns are interchanged. If any two rows of A are equal, or if any two columns of A are equal, $\det(A) = 0$.
- (3) $\det(A)$ is linear in each row of A , and also in each column of A . For each $k \neq \ell$, if one adds any multiple of row (or column) k to row (or column) ℓ of A , the determinant is unchanged.
- (4) For any two $n \times n$ matrices A and B , $\det(AB) = \det(A)\det(B)$.

Proof. We first prove (1). Let τ be any permutation on $\{1, \dots, n\}$. For any n numbers a_1, \dots, a_n , $\prod_{j=1}^n a_j = \prod_{j=1}^n a_{\tau(j)}$. The only difference between the left and the right is that we are doing the multiplication in a different order, but for multiplication of numbers, the order does not matter.

Therefore, for any two permutations σ, τ on $\{1, \dots, n\}$, and any $n \times n$ matrix A ,

$$\sigma(A) = \prod_{j=1}^n A_{\sigma(j), j} = \prod_{j=1}^n A_{\sigma(\tau(j)), \tau(j)} .$$

Now taking $\tau = \sigma^{-1}$, we have $\sigma(A) = \prod_{j=1}^n A_{j, \tau(j)} = \prod_{j=1}^n A_{\tau(j), j}^T = \tau(A^T)$. Since the character of the identity permutation is 1, for $\tau = \sigma^{-1}$, $\chi(\sigma \circ \tau) = 1$ and so $\chi(\tau) = \chi(\sigma)$. Therefore,

$$\det(A) = \sum_{\sigma} \chi(\sigma) \sigma(A) = \sum_{\tau} \chi(\tau) \tau(A^T) = \det(A^T) .$$

This proves (1).

To prove (2), suppose that B is obtained from A by interchanging the k th and ℓ th rows of A . Then we have to show that $\det(B) = -\det(A)$.

To see this, note that $B_{i,j} = A_{\sigma_{k,\ell}(i), j}$, and hence, for any permutation σ , $\sigma(B) = (\sigma \circ \sigma_{k,\ell})(A)$. Since $\sigma_{k,\ell}$ is a pair permutation, $\chi(\sigma \circ \sigma_{k,\ell}) = -\chi(\sigma)$. Therefore,

$$\det(B) = \sum_{\sigma} \chi(\sigma) \sigma(B) = - \sum_{\sigma} \chi(\sigma \circ \sigma_{k,\ell})(\sigma \circ \sigma_{k,\ell})(A) . \quad (8.16)$$

Let τ denote the permutation $\tau := \sigma \circ \sigma_{k,\ell}$. Since $\sigma_{k,\ell}$ is its own inverse, $\sigma = \tau \circ \sigma_{k,\ell}$. That is, the map $\sigma \mapsto \tau := \sigma \circ \sigma_{k,\ell}$ is a one-to-one map of the set of permutations on $\{1, \dots, n\}$. Hence

$$\sum_{\sigma} \chi(\sigma \circ \sigma_{k,\ell})(\sigma \circ \sigma_{k,\ell})(A) = \sum_{\tau} \chi(\tau) \tau(A) = \det(A) . \quad (8.17)$$

(In the sum on the middle, we are summing over τ instead of σ , but τ is just a “dummy” variable; we are summing over *all* permutations on $\{1, \dots, n\}$. Hence $\sum_{\tau} \chi(\tau) \tau(A) = \det(A)$). Combining (8.16) and (8.17) we have $\det(B) = -\det(A)$, that the sign of the permuation changes if two rows are interchanged. By (1), the same is true when any two columns are swapped since the rows of A^T are the columns of A . By what we have just proved, if any two columns or rows are equal, the determinat must change sign if they are swapped. Since swapping identical rows or columns does nothing, the detemrinant is unchanged. The only number that is minus itself is 0, and so in this case we must have $\det(A) = 0$, and this proves (2).

To prove (3), we have to show that if

$$\mathbf{r}_i = \alpha \mathbf{v} + \mathbf{w}$$

then

$$\det \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \alpha \mathbf{v} + \mathbf{w} \\ \vdots \\ \mathbf{r}_n \end{pmatrix} = \alpha \det \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{v} \\ \vdots \\ \mathbf{r}_n \end{pmatrix} + \beta \det \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{w} \\ \vdots \\ \mathbf{r}_n \end{pmatrix} . \quad (8.18)$$

This is true since each product $\sigma(A) = A_{\sigma(1),1}A_{\sigma(2),2}\cdots A_{\sigma(n),n}$ contains exactly one factor coming from the i th row, and hence is a linear function of the entries of the i th row. By definition, $\det(A)$ is a linear combination of the $\sigma(A)$. A linear combination of linear functions is linear, and so the determinant is a linear function of the entries of the i th row.

To prove the final part of (3) for rows, consider (8.18) in the case in which the additive term on the left is in the i th place, $\alpha = 1$, $\mathbf{v} = \mathbf{r}_i$ and $\mathbf{w} = \mathbf{r}_\ell$ for some $\ell \neq i$. Then second the matrix on the right has two rows equal to \mathbf{r}_ℓ , and hence its determinant is zero, by (2). The first matrix on the right is the original matrix A , and we have set $\alpha_4 = 1$, so this proves the assertion for rows. The validity for columns now follows from (1). This proves (3).

We now prove (4). Let $C = AB$. Then

$$\begin{aligned}\sigma(C) &= \prod_{j=1}^n \sum_{k=1}^n A_{\sigma(j),k} B_{k,j} \\ &= \sum_{k_1=1}^n \cdots \sum_{k_n=1}^n (A_{\sigma(1),k_1} B_{k_1,1}) \cdots (A_{\sigma(n),k_n} B_{k_n,n}) \\ &= \sum_{k_1=1}^n \cdots \sum_{k_n=1}^n (A_{\sigma(1),k_1} \cdots A_{\sigma(n),k_n})(B_{k_1,1} \cdots B_{k_n,n})\end{aligned}$$

Therefore,

$$\det(C) = \sum_{\sigma} \sigma(C) \chi(\sigma) = \sum_{k_1=1}^n \cdots \sum_{k_n=1}^n \left(\sum_{\sigma} A_{\sigma(1),k_1} \cdots A_{\sigma(n),k_n} \chi(\sigma) \right) (B_{k_1,1} \cdots B_{k_n,n}).$$

Now notice that $\sum_{\sigma} A_{\sigma(1),k_1} \cdots A_{\sigma(n),k_n} \chi(\sigma)$ is the determinant of the $n \times n$ matrix whose j th column is column k_j of A . By (2), this determinant is zero if $k_j = k_\ell$ for any $j \neq \ell$. Hence, of the n^n terms in the sum over k_1, \dots, k_n , the only terms that can contribute to the sum are the $n!$ terms in which $k_j \neq k_\ell$ whenever $j \neq \ell$. This is the cases exactly when for some permutation τ , $k_j = \tau(j)$ for $j = 1, \dots, n$, and evidently different permutations give different terms. Hence we may replace the sum over k_1, \dots, k_n by a sum over permutation τ :

$$\begin{aligned}\det(C) &= \sum_{\tau} \left(\sum_{\sigma} A_{\sigma(1),\tau(1)} \cdots A_{\sigma(n),\tau(n)} \chi(\sigma) \right) (B_{\tau(1),1} \cdots B_{\tau(n),n}) \\ &= \sum_{\tau} \left(\sum_{\sigma} A_{\sigma(1),\tau(1)} \cdots A_{\sigma(n),\tau(n)} \chi(\sigma) \right) \tau(B)\end{aligned}\tag{8.19}$$

Now define $\hat{\sigma} = \sigma \circ \tau^{-1}$, and note that $\chi(\hat{\sigma}) = \chi(\sigma)\chi(\tau)^{-1}$, so that

$$\chi(\sigma) = \chi(\hat{\sigma})\chi(\tau).\tag{8.20}$$

Also note that $A_{\sigma(1),\tau(1)} \cdots A_{\sigma(n),\tau(n)} = A_{\hat{\sigma}(\tau(1)),\tau(1)} \cdots A_{\hat{\sigma}(\tau(n)),\tau(n)} = A_{\hat{\sigma}(1),1} \cdots A_{\hat{\sigma}(n),n}$ because, one more, the order does not matter in a product of numbers, so the product is the same for all τ . In short,

$$A_{\sigma(1),\tau(1)} \cdots A_{\sigma(n),\tau(n)} = \hat{\sigma}(A).\tag{8.21}$$

Finally notice that for each fixed τ , as σ ranges over the full set of permutations, so does $\hat{\sigma}$. Thus, using (8.20) and (8.21) in (8.19), we obtain $\det(C) = \sum_{\hat{\sigma}} \hat{\sigma}(A) \chi(\hat{\sigma}) \sum_{\tau} \tau(B) \chi(\tau) = \det(A) \det(B)$. \square

Theorem 88. *An $n \times n$ matrix A is invertible if and only if $\det(A) \neq 0$, in which case $\det(A^{-1}) = (\det(A))^{-1}$.*

Proof. Suppose A is invertible. Then $AA^{-1} = I_{n \times n}$, and hence by (4) of Theorem 87, and (8.13)

$$1 = \det(I_{n \times n}) = \det(AA^{-1}) = \det(A)\det(A^{-1}).$$

This shows that neither of $\det(A)$ nor $\det(A^{-1})$ equals zero, and that these two numbers are inverse to each other.

It remains to show that when A is not invertible, then $\det(A) = 0$. Recall that A is not invertible if and only if the columns of A are linearly independent. Hence if A is not invertible, either the first column is zero, or else for some j with $2 \leq j \leq n$, there are numbers t_1, \dots, t_{j-1} such that

$$\mathbf{v}_j = \sum_{\ell=1}^{j-1} t_\ell \mathbf{v}_\ell.$$

Then by subtracting $t_\ell \mathbf{v}_\ell$ from column j for each $\ell = 1, \dots, j-1$, we produce a zero column, but we have not changed the value of the determinant by (3) of Theorem 87. Clearly, the determinant of a matrix with a zero column is zero. \square

We can combine the results of the last two theorems to obtain an effective method for computing determinants. We subtract multiples of one row from another, swapping rows if need be, to produce an upper triangular matrix. The determinant of this is the product of its diagonal entries. This is the same as the original determinant if we did not swap rows at all, or made an even number of swaps, and it is minus the original determinant if we made an odd number of row swaps.

Example 127 (Computing determinants using row operations). *Consider the matrix*

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}.$$

Then subtracting multiples of one row from another, we transform

$$A \rightarrow \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 5 \\ 0 & 2 & 12 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 5 \\ 0 & 0 & 2 \end{bmatrix}$$

By (3), the determinant of the upper triangular matrix on the right is 2. But since our row operations did not change the value of the determinant, this is also the value of $\det(A)$. Hence $\det(A) = 2$. You can readily check that this is what the usual formula gives as well.

We have seen that if a matrix V is orthogonal, then V^T is also orthogonal, and is the inverse of V . Moreover if V and W are both orthogonal then the both take orthonormal sets to orthonormal sets. Hence $\{WV\mathbf{e}_1, \dots, WV\mathbf{e}_n\}$ is orthonormal, and it is the set of columns of WV . Hence WV is orthogonal. That is, the product of any two orthogonal matrices is again an orthogonal matrix. This means that the set of $n \times n$ orthogonal matrices, regarded as a set of transformations of \mathbb{R}^n is a transformation group.

Definition 91 (The orthogonal group on \mathbb{R}^n). *The set of all $n \times n$ orthogonal matrices is called the orthogonal group on \mathbb{R}^n , and is denoted by $O(n)$.*

Theorem 89 (Determinants of orthogonal matrices). *Let $V \in O(n)$. Then*

$$\det(V) = \pm 1 .$$

The set of matrices $V \in O(n)$ such that $\det(V) = 1$, regarded as a set of transformations of \mathbb{R}^n , forms a transformation group on R^n .

Proof. For all $V \in O(n)$, $I = V^T V$. This

$$1 = \det(I) = \det(V^T V) = \det(V^T) \det(V) = (\det(V))^2 ,$$

where we have used Theorem 87. The only solutions of the equation $x^2 = 1$ are ± 1 , and so $\det(V) = \pm 1$. \square

Now suppose $\det(V) = 1$. Then by Theorem 87,

$$\det(V^{-1}) = \det(V^T) = \det(V) = 1 .$$

Hence the inverse of V has the same property. Next, let $V, W \in O(n)$ be such that $\det(V) = \det(W) = 1$. Then by Theorem 87,

$$\det(WV) = \det(W) \det(V) = 1 .$$

Hence the subset of $O(n)$ consisting of matrices with unit determinant is closed under taking inverses and products. It is therefore a transformation group, and, as such, a *subgroup* of $O(n)$.

Definition 92 (The special orthogonal group on \mathbb{R}^n). *The subset of $O(n)$ consisting of orthogonal matrices V with $\det(V) = 1$ is called the special orthogonal group on \mathbb{R}^n , and is denoted by $SO(n)$.*

Example 128 (Two dimensional orthogonal matrices). *Let $U = [\mathbf{u}_1, \mathbf{u}_2] \in O(2)$. Then \mathbf{u}_1 is a unit vector. Hence*

$$\mathbf{u}_1 = (\cos \theta, \sin \theta)$$

for some θ . Since \mathbf{u}_2 must be a unit vector orthogonal to \mathbf{u}_1 , there are only two choices for \mathbf{u}_2 :

$$\mathbf{u}_2 = (-\sin \theta, \cos \theta) \quad \text{or else} \quad \mathbf{u}_2 = (\sin \theta, -\cos \theta) .$$

Thus either we have

$$U = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad \text{or else} \quad U = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix} . \quad (8.22)$$

Note that

$$\det \left(\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \right) = 1 \quad \text{and} \quad \det \left(\begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix} \right) = -1 .$$

Thus, the matrices in $SO(2)$ are precisely the matrices on the left in (8.22), and we recognize these as the two dimensional rotation matrices.

The matrices on the right in (8.22) reflection matrices. Indeed, let $\mathbf{u} = (\cos(\theta/2), \sin(\theta/2)\theta)$. Then the Householder reflection in \mathbb{R}^2 given by \mathbf{u} has the matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} \cos^2(\theta/2) & \cos(\theta/2)\sin(\theta/2) \\ \cos(\theta/2)\sin(\theta/2) & \sin^2(\theta/2) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix},$$

by the double-angle formulas. Thus the matrices in $O(n)$ that are not in $SO(n)$ are precisely the reflection matrices.

Example 129 (Three dimensional orthogonal matrices). Let $U = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3] \in O(3)$, so that, by definition, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an orthonormal basis of \mathbb{R}^3 .

We have seen in Chapter One that there is a linear transformation \mathbf{f} from \mathbb{R}^3 to \mathbb{R}^3 that is the composition of at most 3 Householder reflections such that $(\mathbf{e}_j) = \mathbf{u}_j$ for $j = 1, 2, 3$. This means that the matrix representing \mathbf{f} is the matrix U , and hence U is the product of at most three matrices representing Householder reflections. We have seen that whenever $\mathbf{h}_{\mathbf{u}}$ is a Householder reflection, and $H_{\mathbf{u}} := [\mathbf{h}_{\mathbf{u}}(\mathbf{e}_1), \mathbf{h}_{\mathbf{u}}(\mathbf{e}_2), \mathbf{h}_{\mathbf{u}}(\mathbf{e}_3)]$ is the 3×3 matrix representing $\mathbf{h}_{\mathbf{u}}$,

$$\det(H_{\mathbf{u}}) = \det([\mathbf{h}_{\mathbf{u}}(\mathbf{e}_1), \mathbf{h}_{\mathbf{u}}(\mathbf{e}_2), \mathbf{h}_{\mathbf{u}}(\mathbf{e}_3)]) = \mathbf{h}_{\mathbf{u}}(\mathbf{e}_1) \cdot \mathbf{h}_{\mathbf{u}}(\mathbf{e}_2) \times \mathbf{h}_{\mathbf{u}}(\mathbf{e}_3) = -1.$$

Now suppose U is not the identity matrix. Then U is the product of either 1, 2 or 3 Householder reflection matrices. Since the determinant of each of these is -1 , by Theorem ??, $\det(U) = 1$ if and only if U is the product of exactly 2 Householder reflection matrices.

As we have seen in Chapter Two, the product of any two Householder reflections is a rotation: Each Householder reflection leaves a plane through the origin - the plane of reflection - unchanged. The two planes of reflection meet in a line through the origin which is left unchanged by the composition of the two reflections. This line is the axis of rotation. Thus, $SO(3)$ consists of the 3×3 rotation matrices. Every matrix $U \in O(3)$ that is not in $SO(3)$ is the product of some matrix in $SO(3)$ and a Householder reflection matrix.

8.3 The volume of sets in \mathbb{R}^n and the determinant

The standard unit cube in \mathbb{R}^n is the set C_{unit} deined by

$$C_{\text{unit}} := \left\{ \sum_{j=1}^n t_j \mathbf{e}_j : 0 \leq t_j \leq 1, j = 1, \dots, n \right\} \quad (8.23)$$

Notice that for $n = 2$, C_{unit} is the st of vectors of the form (s, t) , $0 \leq s, t \leq 1$, which is the unit square in the plane with its edges parallel to the coordinate axes and its lower left hand corner at the origin. It will be convenient to refer to C_{unit} as a “cube” regardless of the dimension.

Let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be a set of n vectors in \mathbb{R}^n . The parallelepiped P spanned by $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is the set

$$P := \left\{ \sum_{j=1}^n t_j \mathbf{v}_j : 0 \leq t_j \leq 1, j = 1, \dots, n \right\} \quad (8.24)$$

Let A be the matrix $A := [\mathbf{v}_1, \dots, \mathbf{v}_n]$. As we have seen, $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is linearly independent if and only if A is invertible. For any set $E \subset \mathbb{R}^n$, the *image of E under the linear transformation induced by A* is the set $A(E)$ given by

$$A(E) := \{A\mathbf{x} : \mathbf{x} \in E\}. \quad (8.25)$$

Notice that $\mathbf{x} \in C_{\text{unit}}$ if and only if $\mathbf{x} = \sum_{j=1}^n t_j \mathbf{e}_j$ where $0 \leq t_j \leq 1$ for each j . Then $A\mathbf{x} = \sum_{j=1}^n t_j A\mathbf{e}_j = \sum_{j=1}^n t_j \mathbf{v}_j$ so that $A(C_{\text{unit}})$ is precisely the parallelepiped defined in (8.24).

In this section, we shall first give a careful definition of volume in n -dimensional Euclidean space, and shall introduce the notion of a positively oriented basis in \mathbb{R}^n , which will generalize the notion of a right-handed basis in \mathbb{R}^3 . We shall then prove that $|\det(A)|$ is precisely the volume of the parallelepiped defined in (8.24), and that when A is invertible, so that $\det(A) \neq 0$, $\det(A)$ is positive if and only if the basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is positively oriented. The main tool from Linear algebra that we shall use to do this is the *Singular Value Decomposition*, which is introduced later in this section. We begin with the notion of volume in n -dimensional Euclidean space.

8.3.1 Volume in n -dimensional Euclidean space

A *cube with side length $h > 0$* in \mathbb{R}^n is a set C of the following form: There is an $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be an orthonormal basis of \mathbb{R}^n , and an $\mathbf{x}_0 \in \mathbb{R}^n$ such that $\mathbf{x} \in C$ if and only if

$$\mathbf{x} = \mathbf{x}_0 + \sum_{j=1}^n t_j \mathbf{u}_j \quad \text{where} \quad -h/2 \leq t_j \leq h/2 \quad \text{for all } j = 1, \dots, n.$$

Draw sketches for $n = 2$ and $n = 3$ to be sure you understand the general case. The point \mathbf{x}_0 is the center of the cube. There are $2n$ faces of the cube; for each j , one face consists of points with $t_j = h/2$, and the other with $t_j = -h/2$. The edges are the points where $n - 1$ faces meet. On an edge, for some $j = 1, \dots, n$ t_j takes values in the interval $[-h/2, h/2]$, while for all $k \neq j$, t_k has a fixed value that is either $h/2$ or $-h/2$. Thus each edge is a line segment of length h parallel to some \mathbf{u}_j . We define the *volume* of a cube of side length h in \mathbb{R}^n to be h^n . (Of course when $n = 2$, this is what we would usually call “area” and for $n = 1$ it is what we would usually call “length”, but the current terminology is convenient, and we will use it.)

Now let V be an orthogonal $n \times n$ matrix. Note the for all $1 \leq j, k \leq n$,

$$V\mathbf{u}_j \cdot V\mathbf{u}_k = \mathbf{u}_j \cdot V^T V\mathbf{u}_k = \mathbf{u}_j \cdot \mathbf{u}_k$$

since V is orthogonal. Therefore, $\{V\mathbf{u}_1, \dots, V\mathbf{u}_n\}$ is orthonormal whenever $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is orthonormal, and $V(C)$, the image of C under V , consists of the points \mathbf{y} of the form

$$V\mathbf{x}_0 + \sum_{j=1}^n t_j V\mathbf{u}_j \quad \text{where} \quad -h/2 \leq t_j \leq h/2 \quad \text{for all } j = 1, \dots, n.$$

Hence $V(C)$ is again a cube of side length h , and all such cubes have the same volume, namely h^n . Thus, the transformation of \mathbb{R}^n induced by an orthogonal matrix V has no effect on the volume of a cube.

Now consider any $E \subset \mathbb{R}^n$. A *cube packing* of E is any finite collection $\mathcal{P} := \{C_1, \dots, C_m\}$ of cubes in \mathbb{R}^n that intersect at most in boundary points, and such that for some fixed orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, each cube has edges that are parallel to one of the vectors in this basis, finally such that

$$\bigcup_{C \in \mathcal{P}} C \subset E.$$

We allow \mathcal{P} to be empty. Indeed if E is a subset of a proper subspace of \mathbb{R}^n , then \mathcal{P} cannot contain any cube of any side length $h > 0$.

A *cube-covering* of E is any finite collection $\mathcal{C} := \{C_1, \dots, C_m\}$ of cubes in \mathbb{R}^n such that for some fixed orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, each cube has edges that are parallel to one of the vectors in this basis, and such that

$$E \subset \bigcup_{C \in \mathcal{C}} C.$$

Definition 93 (Volume in \mathbb{R}^n). *Let $E \subset \mathbb{R}^n$. Then the volume of E , $\text{vol}(E)$ is defined if and only if*

$$\inf \left\{ \sum_{C \in \mathcal{C}} \text{vol}(C) , \mathcal{C} \text{ a cube covering of } E \right\} = \sup \left\{ \sum_{C \in \mathcal{C}} \text{vol}(C) , \mathcal{P} \text{ a cube packing of } E \right\},$$

and in this case we define $\text{vol}(E)$ to be this common value.

The first thing to observe, is that if C is a cube of side length h , it has a well-defined volume in the sense of Definition 93. In fact, if one fixes the orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, then however the edges of C are oriented, for each $\epsilon > 0$, there is a cube packing of C using cubes with edges parallel to the vectors in $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ such that the total volume of the cubes in this packing is at least $h^n - \epsilon$. Likewise, there is a cube covering of C by such cubes, including all of the cubes used in the cube packing, such that the total volume of the cubes is no more than $h^n + \epsilon$.

Since ϵ is arbitrary, it follows that C has a well defined volume in the sense of Definition 93, and that volume is h^n .

Now let E be a set that has a well-defined and finite volume $\text{vol}(E)$. For any $\epsilon > 0$, let $\mathcal{P} = \{C_1, \dots, C_m\}$ be a cube packing of E with total volume at least $\text{vol}(E) - \epsilon/2$. Pick any orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n , and for each $j = 1, \dots, m$, let \mathcal{P}_j be a cube-packing of C_j using cubes whose edges align with $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, and such that the total volume of \mathcal{P}_j is at least $\text{vol}(C_j) - \frac{\epsilon}{2m}$. Then $\bigcup_{j=1}^m \mathcal{P}_j$ is a cube packing of E with total volume at least $\text{vol}(E) - \epsilon$. That is, whenever E has a finite well-defined volume, that volume is equal to the supremum of the total cube packing volumes for cube packing \mathcal{P} using cubes that align with any given orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of \mathbb{R}^n . We now prove two important lemmas.

Lemma 26. *Let $E \subset \mathbb{R}^n$ have a well-defined and finite volume. Let V be any orthogonal $n \times n$ matrix, and let $V(E)$ be the image of E under V . Then $V(E)$ has a well defined volume, and $\text{vol}(V(E)) = \text{vol}(E)$.*

Proof. If $\mathcal{P} = \{C_1, \dots, C_m\}$ is a cube packing of E then $\{V(C_1), \dots, V(C_m)\}$ is evidently a cube-packing of $V(E)$ with the same total volume. Likewise, if $\tilde{\mathcal{P}} = \{\tilde{C}_1, \dots, \tilde{C}_{\tilde{m}}\}$ is a cube-packing of $V(E)$, then $\{V^T(\tilde{C}_1), \dots, V^T(\tilde{C}_{\tilde{m}})\}$ is a cube-packing of E with the same total volume. Similar reasoning applies to cube coverings. \square

Lemma 27. Let $E_1 \subset E_2 \subset \mathbb{R}^n$ both have a well-defined and finite volume. Then $\text{vol}(E_1) \leq \text{vol}(E_2)$.

Proof. An cube packing of E_1 is a cube packing of E_2 , and any cube covering of E_2 is a vube covering of E_1 . Taking the supremum over a larger set gives a value at least as large, and tkaing the imfimum over a larger set gives a value at least as low. \square

Our next lemma concerns *scaling transformations*: Let Λ be a diagonal $n \times n$ matrix with positive diagonal entries λ_j , $j = 1, \dots, n$. That is, $\Lambda i, j = \lambda_j$ for $i = j$, and $\Lambda i, j = 0$ for $i \neq j$. In this case we write

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n).$$

Lemma 28. Let $E \subset \mathbb{R}^n$ have a well-defined and finite volume. Let $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_j > 0$ for $j = 1, \dots, n$. Let $\Lambda(E)$ be the image of E under Λ . Then $V(E)$ has a well deifned volume, and $\text{vol}(\Lambda(E)) = \prod_{j=1}^n \lambda_j \text{vol}(E)$.

Proof. First suppose that each λ_j is ratinional. Then, taking a commmon denominator, we may assume that each λ_j is of the form p_j/q where q and p_1, \dots, p_n are positive integers. If C is any cube of side length h with edges that align with $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$, Then $\Lambda(C)$ is exactly the union of $\prod_{j=1}^n p_j$ cubes of volume $(h/q)^n$ that overlap at most on their boundaries. Hence

$$\text{vol}(\Lambda(C)) = \prod_{j=1}^n p_j (h/q)^n = \prod_{j=1}^n \lambda_j h^n = \prod_{j=1}^n \lambda_j \text{vol}(C). \quad (8.26)$$

Now let $\mathcal{P} = \{C_1, \dots, C_m\}$ be any cube packing of E that aligns with $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. Slightly shrinking each cube, making an aritrarily small decrease in the total volume, we may assume that none of them intersect at all. Then $\Lambda(E) = \cup_{j=1}^m \Lambda(C_j)$. For each $\Lambda(C_j)$, we have a cube packing that is also a cube covering by cubes that align with $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. Combining all of these, we obtain a cube packing of $\Lambda(E)$ that has total volume $\prod_{j=1}^n \lambda_j$ times the total volume of \mathcal{P} . Similar reasoning applies to cube coverings.

Now we remove the restriction that each λ_j is rational. For each j , let $\check{\lambda}_j$ and $\hat{\lambda}_j$ be rational and such that $\check{\lambda}_j \leq \lambda_j \leq \hat{\lambda}_j$. Define $\check{\Lambda}$ and $\hat{\Lambda}$ in the natural way. Then any cube C centered at the origin

$$\check{\Lambda}(C) \leq \Lambda(C) \leq \hat{\Lambda}(C).$$

By Lemma 27 and what we have proved so far,

$$\prod_{j=1}^n \check{\lambda}_j \text{vol}(C) \leq \text{vol}(\Lambda(C)) \leq \prod_{j=1}^n \hat{\lambda}_j \text{vol}(C).$$

Since volume is evidently translation invariant, the fact that C is centered at the origin is no restriction, and since $\hat{\lambda}_j - \check{\lambda}_j$ can be made arbitrarily small, we conclude that (8.26) is valid even when the λ_j are not necessarily rational. The proof of the general case now proceeds from (8.26) as before. \square

The next lemma is an immediate consequence of the Singular Value Decomposition Theorem that we prove later in thei section, after giving the geoemetric application.

Lemma 29. Let A be an $n \times n$ invertible matrix. Then there are two $n \times n$ ortogonal matrices U and V , and a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ with each $\lambda_j > 0$ such that $A = U\Lambda V^T$.

Theorem 90. Let A be an $n \times n$ invertible matrix. Let $E \subset \mathbb{R}^n$ have a well-defined and finite volume $\text{vol}(E)$. Then the image of E under A has a well-defined and finite volume $\text{vol}(A(E))$ and

$$\text{vol}(A(E)) = |\det(A)|\text{vol}(E) . \quad (8.27)$$

Proof. Evidently $A(E) = U(\Lambda(V^T(E)))$. Hence, using Lemma 26, then Lemma 28 and then Lemma 26 again,

$$\text{vol}(A(E)) = \text{vol}(\Lambda(V^T(E))) = \prod_{j=1}^n \lambda_j \text{vol}(V^T(E)) = \prod_{j=1}^n \lambda_j \text{vol}(E) .$$

On the other hand, by the product property of determinants and Theorem 89,

$$\det(A) = \det(U) \det(\Lambda) \det(V^T) = \pm \det(\Lambda) = \pm \prod_{j=1}^n \lambda_j .$$

□

Hence, the absolute value of $\det(A)$ is the *volume magnification factor* of the linear transformation induced by A . (This can of course be less than one, in which case the transformation is actually “shrinking” volume.)

Definition 94. Let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be an ordered basis for \mathbb{R}^n ; that is, a ordered set of n linearly independent vectors in \mathbb{R}^n . (The ordering is given by the indices.) The $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is positively oriented in case $\det([\mathbf{v}_1, \dots, \mathbf{v}_n]) > 0$, and is positively oriented in case $\det([\mathbf{v}_1, \dots, \mathbf{v}_n]) < 0$.

Because of the formula expressing the determinant in 3 dimensions in terms of the cross product, this definition is such that for $n = 3$, a basis is positively oriented if and only if it is right-handed, and is negatively oriented if and only if it is left-handed.

A question to which we are going to give short schrift is the question of which bounded subsets of \mathbb{R}^n have a well defined volume. It is not hard to see that if a set E contains an open set and has a piecewise smooth boundary composed of finitely many differentiable surfaces, then E does have a well defined volume, and it is this sort of set with which we shall work in what follows. The reason is the one can cover the boundary with a cubes have a total volume that is an arbitrarily small fraction of the total volume of some cube packing, and then combinigng the packing with the covering of the boundary, one gets a cube covering whose total volume is larger than that of the packing by an arbitrarily small percentage.

8.3.2 The singular value decomposition

Theorem 91 (The Singular Value Decomposition Theorem). Let A be an $m \times n$ matrix. Then there exist matrices U , V and Σ such that

$$A = U\Sigma V^T$$

and (1) $U \in O(m)$, (2) $V \in O(n)$, and (3) Σ is an $m \times n$ diagonal matrix such that $\Sigma_{j,j} = \sigma_j$ where $\sigma_j \geq \sigma_{j+1} \geq 0$ for all $j = 1, \dots, \min\{m, n\} - 1$, and $\Sigma_{j,j} = 0$ if $i \neq j$.

In any such factorization of A , the matrix Σ is always the same. In particular, the numbers $\{\sigma_1, \dots, \sigma_{\min\{m, n\}}\}$ are uniquely determined by A . We call these numbers the singular values of A .

For example, if $m = 3$ and $n = 4$, the matrix σ has the form

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \end{bmatrix}.$$

If $m = 4$ and $n = 3$, the matrix σ has the form

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \\ 0 & 0 & 0 \end{bmatrix}.$$

In both cases, $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$. We call such matrices *diagonal monotone*.

We have already seen one important application of the Singular Value Decomposition in identifying the geometric meaning of the determinant. It has many others, some of which are developed in the exercises.

Proof of Theorem 91. Let A be any $m \times n$ matrix. Form the $(m+n) \times (m+n)$ matrix

$$B := \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}.$$

More explicitly, the 0 entry in the upper left denotes the $m \times m$ zero matrix, and the 0 entry in the lower right denotes the $n \times n$ zero matrix.

We can write any vector $bx \in \mathbb{R}^{m+n}$ in the form

$$\mathbf{x} = (\mathbf{z}, \mathbf{w}) \quad \text{where } \mathbf{z} \in \mathbb{R}^m \text{ and } \mathbf{w} \in \mathbb{R}^n.$$

That is, the entries of \mathbf{z} constitute the first m entries of \mathbf{x} , while the entries of \mathbf{w} constitute the last n entries of \mathbf{x} . Then, by the definition of matrix multiplication,

$$B\mathbf{x} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} (\mathbf{z}, \mathbf{w}) = (A\mathbf{w}, A^T\mathbf{z}). \quad (8.28)$$

The matrix B is a symmetric matrix, and therefore, by the Spectral Theorem, there exists an orthonormal basis of \mathbb{R}^{m+n} consisting of eigenvectors of B . We now produce such a basis with a special structure reflecting the special structure of B .

Suppose that \mathbf{x} is an eigenvector of B so that $B\mathbf{x} = \lambda\mathbf{x}$. Suppose $\lambda \neq 0$. Writing $\mathbf{x} = (\mathbf{z}, \mathbf{w})$, we see from (8.28) that $B\mathbf{x} = \lambda\mathbf{x}$ is equivalent to

$$\mathbf{z} = \lambda A\mathbf{w} \quad \text{and} \quad \mathbf{w} = \lambda A^T\mathbf{z}. \quad (8.29)$$

It then follows that the vector $(\mathbf{z}, -\mathbf{w})$ is an eigenvector of B with eigenvalue $-\lambda$. Moreover, since eigenvectors corresponding to distinct eigenvalues of symmetric matrices are orthogonal, $(\mathbf{z}, -\mathbf{w}) \cdot (\mathbf{z}, \mathbf{w}) = 0$, which means that $\|\mathbf{z}\|^2 = \|\mathbf{w}\|^2$.

By the Spectral Theorem, there is an orthonormal basis of \mathbb{R}^{m+n} consisting of eigenvectors of B . Relabeling the vectors, we may suppose that exactly the first r of these have strictly positive

eigenvalues $\lambda_1 \geq \dots \geq \lambda_r$, and all of the others have zero or negative eigenvalues. Denote these first r eigenvectors as $\{\mathbf{x}_1, \dots, \mathbf{x}_r\}$. By what we have seen just above, for each j we have $\mathbf{x}_j = (\mathbf{z}_j, \mathbf{w}_j)$ where $\|\mathbf{z}_j\| = \|\mathbf{w}_j\|$. Without loss of generality, we may take $\|\mathbf{z}_j\| = \|\mathbf{w}_j\| = 1$. For $1 \leq j < k \leq r$, we have $0 = \mathbf{x}_j \cdot \mathbf{x}_k = \mathbf{z}_j \cdot \mathbf{z}_k + \mathbf{w}_j \cdot \mathbf{w}_k$. But since $(\mathbf{z}_j, \mathbf{w}_j)$ and $(\mathbf{z}_k, -\mathbf{w}_k)$ are necessarily orthogonal (one having a positive eigenvalue and the other a negative eigenvalue), we also have $0 = \mathbf{z}_j \cdot \mathbf{z}_k + \mathbf{w}_j \cdot \mathbf{w}_k$. Therefore, both $\mathbf{z}_j \cdot \mathbf{z}_k = 0$ and $\mathbf{w}_j \cdot \mathbf{w}_k = 0$. That is, $\{\mathbf{z}_1, \dots, \mathbf{z}_r\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_r\}$ are orthonormal. Extend these to orthonormal bases $\{\mathbf{z}_1, \dots, \mathbf{z}_m\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ of \mathbb{R}^m and \mathbb{R}^n respectively. Vectors of the form $(\mathbf{z}_j, 0)$ or $(0, \mathbf{w}_j)$ for $j > r$ are orthogonal to all of the eigenvectors of B with non-zero eigenvalues, and hence they are in the null space of B .

Now consider any $\mathbf{y} \in \mathbb{R}^n$. We expand

$$\mathbf{y} = \sum_{j=1}^n (\mathbf{y} \cdot \mathbf{w}_j) \mathbf{w}_j .$$

consequently

$$(0, \mathbf{y}) = \frac{1}{2} \sum_{j=1}^n (\mathbf{y} \cdot \mathbf{w}_j) ((\mathbf{z}_j, \mathbf{w}_j) + (-\mathbf{z}_j, \mathbf{w}_j)) .$$

Therefore,

$$\begin{aligned} B(0, \mathbf{y}) &= \frac{1}{2} \sum_{j=1}^n (\mathbf{y} \cdot \mathbf{w}_j) (B(\mathbf{z}_j, \mathbf{w}_j) + B(-\mathbf{z}_j, \mathbf{w}_j)) \\ &= \frac{1}{2} \sum_{j=1}^r (\mathbf{y} \cdot \mathbf{w}_j) \lambda_j ((\mathbf{z}_j, \mathbf{w}_j) - B(-\mathbf{z}_j, \mathbf{w}_j)) \\ &= \sum_{j=1}^r (\mathbf{y} \cdot \mathbf{w}_j) \lambda_j \mathbf{z}_j . \end{aligned}$$

However, $B(0, \mathbf{y}) = A\mathbf{y}$, and so we have $A\mathbf{y} = \sum_{j=1}^r (\mathbf{y} \cdot \mathbf{w}_j) \lambda_j \mathbf{z}_j$. To write this as a matrix factorization, let $U = [\mathbf{w}_1, \dots, \mathbf{w}_m]$, $V = [\mathbf{z}_1, \dots, \mathbf{z}_n]$, and let Σ be the $m \times n$ with $\Sigma_{j,j} = \lambda_j$ for $1 \leq j \leq r$, and with all other entries zero. Then $A\mathbf{y} = \sum_{j=1}^r (\mathbf{y} \cdot \mathbf{w}_j) \lambda_j \mathbf{z}_j$ is equivalent to $A = U\Sigma V^T$. \square

8.4 Exercises

8.1 Consider the following permutations

$$\sigma_1 = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 4 & 5 & 6 & 2 \end{matrix} \quad \sigma_2 = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 3 & 6 & 5 & 2 & 1 \end{matrix} \quad \sigma_3 = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 1 & 2 & 3 \end{matrix}$$

(a) Compute $D(\sigma_j)$ and $\chi(\sigma_j)$ for $j = 1, 2, 3$.

(b) For each $j = 1, 2, 3$, find a way to write σ_j as a product of pair permutations.

(c) Compute the value of $\chi(\sigma_1 \circ (\sigma_2 \circ \sigma_3)^{-1})$.

8.2 Consider the following permutations

$$\sigma_1 = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 1 & 3 & 5 \end{matrix} \quad \sigma_2 = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 1 & 6 & 4 & 2 & 3 \end{matrix} \quad \sigma_3 = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 1 & 5 & 2 & 6 & 3 \end{matrix}$$

- (a) Compute $D(\sigma_j)$ and $\chi(\sigma_j)$ for $j = 1, 2, 3$.
- (b) For each $j = 1, 2, 3$, find a way to write σ_j as a product of pair permutations.
- (c) Compute the value of $\chi(\sigma_1 \circ (\sigma_2 \circ \sigma_3)^{-1})$.

8.3 Let σ_1, σ_2 and σ_3 be the permutations defined in Exercise 1. Compute the distances $\varrho(\sigma_1, \sigma_2)$, $\varrho(\sigma_2, \sigma_3)$ and $\varrho(\sigma_3, \sigma_1)$. Also, find geodesics from σ_1 to σ_2 , from σ_2 to σ_3 , and from σ_3 to σ_1 .

8.4 Let σ_1, σ_2 and σ_3 be the permutations defined in Exercise 2. Compute the distances $\varrho(\sigma_1, \sigma_2)$, $\varrho(\sigma_2, \sigma_3)$ and $\varrho(\sigma_3, \sigma_1)$. Also, find geodesics from σ_1 to σ_2 , from σ_2 to σ_3 , and from σ_3 to σ_1 .

8.5 The *order reversing permutation* σ_* in \mathcal{S}_n is the permutation defined by

$$\sigma_* := \begin{matrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{matrix}.$$

In other words,

$$\sigma_*(k) = n - k + 1 \quad \text{for all } k = 1, \dots, n.$$

- (a) Show that $D(\sigma_*) = n(n-1)/2$, and that for all $\sigma \in \mathcal{S}_n$, $D(\sigma) < n(n-1)/2$ unless $\sigma = \sigma_*$.
- (b) Prove that

$$\max\{ \varrho(\sigma_1, \sigma_2) : \sigma_1, \sigma_2 \in \mathcal{S}_n \} = n(n-1)/2.$$

In other words, any two permutations in \mathcal{S}_n are connected by a path of at most $n(n-1)/2$ steps, and there exist pairs of permutations such that the shortest path connecting them has this many steps. This is often expressed by saying that *the diameter of \mathcal{S}_n is $n(n-1)/2$* .

8.6 Show that the set \mathcal{A}_n consisting of all even permutations in \mathcal{S}_n is a transformation group on $\{1, \dots, n\}$. \mathcal{A}_n is called the *alternating group of order n* . Show that there are exactly $n!/2$ permutations in \mathcal{A}_n , and show that the set of all odd permutations is not a transformation group.

8.7 For each $\sigma \in \mathcal{S}_n$, define the $n \times n$ matrix P_σ by

$$P_\sigma := [\mathbf{e}_{\sigma(1)}, \mathbf{e}_{\sigma(2)}, \dots, \mathbf{e}_{\sigma(n)}]; \tag{8.30}$$

that is, the j th column of P_σ is $\mathbf{e}_{\sigma(j)}$. The $n!$ matrices P_σ with $\sigma \in \mathcal{S}_n$ are called the *permutation matrices*. Prove that for all $\sigma \in \mathcal{S}_n$, $\det(P_\sigma) = \chi(\sigma)$.

Chapter 9

FLUX AND CIRCULATION, DIVERGENCE AND CURL

9.1 Flows and flux

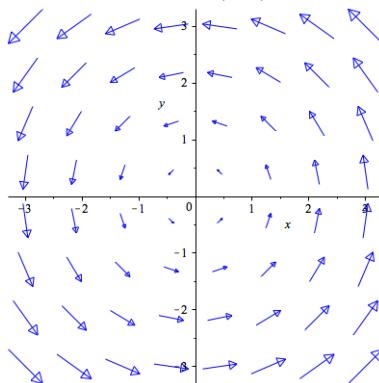
9.1.1 Vector fields and flows

Definition 95 (Vector field). *Let U be an open subset of \mathbb{R}^n . A function \mathbf{F} defined on U with values in \mathbb{R}^n is called a vector field on \mathbb{R}^n .*

For example, let $U = \mathbb{R}^2$, and define

$$\mathbf{F}(x, y) = (-y, x). \quad (9.1)$$

A vector field can be plotted by choosing certain points (x, y) , and then drawing an arrow with its tail at the point (x, y) , such that its magnitude and direction indicate the magnitude and direction of $\mathbf{F}(x, y)$. To avoid a cluttered plot, this should not be done for *too many* points. If it is done for a regular grid that is not too dense, the result usually portrays a clear picture of the vector field \mathbf{F} . Here is such a plot for the vector field \mathbf{F} defined in (9.1):



One can see that each of the arrows is tangent to the centered circle passing through its tail. Thinking of the arrows as velocity vectors, one can see that the graph is a portrait of circular motion. The length of the arrow at \mathbf{x} represents the magnitude $\|\mathbf{F}(\mathbf{x})\|$ of the vector $\mathbf{F}(\mathbf{x})$. In this case $\|\mathbf{F}(\mathbf{x})\| = \|\mathbf{x}\|$, which is represented by the length of each arrow being proportional to the distance between its tail and the origin.

If we think of the vectors $\mathbf{F}(\mathbf{x})$ as representing the velocity of a particle as it passes through \mathbf{x} , we may regard the plot of this vector field as a kind of portrait of circular motion, where the speed on each circle is proportional to the radius of the circle, so that the angular speed is the same for all of the circles. *That is, the motion being described is that given by rotating points in the plane at a constant angular speed.*

Definition 96 (Flow curves of a vector field). *Let U be an open subset of \mathbb{R}^n , and let \mathbf{F} be a continuous vector field defined on U . Let $\mathbf{x}(t)$ be a continuously differentiable curve in \mathbb{R}^n defined for $t \in (a, b)$. Suppose that $\mathbf{x}(t) \in U$ for all $t \in (a, b)$ and the*

$$\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}(t)) \quad \text{for all } t \in (a, b) . \quad (9.2)$$

Then $\mathbf{x}(t)$ is a flow curve of the vector field \mathbf{F} . (Flow curves are often also called integral curves.)

For the vector field \mathbf{F} given by (9.1) there is a unique flow curve $\mathbf{x}(t)$, defined for all t , passing through each \mathbf{x}_0 at $t = 0$. To see this, let $\mathbf{x}_0 = (x_0, y_0)$ be given, and suppose that $\mathbf{x}(t)$ is a continuously differentiable curve in \mathbb{R}^2 that satisfies

$$\mathbf{x}'(t) = (x'(t), y'(t)) = \mathbf{F}(\mathbf{x}(t)) = (-y(t), x(t)) \quad \text{and} \quad (x(0), y(0)) = (x_0, y_0) .$$

Then differentiating once more, $(x''(t), y''(t)) = (-x(t), -y(t))$ so that $x''(t) = -x(t)$ and $y''(t) = -y(t)$. We can satisfy $x''(t) = -x(t)$ by taking $x(t) = \alpha \cos(t) + \beta \sin(t)$ for any numbers α and β . Differentiating, and using the equation $x'(t) = -y(t)$, we then see that $y(t) = \alpha \sin(t) - \beta \cos(t)$. Now setting $t = 0$ and using the conditions $x(0) = x_0$ and $y(0) = y_0$, we see that $\alpha = x_0$ and $\beta = -y_0$. Thus, we have the solution

$$\mathbf{x}(t) = (x_0 \cos t - y_0 \sin t, x_0 \sin t + y_0 \cos t) .$$

We can write this in matrix notation as

$$\mathbf{x}(t) = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} (x_0, y_0) . \quad (9.3)$$

The matrix in (9.3) is the matrix representing counterclockwise rotation through the angle t in the plane. If one takes any point (x_0, y_0) , and then for each t produces $\mathbf{x}(t)$ by applying the rotation matrix $\begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}$ to it, the result is a flow curve of the vector field \mathbf{F} that is given in (9.1).

Are there any other flow curves for the vector field $\mathbf{F}(x, y) = (-y, x)$ besides the ones specified in (9.3)? No, and this is crucially important in what follows: Through each point \mathbf{x}_0 in the plane, there is exactly one flow curve $\mathbf{x}(t)$ satisfying $\mathbf{x}(0) = \mathbf{x}_0$.

9.1.2 Lipschitz vector fields on \mathbb{R}^n

Definition 97 (Lipschitz vector fields). *Let \mathbf{F} be a vector field defined on \mathbb{R}^n . Then \mathbf{F} is a Lipschitz vector field on \mathbb{R}^n in case there is a finite constant L such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$\|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|. \quad (9.4)$$

Example 130 (Linear vector fields). *A vector field \mathbf{F} on \mathbb{R}^n is linear in case there is an $n \times n$ matrix A such that $\mathbf{F}(\mathbf{x}) = A\mathbf{x}$. For example, the vector field $\mathbf{F}(x, y) = (-y, x)$ is linear with*

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Every linear vector field $\mathbf{F}(\mathbf{x}) = A\mathbf{x}$ is Lipschitz. This is because $\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) = Ay - Ax = A(\mathbf{y} - \mathbf{x})$, and hence

$$\|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x})\| = \|A(\mathbf{y} - \mathbf{x})\| \leq \|A\|_{\text{F}}\|\mathbf{y} - \mathbf{x}\|.$$

In particular, (9.4) is valid with $L = \|A\|_{\text{F}}$.

The inequality $\|Az\| \leq \|A\|_{\text{F}}\|z\|$ that we are using here, which is essentially the Cauchy-Schwarz inequality, is always true, and always gives a valid constant L , but it may not give the best one possible.

In fact for $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, $\|A\|_{\text{F}} = \sqrt{2}$, but it is easy to see that $\|Ay - Ax\| = \|A(\mathbf{y} - \mathbf{x})\| = \|\mathbf{y} - \mathbf{x}\|$ since A is an orthogonal matrix.

We now come to the first of two important theorems about Lipschitz vector fields:

Theorem 92 (Uniqueness of flow curves). *Let \mathbf{F} be a vector field that satisfies (9.4). Let $\mathbf{x}(t)$ and $\mathbf{y}(t)$ be flow curves for \mathbf{F} defined for $-a < t < a$ for some $a > 0$. Let \mathbf{x}_0 denote $\mathbf{x}(0)$, and let \mathbf{y}_0 denote $\mathbf{y}(0)$. Then for all $-a < t < a$,*

$$e^{-|t|L}\|\mathbf{y}_0 - \mathbf{x}_0\| \leq \|\mathbf{y}(t) - \mathbf{x}(t)\| \leq e^{|t|L}\|\mathbf{y}_0 - \mathbf{x}_0\|. \quad (9.5)$$

Both inequalities in (9.5) tell us something important about flow curves: From the inequality on the right, we see that if $\mathbf{x}_0 = \mathbf{y}_0$, then $\mathbf{x}(t) = \mathbf{y}(t)$ for all $-a < t < a$, so that the two flow curves are the same. In other words, when \mathbf{F} satisfies (9.4), there is at most one flow curve through each point $\mathbf{x}_0 \in \mathbb{R}^n$. But that is not all: fix any $\epsilon > 0$, and define $\delta(\epsilon) = e^{|t|L}\epsilon$. Then

$$\|\mathbf{y}(0) - \mathbf{x}(0)\| \leq \delta(\epsilon) \Rightarrow \|\mathbf{y}(t) - \mathbf{x}(t)\| \leq \epsilon. \quad (9.6)$$

In other words, if the initial points \mathbf{x}_0 and \mathbf{y}_0 are sufficiently close, then the flow curves through them will be close at time t . The inequality on the left in (9.5) tells us that the flow curves never cross: If $\mathbf{x}(0) \neq \mathbf{y}(0)$, then it is impossible to have $\mathbf{x}(t) = \mathbf{y}(t)$ for any t .

Proof of Theorem 92. Define $f(t) = \|\mathbf{y}(t) - \mathbf{x}(t)\|^2$. Differentiating, we compute

$$f'(t) = 2(\mathbf{y}(t) - \mathbf{x}(t)) \cdot (\mathbf{y}'(t) - \mathbf{x}'(t)) = 2(\mathbf{y}(t) - \mathbf{x}(t)) \cdot (\mathbf{F}(\mathbf{y}(t)) - \mathbf{F}(\mathbf{x}(t))).$$

Then by the Cauchy-Schwarz inequality and then the Lipschitz condition on \mathbf{F} ,

$$|(\mathbf{y}(t) - \mathbf{x}(t)) \cdot (\mathbf{F}(\mathbf{y}(t)) - \mathbf{F}(\mathbf{x}(t)))| \leq \|\mathbf{y}(t) - \mathbf{x}(t)\| \|\mathbf{F}(\mathbf{y}(t)) - \mathbf{F}(\mathbf{x}(t))\| \leq L\|\mathbf{y}(t) - \mathbf{x}(t)\|^2.$$

That is,

$$-2Lf(t) \leq f'(t) \leq 2Lf(t) . \quad (9.7)$$

The inequality on the right in (9.7) can be written as $f'(t) - 2Lf(t) \leq 0$. Multiplying through by e^{-2tL} , we have $(f'(t) - 2Lf(t))e^{-2tL} \leq 0$. Since the left hand side is the derivative of $f(t)e^{-2tL}$, we see that $f(t)e^{-2tL}$ is a non-increasing function of t . In particular, for all $t > 0$, $f(t)e^{-2tL} \leq f(0)$ and for all $t < 0$, $f(t)e^{-2tL} \geq f(0)$.

The inequality on the left in (9.7) can be written as $f'(t) + 2Lf(t) \geq 0$. Multiplying through by e^{2tL} , we have $(f'(t) + 2Lf(t))e^{2tL} \geq 0$. Since the left hand side is the derivative of $f(t)e^{2tL}$, we see that $f(t)e^{2tL}$ is a non-decreasing function of t . In particular, for all $t > 0$, $f(t)e^{2tL} \geq f(0)$ and for all $t < 0$, $f(t)e^{2tL} \geq f(0)$.

Altogether, we have proved $e^{-2|t|L}f(0) \leq f(t) \leq e^{-2|t|L}f(0)$. Taking square roots, we obtain (9.6) \square

Without the Lipschitz condition, uniqueness may fail to be true; see Example 131 below. In fact, the Lipschitz condition is quite close to being necessary, as well as sufficient.

The second important theorem on Lipschitz vector fields says that flow curves exist and are defined for all t :

Theorem 93 (Picard's Theorem). *Let \mathbf{F} be a Lipschitz vector field on \mathbb{R}^n , and let $\mathbf{x}_0 \in \mathbb{R}^n$. There there is a continuously differentiable curve $\mathbf{x}(t)$ defined for all t such that $\mathbf{x}(0) = \mathbf{x}_0$ and $\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}(t))$ for all t .*

We shall not prove this Theorem 93 here; we postpone until we systematically study differential equations. For our present purposes, we only need to know that flow curves exist, and are unique. In some cases, such as the case $\mathbf{F}(x, y) = (-y, x)$ we can explicitly solve for the flow curves.

Picard's Theorem is intuitively quite plausible since the equation $\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}(t))$ is essentially a set of instructions for constructing a curve $\mathbf{x}(t)$ passing through \mathbf{x}_0 at time $t = 0$. To see this, pick a small time step $h > 0$. Since any flow curve satisfies $\mathbf{x}'(0) = \mathbf{F}(\mathbf{x}_0)$, it is the case that $\mathbf{x}(h) \approx \mathbf{x}_0 + \mathbf{F}(\mathbf{x}_0)h$. Define $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{F}(\mathbf{x}_0)h$. Now we inductively define a sequence of points $\{\mathbf{x}_n\}$ as follows: Given \mathbf{x}_n , we use the vector field to specify the next step:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{F}(\mathbf{x}_n)h .$$

For $n \geq 0$ define $t_n = nh$. Now define a continuous curve by “connecting the dots”: For $t_n \leq t \leq t_{n+1}$ define

$$\mathbf{x}^{(h)}(t) = \frac{t_{n+1} - t}{h}\mathbf{x}_n + \frac{t - t_n}{h}\mathbf{x}_{n+1} .$$

It can be shown that $\mathbf{x}(t) = \lim_{h \rightarrow 0} \mathbf{x}^{(h)}(t)$ exists and is the flow curve we seek. (The same construction can be adapted for $t < 0$ as well). This is the *Euler scheme* for solving such equations. Using the Euler scheme with a small time step h , or a refinement of the Euler scheme, it is easy to numerically plot flow curves.

The following example in which \mathbf{F} is not Lipschitz will be useful for understanding the role of the Lipschitz condition.

Example 131. Consider the vector field $\mathbf{F}(x, y) = (x^2, y^2)$. Then $\|\mathbf{F}(x, 0) - \mathbf{F}(0, 0)\| = x^2 = |x|\|(x, 0) - (0, 0)\|$, and since $|x|$ can be arbitrarily large we see that $\|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x})\|$ can be an arbitrarily large multiple of $\|\mathbf{y} - \mathbf{x}\|$. Hence, \mathbf{F} is not a Lipschitz vector field.

It is easy to solve the equation $\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}(t))$. This is the uncoupled system

$$\begin{aligned} x'(t) &= x^2(t) \\ y'(t) &= y^2(t). \end{aligned}$$

It suffices to consider the equation for $x(t)$. First, if $x(0) = 0$, the constant function $x(t) = 0$ for all t solves $x'(t) = x^2(t)$ with $x(0) = 0$. Hence we have an integral curve in this case. Now suppose $x(0) \neq 0$. Then by continuity, for t sufficiently small, $x(t) \neq 0$. and we can rewrite $x'(t) = x^2(t)$ as

$$\frac{x'(t)}{x^2(t)} = 1$$

The advantage of this is that the left hand side is the derivative of $-1/x(t)$. Hence $(1/x(t))' = -1$. Suppose $x(t)$ is any solution such that $x(s) \neq 0$ for any s between 0 and t . Integrating $(1/x(s))' = -1$ from $s = 0$ to $s = t$ we find

$$\frac{1}{x(0)} - \frac{1}{x(t)} = t.$$

Solving for $x(t)$ we find

$$x(t) = \frac{x(0)}{1 - x(0)t}.$$

Because $x(0) \neq 0$, this is never equal to zero, but the denominator goes to zero as t approaches $1/x(0)$. Hence the solution “blows up” in a finite time. By Picard’s Theorem, this never happens for Lipschitz vector fields.

We now come to the *flow transformations* associated to a Lipschitz vector field \mathbf{F} on \mathbb{R}^n . For $t \in \mathbb{R}$ and \mathbf{x}_0 in \mathbb{R}^n , let $\mathbf{x}(t)$ be the unique flow curve of \mathbf{F} such that $\mathbf{x}(0) = \mathbf{x}$. Then for each t , define a function Φ^t on \mathbb{R}^n with values in \mathbb{R}^n by

$$\Phi^t(\mathbf{x}) = \mathbf{x}(t).$$

The time t can be either positive or negative. If $t > 0$, $\Phi^t(\mathbf{x})$ is the point that \mathbf{x} gets carried to at time t by the flow. If $t < 0$, $\Phi^t(\mathbf{x})$ is the point that gets carried to \mathbf{x} by the flow at time t . The time t is written as an exponent to suggest an analogy with the exponential function; this will be explained below.

Example 132 (Flow transformations generated by $\mathbf{F}(x, y) = (-y, x)$). We have already seen that for $\mathbf{F}(x, y) = (-y, x)$, the unique solution of $\mathbf{x}'(t) = (\mathbf{x}(t))$ with $\mathbf{x}(0) = (x_0, y_0)$ is

$$\mathbf{x}(t) = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} (x_0, y_0).$$

It follows that for each $t \in \mathbb{R}$, Φ^t is the linear transformation whose matrix is $\begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}$.

We now consider two more examples in which we can give an explicit formula for $\Phi^t(\mathbf{x})$.

Example 133 (Flow transformations generated by a constant vector field). *Let $\mathbf{A} \in \mathbb{R}^n$, and then let $\mathbf{F}(\mathbf{x}) = \mathbf{A}$ for all \mathbf{x} so that \mathbf{F} is a constant vector field. The flow curve $\mathbf{x}(t)$ passing through \mathbf{x}_0 at $t = 0$ satisfies $\mathbf{x}'(t) = \mathbf{A}$ and $\mathbf{x}(0) = \mathbf{x}_0$. Therefore, $\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{A}$. It follows that the flow transformation generated by \mathbf{F} is given by $\Phi^t(\mathbf{x}) = \mathbf{x} + t\mathbf{A}$.*

Example 134 (Flow transformations generated by $\mathbf{F}(\mathbf{x}) = \mathbf{x}$). *Let $\mathbf{F}(\mathbf{x}) = \mathbf{x}$ for all \mathbf{x} . The flow curve $\mathbf{x}(t)$ passing through \mathbf{x}_0 at $t = 0$ satisfies $\mathbf{x}'(t) = \mathbf{x}(t)$ and $\mathbf{x}(0) = \mathbf{x}_0$. Therefore, $\mathbf{x}(t) = e^t \mathbf{x}_0$. It follows that the flow transformation generated by \mathbf{F} is given by $\Phi^t(\mathbf{x}) = e^t \mathbf{x}$.*

9.1.3 Flux across an oriented curve in \mathbb{R}^2 .

Definition 98 (Smooth simple planar curve). *A smooth planar curve \mathcal{C} is a subset of \mathbb{R}^2 such that for each $\mathbf{x}_0 \in \mathcal{C}$, there is an interval (a, b) and a continuously differentiable parameterized curve $\mathbf{x}(t)$ defined for $t \in (a, b)$ and a $t_0 \in (a, b)$ such that:*

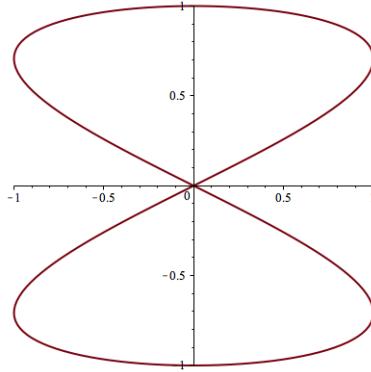
- (i) $\mathbf{x}(t_0) = \mathbf{x}_0$, so that the parameterized curve passes through \mathbf{x}_0 .
- (ii) $\mathbf{x}(t) \in \mathcal{C}$ for all $t \in (a, b)$, so that every point on the parameterized curve lies in \mathcal{C} .
- (iii) $\mathbf{x}(s) \neq \mathbf{x}(t)$ for any $a < s < t < b$, so that the parameterized curve is one to one onto the portion of \mathcal{C} that it covers.
- (iv) There is an $r > 0$ such that every $\mathbf{x} \in \mathcal{C}$ with $\|\mathbf{x} - \mathbf{x}_0\| < r$ is of the form $\mathbf{x}(t)$ for some $t \in (a, b)$. (This requirement rules out self-intersections, among other things, see the discussion below.)

An orientation of \mathcal{C} is a continuous specification of one preferred unit tangent $\mathbf{T}(\mathbf{x})$ direction at each point \mathbf{x} of \mathcal{C} . In this case, we refer to $\mathbf{T}(\mathbf{x})$ as the unit tangent vector of the oriented curve \mathcal{C} at \mathbf{x} , and we define $\mathbf{N}(\mathbf{x}) = -\mathbf{T}(\mathbf{x})^\perp$ which we refer to as the unit normal vector of the oriented curve \mathcal{C} at \mathbf{x}_0 . We say that a parameterization $\mathbf{x}(t)$ of \mathcal{C} (or part of \mathcal{C}) is consistent with the orientation of \mathcal{C} in case

$$\mathbf{T}(\mathbf{x}(t)) = \frac{1}{\|\mathbf{x}'(t)\|} \mathbf{x}'(t)$$

for all parameter values t . (Recall that we have defined $(x, y)^\perp = (-y, x)$, so that $(x, y)^\perp$ is the counterclockwise rotation of (x, y) through $\pi/2$. Then $-(x, y)^\perp$ is given by the corresponding clockwise rotation.) With the convention that $\mathbf{N} = -\mathbf{T}^\perp$ so that $\mathbf{T} = \mathbf{N}^\perp$, the orientation can be equivalently specified by giving the preferred unit normal at each point, which is required to be continuous so that $\mathbf{T} = -\mathbf{N}^\perp$ is continuous.

The fact that there are exactly two unit tangent vectors defined at each point on \mathcal{C} is true because near each point we have a continuously differentiable parameterization, and because the definition rules out self-intersection of curves. For example, the “figure eight” curve pictured below is not a smooth simple curve.



This curve has a nice parameterization: Let $\mathbf{x}(t) = (\sin(2t), \sin(t))$. If we restrict t to the interval $(0, \pi)$, we get a one to one parameterization of the upper loop. If we restrict t to the interval $(\pi, 2\pi)$, we get a one to one parameterization of the lower loop. The only point not yet covered is $(0,0) = \mathbf{x}(0) = \mathbf{x}(\pi)$. We can cover this in a one to one manner by restricting t to the interval $(\pi/2, 3\pi/2)$ or to $(3\pi/2, 5\pi/2)$. Then through each point $\mathbf{x}_0 \in \mathcal{C}$, we have a parameterization of part of \mathcal{C} that satisfies (i), (ii) and (iii). However, (iv) is not satisfied since two “branches” of \mathcal{C} cross at $\mathbf{x}_0 = (0,0)$, and no one to one parameterization can cover all of the points of \mathcal{C} in $B_r((0,0))$ for any $r > 0$.

Moreover, note that while $\mathbf{x}(0) = \mathbf{x}(\pi) = (0,0)$, $\mathbf{x}'(0) = (2,1)$ and $\mathbf{x}'(\pi) = (2,-1)$: The curve passes through $(0,0)$ in two linearly independent directions, and there is no uniquely defined tangent line at $(0,0) \in \mathcal{C}$.

Condition (iv) is also what guarantees that the curve is really one dimensional. Without this condition, \mathbb{R}^2 itself would satisfy the definition: Through each point (x_0, y_0) there passes the $\mathbf{x}(t) = (x_0 + t, y_0)$, say, but it does not cover all $B_r((x_0, y_0))$ for any $r > 0$.

Example 135. *The unit circle is smooth simple planar curve: Consider $\mathbf{x}_1(t) = (\cos t, \sin t)$ for $0 < t < 2\pi$, and $\mathbf{x}_2(t) = (\sin t, \cos t)$. Then every point in \mathcal{C} except $(1,0)$ is of the form $\mathbf{x}(t)$ for exactly one $t \in (0, 2\pi)$, and every point in \mathcal{C} except $(0,1)$ is of the form $\mathbf{x}(t)$ for exactly one $t \in (0, 2\pi)$. Together, every point is covered by the two parameterizations. and clearly each one covers all points on \mathcal{C} that are sufficiently close to any point either covers.*

The unit circle is a closed curve; it divides the plane into an “inside” and an “outside” part. In this case we may orient the curve so the the unit normal vector \mathbf{N} points outward or inward. The conventional choice for closed curves is to choose the outward normal. Then since $\mathbf{N} = -\mathbf{T}^\perp$, $\mathbf{T} = \mathbf{N}^\perp$, and so this orientation corresponds to counterclockwise motion.

Notice that the motion along $\mathbf{x}_1(t)$ is counterclockwise, so this parameterization is consistent with the outward-normal orientation. However, the motion along $\mathbf{x}_2(t)$ is clockwise, so this parameterization is not consistent with the outward-normal orientation. We can fix this by reversing the parameterization: Define $\mathbf{x}_3(t) = \mathbf{x}_2(2\pi - t)$ for $0 < t < 2\pi$. This parameterization is consistent with the orientation. Such a reversal may always be employed to bring parametrization in line with an specified orientation.

Now we come to the concept of the *flux across a a simple oriented curve \mathcal{C} generated by a vector*

field \mathbf{F} . (We will always suppose that \mathcal{C} is smooth and \mathbf{F} is continuously differentiable and Lipschitz, at least on an open set U containing \mathcal{C} , even when we do not explicitly mention this.)

Let us suppose for the moment that \mathcal{C} is covered by a single parameterization $\mathbf{x}(t)$, $t \in (a, b)$, and that the parameterization is consistent with the orientation of \mathcal{C} .

Now consider a parameterized “patch” of \mathbb{R}^2 given by “pushing \mathcal{C} along” the flow generated by \mathbf{F} for a short time interval. That is, define

$$\mathbf{X}(u, v) = \Phi^u(\mathbf{x}(v))$$

where Φ is the flow transformation generated by \mathbf{F} .

Example 136 (Pushing an oriented curve along a flow). Let \mathcal{C} be the vertical line segment parameterized by $\mathbf{x}(t) = (1, 4t - 2)$ for $t \in (0, 1)$. Give \mathcal{C} the orientation that is consistent with this parameterization. Since $\mathbf{T} = (0, 1)$ at each point of \mathcal{C} , $\mathbf{N} = (1, 0)$ at each point of \mathcal{C} .

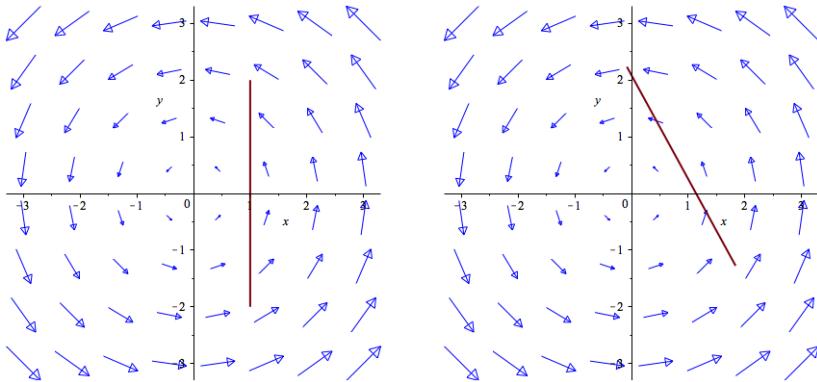
Let $\mathbf{F}(x, y) = (-y, x)$ as in (9.1). In Example 132, we have computed the flow transformation generated by \mathbf{F} and found the

$$\Phi^u(\mathbf{x}) = \begin{bmatrix} \cos u & -\sin u \\ \sin u & \cos u \end{bmatrix} \mathbf{x}.$$

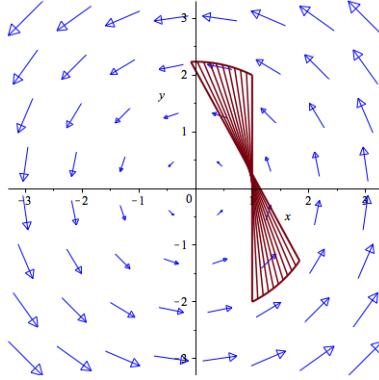
Replacing \mathbf{x} by $\mathbf{x}(v)$, we have

$$\mathbf{X}(u, v) = \begin{bmatrix} \cos u & -\sin u \\ \sin u & \cos u \end{bmatrix} (1, 4v - 2) = (\cos u - (4v - 2)\sin u, \sin u + (4v - 2)\cos u).$$

The next two plots show the initial curve \mathcal{C} against the background of the vector field \mathbf{F} that will “push” it along, and the rotated segment that results from running the flow generated by \mathbf{F} for a time interval of $u = 1/2$, so that the transformed curve is parameterized by $\mathbf{x}(v) = \mathbf{x}(1/2, v)$, $v \in (0, 1)$.



As the flow transforms the initial curve into the final curve, it “sweeps out” a patch of the plane. For $t > 0$, let \mathcal{D}_t denote the patch swept out for u in $(0, t)$. Here is a plot showing the boundary of the patch $\mathcal{D}_{1/2}$ against background of the vector field \mathbf{F} , and also the images of the initial curve \mathcal{C} for $u = n/20$, $n = 1, 2, \dots, 10$:



The region $\mathcal{D}_{1/2}$ consists of three parts: Define

$$\begin{aligned}\mathcal{D}_{1/2}^- &= \{(x(u, v), y(u, v)) : u \in (0, 1/2) \text{ and } v \in (1/2, 1)\} \\ \mathcal{D}_{1/2}^+ &= \{(x(u, v), y(u, v)) : u \in (0, 1/2) \text{ and } v \in (0, 1/2)\} \\ \mathcal{D}_{1/2}^0 &= \{(x(u, v), y(u, v)) : u \in (0, 1/2) \text{ and } v = 1/2\}\end{aligned}$$

Then $\mathcal{D}_{1/2} = \mathcal{D}_{1/2}^+ \cup \mathcal{D}_{1/2}^0 \cup \mathcal{D}_{1/2}^-$. The component $\mathcal{D}_{1/2}^0$ is an arc of the unit circle. If

The two regions $\mathcal{D}_{1/2}^+$ and $\mathcal{D}_{1/2}^-$ are not disjoint: The intersection is the open set bounded by the the unit circle (the relevant part being the arc $\mathcal{D}_{1/2}^0$), and the tangent lines to the unit circle at $(1, 0)$ and $(\cos(1/2), \sin(1/2))$.

The meaning of $\mathcal{D}_{1/2}^-$ is that it is the region “swept out” by points on the upper half of \mathcal{C} , which are initially moving toward the “negative side” of \mathcal{C} , since $\mathbf{N} = (1, 0)$ all along \mathcal{C} , these points are initially moving toward the left. Likewise, the meaning of $\mathcal{D}_{1/2}^+$ is that it is the region “swept out” by points on the lower half of \mathcal{C} , which are initially moving toward the “positive side” of \mathcal{C} . The word *initially* is crucial here. Fix a small $\epsilon > 0$, and consider the flow curve parameterized by $\mathbf{x}(u, 1/2 - \epsilon/4)$ through the point $(1, -\epsilon)$ on the lower half of \mathcal{C} . From the computations above,

$$\mathbf{x}(u, 1/2 - \epsilon/4) = (\cos u + \epsilon \sin u, \sin u - \epsilon \cos u).$$

Differentiating in u , the initial velocity is $(\epsilon, 1)$, and the dot product with \mathbf{N} is strictly positive. This confirms by calculation what is clear from the plots: Initially, the motion is toward the positive side of \mathcal{C} . However, for $u = 1/2$ the position is $\mathbf{x}(1/2, 1/2 - \epsilon/4) = (\cos(1/2) + \epsilon \sin(1/2), \sin(1/2) - \epsilon \cos(1/2))$, and since $\cos(1/2) < 1$, for sufficiently small ϵ , this point lies to the left of \mathcal{C} .

Flux is a measure of the rate at which the flow generated by a vector field \mathbf{F} sweeps area across an oriented curve \mathcal{C} , keeping track of the net amount of area that is swept to the positive side.

Before introducing general definitions, we do an actual computation of the quantities involved.

Example 137. Continuing with the notation introduced in Example 136, we now compute the areas of $\mathcal{D}_{1/2}^+$ and $\mathcal{D}_{1/2}^-$.

This is very easy to do using the change of variables formula since $\mathcal{D}_{1/2}^+$ is the image of the rectangle $[0, 1/2] \times [0, 1/2]$ in the u, v plane under the transformation sending (u, v) to $\mathbf{X}(u, v)$. From our formula for $\Phi^u(\mathbf{x})$ and the definition of $\mathbf{X}(u, v)$, we have the formula

$$\mathbf{X}(u, v) = (\cos u - (4v - 2) \sin u, \sin u + (4v - 2) \cos u),$$

we compute

$$[\mathcal{D}_{\mathbf{x}}(u, v)] = \begin{bmatrix} -\sin u - (4v - 2) \cos u & \cos u - (4v - 2) \sin u \\ -4 \sin u & 4 \cos u \end{bmatrix},$$

and then

$$\det([\mathcal{D}_{\mathbf{x}}(u, v)]) = (2 - 4v)(\cos^2 u - \sin^2 u) = (2 - 4v) \cos(2u)$$

Note that for $u \in (0, 1/2)$, $\cos(u) > 0$. Hence the determinant is positive if and only if $v < 1/2$, and negative if and only if $v > 1/2$. Hence the determinant is positive on $\mathcal{D}_{1/2}^+$, and negative on $\mathcal{D}_{1/2}^-$.

On both regions, the Jacobian $\frac{\partial(x, y)}{\partial(u, v)}$ is given by $\frac{\partial(x, y)}{\partial(u, v)} = 4|1/2 - v| \cos(2u)$ and we find

$$\text{area}(\mathcal{D}_{1/2}^-) = \text{area}(\mathcal{D}_{1/2}^+) = 4 \left(\int_0^{1/2} (1/2 - v) dv \right) \left(\int_0^{1/2} \cos(2u) du \right) = \frac{1}{4} \sin 1.$$

In this case, the net area, defined to be the difference of the positive area and the negative area, is exactly zero. Note also that since $\mathcal{D}_{1/2}^-$ and $\mathcal{D}_{1/2}^+$ are not disjoint, the total area of $\mathcal{D}_{1/2}^- \cup \mathcal{D}_{1/2}^+$ is not the sum of the two area we just computed, but this minus the area of the overlap. By elementary geometric reasoning, the area of the overlap is $\tan(1/4) - 1/4$, and hence the total area, as opposed to the net area, is $\sin 1 + 1/4 - \tan(1/4)$. The net area is somewhat simpler to work with; we need no concern ourselves with overlap since it is canceled out anyhow.

Now that we have familiarized ourselves with the concept of the net area pushed across an oriented curve \mathcal{C} by the flow generated by a vector field \mathbf{F} , we turn to flux, which is defined to be the rate at which this is taking place.

Suppose that \mathcal{C} is covered by a single parameterization $\mathbf{z}(t)$, $t \in (a, b)$, and such that $\mathbf{F}(\mathbf{z}(t)) \cdot \mathbf{N}(\mathbf{z}(t))$ is either strictly positive or else strictly negative for all $t \in (a, b)$. (We are saving the variable \mathbf{x} for a change of variables in the plane that is coming up, and so we denote position along the curve \mathcal{C} by \mathbf{z} .)

In general, we can break \mathcal{C} into disjoint segments on which the conditions are satisfied, and we can then consider the segments one at a time so that there is no genuine loss of generality in making this assumption.

Definition 99 (Flux across an oriented curve \mathcal{C}). *Let \mathcal{C} be an oriented smooth simple curve with unit normal \mathbf{N} . Let \mathbf{F} be a continuously differentiable Lipschitz vector field on \mathbb{R}^2 . Suppose that either (i) $\mathbf{N} \cdot \mathbf{F} > 0$ everywhere along \mathcal{C} , or (ii) $\mathbf{N} \cdot \mathbf{F} < 0$ everywhere along \mathcal{C} . For $v \in \mathbb{R}$, let Φ^v be the flow transformation generated by \mathbf{F} at time v . For $t > 0$ define the set \mathcal{D}_t in the plane by*

$$\mathcal{D}_t = \{\Phi^v(\mathbf{z}) : v \in (0, t), \mathbf{z} \in \mathcal{C}\}. \quad (9.8)$$

This is the set of points that is “swept out” in the time interval $(0, t)$ as the flow pushes points across \mathcal{C} . We define the flux across \mathcal{C} generated by \mathbf{F} , $\text{flux}(\mathcal{C}, \mathbf{F})$ to be the quantity

$$\text{flux}(\mathcal{C}, \mathbf{F}) = \pm \frac{d}{dt} \text{area}(\mathcal{D}_t) \Big|_{t=0}, \quad (9.9)$$

where the sign in front of the derivative is the sign of $\mathbf{F} \cdot \mathbf{N}$ along \mathcal{C} . If \mathcal{C} is a simple oriented curve that is a disjoint union of components \mathcal{C}_j , $j = 1, \dots, N$, satisfying the conditions above, we define

$$\text{flux}(\mathcal{C}, \mathbf{F}) = \sum_{j=1}^N \text{flux}(\mathcal{C}_j, \mathbf{F}). \quad (9.10)$$

Fortunately, it is not necessary to compute $\text{area}(\mathcal{D}_t)$ to compute $\text{flux}(\mathcal{C}, \mathbf{F})$:

Theorem 94. Let \mathcal{C} be a smooth simple oriented curve, and let \mathbf{F} be a continuously differentiable vector field on \mathbb{R}^2 . Then the flux across \mathcal{C} generated by \mathbf{F} is given by

$$\text{flux}(\mathcal{C}, \mathbf{F}) = \int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds, \quad (9.11)$$

which is the integral of the function $\mathbf{F} \cdot \mathbf{N}$ along the curve with respect to arc length.

Proof. We the area of \mathcal{D}_t as an integral over a simple rectangle by making a change of variables.

Introduce the coordinate transformation $\mathbf{x}(u, v) = \Phi^v(\mathbf{z}(u))$. Then \mathcal{D}_t is the image of $(0, t) \times (a, b)$ under this transformation. Therefore, by the change of variables formula,

$$\text{area}(\mathcal{D}_t) = \int_{(0,t) \times (a,b)} |\det[D\mathbf{x}(u, v)]| d^2\mathbf{u}. \quad (9.12)$$

The Jacobian matrix of this transformation is $[D\mathbf{x}(u, v)] = \begin{bmatrix} \frac{\partial \mathbf{x}(u, v)}{\partial u} & \frac{\partial \mathbf{x}(u, v)}{\partial v} \end{bmatrix}$. For fixed u , $\mathbf{x}(u, v)$ traces out a flow curve of \mathbf{F} as v varies, and so $\frac{\partial \mathbf{x}(u, v)}{\partial v} = \mathbf{F}(\mathbf{x}(u, v))$. This tells us the second column of $[D\mathbf{x}(u, v)]$.

Next, since Φ^0 is the identity transformation $\mathbf{x}(u, 0) = \mathbf{z}(u)$ and so $\frac{\partial}{\partial u} \mathbf{x}(u, 0) = \mathbf{z}'(u)$. Therefore, we can evaluate the Jacobian determinant at $v = 0$:

$$\det[D\mathbf{x}(u, 0)] = \det([\mathbf{z}'(u), \mathbf{F}(\mathbf{z}(u))]) = -(\mathbf{z}'(u))^{\perp} \cdot \mathbf{F}(\mathbf{z}(u)).$$

Since the orientation is consistent with the parameterization, $-(\mathbf{z}'(u))^{\perp} = \|\mathbf{z}'(u)\| \mathbf{N}(\mathbf{z}(u))$, and so

$$\det[D\mathbf{x}(u, 0)] = \|\mathbf{z}'(u)\| \mathbf{F}(\mathbf{z}(u)) \cdot \mathbf{N}(\mathbf{z}(u)).$$

This has the same sign as $\mathbf{F} \cdot \mathbf{N}$ along \mathcal{C} . Moreover, by continuity if $\mathbf{F} \cdot \mathbf{N}$ is strictly positive, then also $\det[D\mathbf{x}(u, v)] > 0$ for all sufficiently small v , and if $\mathbf{F} \cdot \mathbf{N}$ is strictly negative, then also $\det[D\mathbf{x}(u, v)] < 0$ for all sufficiently small v

Therefore, (9.13) becomes

$$\pm \text{area}(\mathcal{D}_t) = \pm \int_{(0,t) \times (a,b)} \left(\frac{\partial \mathbf{x}(u, v)}{\partial u} \right)^{\perp} \cdot \mathbf{F}(\mathbf{x}(u)) d^2\mathbf{u} = \pm \int_0^t \left(\int_a^b \left(\frac{\partial \mathbf{x}(u, v)}{\partial u} \right)^{\perp} \cdot \mathbf{F}(\mathbf{x}(u)) du \right) dv \quad (9.13)$$

where the sign on the left is the sign of $\mathbf{F} \cdot \mathbf{N}$ along \mathcal{C} .

Now by the Fundamental Theorem of Calculus,

$$\begin{aligned} \pm \frac{d}{dt} \text{area}(\mathcal{D}_t) \Big|_{t=0} &= - \int_a^b \left(\frac{\partial \mathbf{x}(u, 0)}{\partial u} \right)^{\perp} \cdot \mathbf{F}(\mathbf{x}(u)) du \\ &= - \int_a^b (\mathbf{z}'(u))^{\perp} \cdot \mathbf{F}(\mathbf{z}(u)) du \\ &= \int_a^b \mathbf{F}(\mathbf{z}(u)) \cdot \mathbf{N}(\mathbf{z}(u)) \|\mathbf{z}'(u)\| du = \int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds \end{aligned}$$

where in the last line we have used the formula $ds = \|\mathbf{z}'(u)\|du$ for the arc length element along $\mathbf{z}(u)$. \square

It is often easy to compute flux integrals once one has parameterized the oriented curve \mathcal{C} . While arc length integrals tends to be complicated even for simple curves, Nds is often simpler than ds itself. To see why, let $\mathbf{x}(t) = (x(t), y(t))$ be a parameterization of \mathcal{C} (or a segment of a large curve that we are considering). Then the unit tangent and unit normal vectors are $\mathbf{T}(t) = \frac{1}{\|\mathbf{x}'(t)\|}\mathbf{x}'(t)$ and $\mathbf{N}(t) = -\frac{1}{\|\mathbf{x}'(t)\|}\mathbf{x}'(t)^\perp$ if the parameterization is consistent with the orientation, while $\mathbf{T}(t) = -\frac{1}{\|\mathbf{x}'(t)\|}\mathbf{x}'(t)$ and $\mathbf{N}(t) = \frac{1}{\|\mathbf{x}'(t)\|}\mathbf{x}'(t)^\perp$ otherwise. The arc length element is $ds = \|\mathbf{x}'(t)\|dt$, and so

$$Nds = \pm \mathbf{x}'(t)dt , \quad (9.14)$$

where the minus sign is correct if the parameterization is consistent with the orientation, and the plus sign otherwise.

That is, whenever \mathcal{C} is parameterized by $\mathbf{x}(t)$ with $t \in (a, b)$, and the parameterization is consistent with the orientation,

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = - \int_a^b \mathbf{F}(\mathbf{x}(t)) \cdot (\mathbf{x}'(t))^\perp dt . \quad (9.15)$$

Example 138. Let $\mathbf{F}(x, y) = (-y, x)$. Let \mathcal{C} be the line segment running from $(1, -2)$ to $(1, 2)$, oriented so that the positive side is to the left. Then $\mathbf{x}(t) = (1, 2t - 4)$, $t \in (0, 1)$ is a parameterization of \mathcal{C} that is consistent with the orientation.

We compute $\mathbf{F}(\mathbf{x}(t)) = (4 - 2t, 1)$ and $(\mathbf{x}'(t))^\perp = (-2, 0)$. Hence $\mathbf{F}(\mathbf{x}(t)) \cdot (\mathbf{x}'(t))^\perp = 4t - 8$. Finally, by (9.15),

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = \int_0^1 (8 - 4t) dt = 0 ,$$

which is consistent with what we found earlier.

We now consider an example with a more complicated vector field \mathbf{F} :

Example 139. Let $\mathbf{F}(x, y) = (xy, x^2 - y^2)$. Let \mathcal{C} be the unit circle centered at $(1, 1)$, oriented so that the unit normal is outward. The standard parameterization of \mathcal{C} is $\mathbf{x}(t) = (1 + \cos t, 1 + \sin t)$, $t \in (0, 2\pi)$. This parameterization is consistent with the orientation.

We compute

$$\mathbf{F}(\mathbf{x}(t)) = (1 + \cos t + \sin t + \cos t \sin t, \cos^2 t - \sin^2 t + 2(\cos t - \sin t))$$

and $(\mathbf{x}'(t))^\perp = (\cos t, \sin t)$. Hence

$$\mathbf{F}(\mathbf{x}(t)) \cdot (\mathbf{x}'(t))^\perp = \cos t(1 + \cos t + \sin t + \cos t \sin t) + \sin t(\cos^2 t - \sin^2 t + 2(\cos t - \sin t)) .$$

Finally, by (9.15), $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds$ is the integral of this over $(0, 2\pi)$. There are many terms, but most integrate to zero by symmetry. What remains is

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = \int_0^{2\pi} [\cos^2 t - 2\sin^2 t] dt = -\pi .$$

In this example, more area is being swept into the disc than is being swept out.

9.1.4 Flux across oriented surfaces in \mathbb{R}^3

The definition of a smooth, simple surface in \mathbb{R}^2 is very much like the definition of a simple closed curve in \mathbb{R}^2 :

Definition 100 (Smooth simple surface in \mathbb{R}^3). A smooth surface in \mathbb{R}^3 \mathcal{S} is a subset of \mathbb{R}^3 such that for each $\mathbf{x}_0 \in \mathcal{S}$, there is an open rectangle $(a, b) \times (c, d) \subset \mathbb{R}^2$ and a continuously differentiable parameterized surface $\mathbf{X}(u, v)$ defined for $(u, v) \in (a, b) \times (c, d)$ and a $(u_0, v_0) \in (a, b) \times (c, d)$ such that:

- (i) $\mathbf{X}(u_0, v_0) = \mathbf{x}_0$, so that the parameterized surface passes through \mathbf{x}_0 .
- (ii) $\mathbf{X}(u, v) \in \mathcal{C}$ for all $(u, v) \in (a, b) \times (c, d)$, so that every point on the parameterized surface lies in \mathcal{S} .
- (iii) $\mathbf{X}(u_1, v_1) \neq \mathbf{X}(u_2, v_2)$ for any distinct points (u_1, v_1) and (u_2, v_2) in $(a, b) \times (c, d)$ so that the parameterized curve is one to one onto the portion of \mathcal{S} that it covers.
- (iv) There is an $r > 0$ such that every $\mathbf{x} \in \mathcal{S}$ with $\|\mathbf{x} - \mathbf{x}_0\| < r$ is of the form $\mathbf{X}(u, v)$ for some $(u, v) \in (a, b) \times (c, d)$.

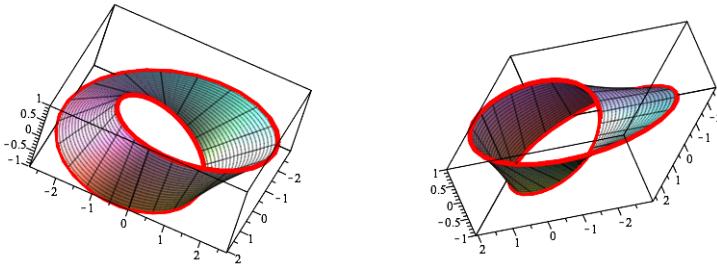
A smooth simple surface \mathcal{S} is orientable in casr it is possible to continuously assign a unit normal vector \mathbf{N} to each point of \mathcal{S} . In this case, the specification of \mathbf{N} is the orientation od \mathcal{S} , and \mathcal{S} becomes an oriented surface. We say that a parameterization $\mathbf{X}(u, v)$ of \mathcal{S} (or part of \mathcal{S}) is consistent with the orientation of \mathcal{S} in case $\mathbf{N}(\mathbf{X}(u, v))$ is a positive multiple of $\mathbf{X}_u \times \mathbf{X}_v(\mathbf{X}(u, v))$ for all parameter values u, v .

Not evey smooth simple surface is orientable. The classic example is the Möbius band:

Example 140 (Möbius band). Consider the function

$$\mathbf{X}(u, v) = (\cos u(2 + v \sin(u/2)), \sin u(2 + v \sin(u/2)), v \cos(u/2)), \quad u \in [0, 2\pi], \quad v \in (-1, 1) \quad (9.16)$$

The image of this function in \mathbb{R}^3 is a Möbius band \mathcal{S} . It is a “one sided surface”, and its boundary is a single edge. The bounding the curve \mathcal{C} given by fixing $v = 1$ in the parameterization, and letting u vary over $[0, 2\pi]$. The next plots show two views of the surface \mathcal{S} and its bounding curve \mathcal{C} .



Moving along the surface, one can get to what is locally the “other side” without crossing through the surface. Globally, the surface \mathcal{S} has only one side, and hence it connot be oriented – we cannot

designate one side as positive and one side as negative. This is very different from the unit sphere, which has an “inside” and an “outside”.

One might wonder why we cannot simply use $\mathbf{X}_u \times \mathbf{X}_v(u, v)$ to define $\mathbf{N}(u, v)$. The problem is that (9.16) is not a parameterization of \mathcal{S} . Note that for $v = 0$,

$$\mathbf{X}(u, 0) = (\cos u, \sin u, 0) \quad u \in [0, 2\pi] ,$$

and the same point $(1, 0, 0)$ in \mathcal{S} is two different sets of parameter values: $(u, v) = (0, 0)$ and $(u, v) = (2\pi, 0)$. A parameterization is required to be one-to-one. What is worse is that calculating one finds

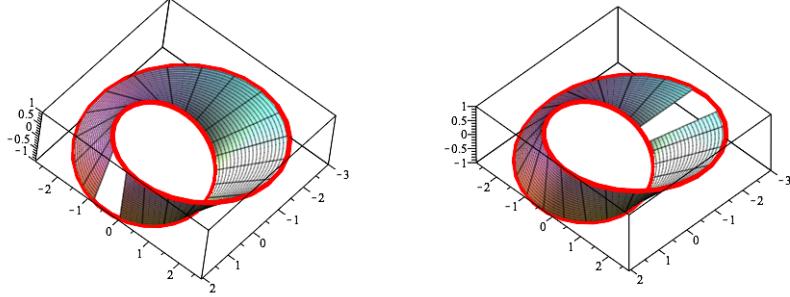
$$\mathbf{X}_u \times \mathbf{X}_v(u, 0) = 2(\cos(u/2) \cos u, \cos(u/2) \sin u, \sin(u/2)) .$$

so that evidently

$$\mathbf{X}_u \times \mathbf{X}_v(u, 0) = -\mathbf{X}_u \times \mathbf{X}_v(u + 2\pi, 0) .$$

Since $\mathbf{X}(0, 0) = \mathbf{X}(2\pi, 0) = (1, 0, 0)$, the function $\mathbf{X}_u \times \mathbf{X}_v(u, 0)$ assigns both normal directions to $(1, 0, 0) \in \mathcal{S}$.

If we restrict u to make the function $\mathbf{X}(u, v)$ one-to-one, we get a proper parameterization of part of the Möbius band. The next two figures show the parts covered by restricting $\mathbf{X}(u, v)$ to $(0, 2\pi/8) \times (-1, 1)$ and $(\pi, 23\pi/8) \times (-1, 1)$ respectively, and also the whole bounding curve \mathcal{C} to more clearly display what is left out:



Each of these parts is a two sided orientable surface. However there is no way to choose the orientations so that they are compatible on their overlap.

Given a smooth simple surface \mathcal{S} , and a Lipschitz vector field \mathbf{F} on \mathbb{R}^3 , we can use the flow generated Φ^t by \mathbf{F} to “sweep out” a region in \mathbb{R}^3 . The flux is the net rate rate at which volume is being swept out, taking into account whether it is being swept to the positive or negative side of \mathcal{S} :

Definition 101 (Flux across an oriented surface \mathcal{S}). Let \mathcal{S} be an oriented smooth simple surface with unit normal \mathbf{N} . Let \mathbf{F} be a continuously differentiable Lipschitz vector field on \mathbb{R}^3 . Suppose that either (i) $\mathbf{N} \cdot \mathbf{F} > 0$ everywhere along \mathcal{S} , or (ii) $\mathbf{N} \cdot \mathbf{F} < 0$ everywhere along \mathcal{S} . For $w \in \mathbb{R}$, let Φ^w be the flow transformation generated by \mathbf{F} at time w . For $t > 0$ define the set \mathcal{V}_t in the plane by

$$\mathcal{V}_t = \{\Phi^w(\mathbf{x}) : w \in (0, t), \mathbf{x} \in \mathcal{S}\} , \quad (9.17)$$

and let $\text{vol}(\mathcal{V}_t)$ denote the volume of \mathcal{V}_t , the set of points that is “swept out” in the time interval $(0, t)$ as the flow pushes \mathcal{S} around in \mathbb{R}^3 . We define the flux across \mathcal{S} generated by \mathbf{F} , $\text{flux}(\mathcal{S}, \mathbf{F})$ to be the quantity

$$\text{flux}(\mathcal{S}, \mathbf{F}) = \pm \frac{d}{dt} \text{vol}(\mathcal{V}_t) \Big|_{t=0}, \quad (9.18)$$

where the sign in front of the derivative is the sign of $\mathbf{F} \cdot \mathbf{N}$ along \mathcal{S} . If \mathcal{S} is a simple oriented curve that is a disjoint union of component \mathcal{C}_j , $j = 1, \dots, N$, satisfying the conditions above, we define

$$\text{flux}(\mathcal{C}, \mathbf{F}) = \sum_{j=1}^N \text{flux}(\mathcal{S}_j, \mathbf{F}). \quad (9.19)$$

The notion of flux depends on the orientations of the surface \mathcal{S} . It makes not sense at all to talk about flux across a Möbius band because it makes no sense to talk about a positive and negative side of a Möbius band. However, when \mathcal{S} is the boundary of a region in \mathbb{R}^3 ; e.g., the unit sphere which is the boundary of the unit ball, then \mathcal{S} is always orientable: One can choose \mathbf{N} to be the outward normal.

Fortunately, in analogy with what we saw in \mathbb{R}^2 , it is not necessary to compute $\text{vol}(\mathcal{V}_t)$ to compute $\text{flux}(\mathcal{S}, \mathbf{F})$:

Theorem 95. *Let \mathcal{S} be a smooth simple oriented surface, and let \mathbf{F} be a continuously differentiable vector field on \mathbb{R}^3 . Then the flux across \mathcal{S} generated by \mathbf{F} is given by*

$$\text{flux}(\mathcal{S}, \mathbf{F}) = \int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS, \quad (9.20)$$

which is the integral of the function $\mathbf{F} \cdot \mathbf{N}$ along \mathcal{S} with respect to surface area.

Proof. Suppose first that \mathcal{S} is covered by a single parameterization $\mathbf{X}(u, v)$, $(u, v) \in U \subset \mathbb{R}^2$. Later we can apply our conclusions to each component of a more general surface that is a disjoint union of such pieces using (9.19).

Let $\mathbf{u} = (u, v, w)$, and define the transformation $\mathbf{x}(\mathbf{u})$

$$\mathbf{x}(u, v, w) = \Phi^w(\mathbf{X}(u, v)).$$

Then \mathcal{V}_t is the image of $(0, t) \times U \subset \mathbb{R}^3$ under this transformation, and we may use the change of variables formula

$$\text{vol}(\mathcal{V}_t) = \int_{(0,t) \times U} |\det[D\mathbf{x}(u, v, w)]| d^3\mathbf{u}$$

to compute the volume of \mathcal{V}_t .

Let U_+ be the subset of U consisting of (u, v) such that $\mathbf{F}(\mathbf{X}(u, v)) \cdot \mathbf{N}(\mathbf{X}(u, v)) > 0$. For $(u_0, v_0) \in U_+$, the flow curve $\mathbf{x}(w) = \Phi^w(\mathbf{X}(u_0, v_0))$ is initially moving to the *positive* side of \mathcal{S} . Let \mathcal{V}_t^+ be the volume swept out by this part of the surface. .

Let U_- be the subset of U consisting of (u, v) such that $\mathbf{F}(\mathbf{X}(u, v)) \cdot \mathbf{N}(\mathbf{X}(u, v)) < 0$. For $(u_0, v_0) \in U_-$, the flow curve $\mathbf{x}(w) = \Phi^w(\mathbf{X}(u_0, v_0))$ is initially moving to the *negative* side of \mathcal{S} . Let \mathcal{V}_t^- be the volume swept out by this part of the surface.

There is no need to introduce U_0 , the part of U on which $\mathbf{F}(\mathbf{X}(u, v)) \cdot \mathbf{N}(\mathbf{X}(u, v)) = 0$ since this part of the surface will not make a contribution either way. We define the *flux across \mathcal{S}* generated

by \mathbf{F} by

$$\text{flux}(\mathcal{S}, \mathbf{F}) = \frac{d}{dt} \text{vol}(\mathcal{V}_t^+) \Big|_{t=0} - \frac{d}{dt} \text{vol}(\mathcal{V}_t^-) \Big|_{t=0}.$$

To calculate this using the change of variables formula, we consider the Jacobian matrix

$$[D\mathbf{x}(u, v, w)] = \left[\frac{\partial \mathbf{x}(u, v, w)}{\partial u}, \frac{\partial \mathbf{x}(u, v, w)}{\partial v}, \frac{\partial \mathbf{x}(u, v, w)}{\partial w} \right].$$

Since for fixed u and v , $\mathbf{x}(u, v, w)$ traces out a flow curve of \mathbf{F} as w varies,

$$\frac{\partial \mathbf{x}(u, v, w)}{\partial w} = \mathbf{F}(\mathbf{x}(u, v, w)).$$

This tells us the third column of the matrix. Next, since Φ^0 is the identity transformation $\mathbf{x}(u, v, 0) = \mathbf{X}(u, v)$ and so

$$\frac{\partial}{\partial u} \mathbf{x}(u, v, 0) = \mathbf{X}_u(u, v) \quad \text{and} \quad \frac{\partial}{\partial v} \mathbf{x}(u, v, 0) = \mathbf{X}_v(u, v).$$

Therefore, we can evaluate the Jacobian determinant at $w = 0$:

$$\det[D\mathbf{x}(u, v, 0)] = \det([\mathbf{X}_u(u, v), \mathbf{X}_v(u, v), \mathbf{F}(\mathbf{X}(u, v))]) = \mathbf{X}_u \times \mathbf{X}_v \cdot \mathbf{F}(\mathbf{X})(u, v).$$

Therefore, since $\mathbf{X}_u \times \mathbf{X}_v(u, v)$ is a positive multiple of $\mathbf{N}(\mathbf{X}(u, v))$, we see that $\det[D\mathbf{x}(u, v, 0)] > 0$ for $(u, v) \in U_+$ and $\det[D\mathbf{x}(u, v, 0)] < 0$ for $(u, v) \in U_-$. By continuity, the same is true for all sufficiently small values of w . Therefore,

$$\frac{d}{dt} \text{vol}(\mathcal{V}_t^+) \Big|_{t=0} = \frac{d}{dt} \left(\int_{(0,t) \times U_+} \det[D\mathbf{x}(u, v, w)] d^3 \mathbf{u} \right) \Big|_{t=0}$$

and

$$-\frac{d}{dt} \text{vol}(\mathcal{V}_t^-) \Big|_{t=0} = \frac{d}{dt} \left(\int_{(0,t) \times U_-} \det[D\mathbf{x}(u, v, w)] d^3 \mathbf{u} \right) \Big|_{t=0}$$

where we have dropped the absolute values signs on the determinant, but taken the negative sign of the determinant into account in the second formula with the initial minus sign.

Now it is possible to combine terms to obtain

$$\text{flux}(\mathcal{S}, \mathbf{F}) = \frac{d}{dt} \left(\int_{(0,t) \times U} \det[D\mathbf{x}(u, v, w)] d^3 \mathbf{u} \right) \Big|_{t=0}. \quad (9.21)$$

Just as in two dimensions, the sign of the Jacobian determinant automatically takes into account the orientation so that we get the correct net flux simply by dropping the absolute value sign in the change of variables formula. Finally,

$$\int_{(0,t) \times U} \det[D\mathbf{x}(u, v, w)] d^3 \mathbf{u} = \int_0^t \left(\int_U \det[D\mathbf{x}(u, v, w)] du dv \right) dw,$$

and so

$$\begin{aligned} \frac{d}{dt} \left(\int_{(0,t) \times U} \det[D\mathbf{x}(u, v, w)] d^3 \mathbf{u} \right) \Big|_{t=0} &= \int_U \det[D\mathbf{x}(u, v, 0)] du dv \\ &= \int_U \mathbf{X}_u \times \mathbf{X}_v \cdot \mathbf{F}(\mathbf{X})(u, v) du dv \\ &= \int_U \mathbf{N}(\mathbf{X}) \cdot \mathbf{F}(\mathbf{X}) \|\mathbf{X}_u \times \mathbf{X}_v\| du dv \\ &= \int_{\mathcal{S}} \mathbf{N} \cdot \mathbf{F} dS. \end{aligned}$$

Combining this with (9.21) we have $\text{flux}(\mathcal{S}, \mathbf{F}) = \int_{\mathcal{S}} \mathbf{N} \cdot \mathbf{F} dS$. \square

9.1.5 Computing flux integrals

Flux integrals are often easier to compute than surface area integrals since the combination $\mathbf{N}dS$ is in some ways simpler than dS itself. To see why, let $\mathbf{X}(u, v)$ be a parameterization of an oriented surface \mathcal{S} . Then

$$\mathbf{N}(u, v) = \pm \frac{1}{\|\mathbf{X}_u \times \mathbf{X}_v(u, v)\|} \mathbf{X}_u \times \mathbf{X}_v(u, v) \quad \text{and} \quad dS = \|\mathbf{X}_u \times \mathbf{X}_v(u, v)\| du dv ,$$

so that

$$\mathbf{N}dS = \pm \mathbf{X}_u \times \mathbf{X}_v(u, v) du dv \quad (9.22)$$

where the plus sign is valid if the parameterization is consistent with the orientation of \mathcal{S} , and the minus sign otherwise. The good thing that happens is that two factors of $\|\mathbf{X}_u \times \mathbf{X}_v(u, v)\|$ cancel out when forming $\mathbf{N}dS$.

Example 141. Let \mathbf{F} be the vector field $\mathbf{F}(x, y, z) = (2xyz - y^2, x^2z - 2xy, x^2y)$. Let \mathcal{S} be the part of the paraboloid $z = 1 - x^2 - y^2$ that lies above the x, y plane, oriented so that its unit normal \mathbf{N} points upward. To compute that flux integral $\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS$ directly, the first step is to parameterize \mathcal{S} . Using cylindrical coordinates, and using $z = 1 - r^2$ to eliminate z , we obtain

$$\mathbf{X}(r, \theta) = (r \cos \theta, r \sin \theta, 1 - r^2) \quad \text{with } r \in (0, 1) \theta \in (0, 2\pi) .$$

Differentiating,

$$\mathbf{X}_r(r, \theta) = (\cos \theta, \sin \theta, -2r) \quad \text{and} \quad \mathbf{X}_{\theta}(r, \theta) = (-r \sin \theta, r \cos \theta, 0) .$$

Then

$$\mathbf{X}_r \times \mathbf{X}_{\theta}(r, \theta) = (2r^2 \cos \theta, 2r^2 \sin \theta, r) .$$

Since the third component is $r > 0$, this points upward, and so the parameterization is consistent with the orientation. Hence

$$\mathbf{N}dS = (2r^2 \cos \theta, 2r^2 \sin \theta, r) dr d\theta .$$

We now evaluate \mathbf{F} on the surface:

$$\mathbf{F}(\mathbf{X}(r, \theta)) = (2r^2(1 - r^2) \cos \theta \sin \theta - r^2 \sin^2 \theta, r^2(1 - r^2) \cos^2 \theta - 2r^2 \cos \theta \sin \theta, r^3 \cos^2 \theta \sin \theta) .$$

Hence the flux element is

$$\begin{aligned} \mathbf{F}(\mathbf{X}(r, \theta)) \cdot \mathbf{N}(\mathbf{X}(r, \theta)) dr d\theta &= [2r^2(1 - r^2) \cos \theta \sin \theta - r^2 \sin^2 \theta][2r^2 \cos \theta] dr d\theta \\ &+ [r^2(1 - r^2) \cos^2 \theta - 2r^2 \cos \theta \sin \theta][2r^2 \sin \theta] dr d\theta \\ &+ [r^3 \cos^2 \theta \sin \theta][r] dr d\theta . \end{aligned}$$

We now integrate. However, since

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

the integral of each term is zero, and finally we find $\text{flux}(\mathbf{F}, \mathcal{S}) = 0$.

9.2 The Divergence Theorem

9.2.1 The Divergence Theorem in the plane

The flux through a simple closed curve \mathcal{C} in the plane that is generated by a vector field \mathbf{F} can be computed by integrating a *flux density* over the region \mathcal{D} enclosed by \mathcal{C} . The flux density is the *divergence of \mathbf{F}* that we now define, once and for all, in \mathbb{R}^n for general n .

Definition 102 (Divergence of a vector field). *Let \mathbf{F} be a continuously differentiable vector field on \mathbb{R}^n . The divergence of \mathbf{F} is the real valued function $\text{div}\mathbf{F}$ defined by*

$$\text{div}\mathbf{F}(\mathbf{x}) = \sum_{j=1}^n \frac{\partial}{\partial x_j} \mathbf{e}_j \cdot \mathbf{F}(\mathbf{x}) \quad (9.23)$$

It is tradition in Vector Calculus to use P and Q for the components of two dimensional vector fields, and P , Q and R for the components of three dimensional vector field, so that we would write a two dimensional vector field as $\mathbf{F}(x, y) = (P(x, y), Q(x, y))$ and then

$$\text{div}\mathbf{F}(x, y) = \frac{\partial P(x, y)}{\partial x} + \frac{\partial Q(x, y)}{\partial y} .$$

We would write a three dimensional vector field as $\mathbf{F}(x, y, z) = (P(x, y, z), Q(x, y, z), R(x, y, z))$ and then

$$\text{div}\mathbf{F}(x, y, z) = \frac{\partial P(x, y, z)}{\partial x} + \frac{\partial Q(x, y, z)}{\partial y} + \frac{\partial R(x, y, z)}{\partial z} .$$

The Divergence Theorem in the plane says that the divergence is a flux density in that the total flux out of a region \mathcal{D} , flowing across its boundary \mathcal{C} , is given by integrating the divergence if \mathbf{F} over \mathcal{D} :

Theorem 96 (The Divergence Theorem in \mathbb{R}^2). *Let \mathcal{C} be a simple closed curve in the plane, and let \mathcal{D} be the domain enclosed by \mathcal{C} . Let \mathbf{N} be the outward normal along \mathcal{C} . Let \mathbf{F} be a continuously differentiable Lipschitz vector field on \mathbb{R}^2 . Then*

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = \int_{\mathcal{D}} \text{div}\mathbf{F}(\mathbf{x}) d^2\mathbf{x} . \quad (9.24)$$

Example 142. *Let $\mathbf{F}(x, y) = (xy, x^2 - y^2)$. Let \mathcal{C} be the unit circle centered at $(1, 1)$, oriented so that the unit normal is outward. We compute $\text{div}\mathbf{F}(x, y) = -y$. Therefore, if \mathcal{D} is the unit disc centered at $(1, 1)$, the domain bounded by \mathcal{C} ,*

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = - \int_{\mathcal{D}} y d^2\mathbf{x} .$$

By symmetry, the average value of y in \mathcal{D} is 1. Hence

$$\frac{\int_{\mathcal{D}} y d^2\mathbf{x}}{\int_{\mathcal{D}} 1 d^2\mathbf{x}} = 1 .$$

Thus, $-\int_{\mathcal{D}} y d^2\mathbf{x} = -\int_{\mathcal{D}} 1 d^2\mathbf{x} = \pi$, and again we obtain $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = -\pi$ but the integrals are much simpler.

The Divergence Theorem is often useful for computing the flux across *open* curves as well as the closed curves to which it applies directly. The next example explains how.

Example 143. Let \mathcal{C}_1 be the part of the parabola $y = 4 - x^2$ lying above the x -axis, and oriented so that \mathbf{N} points upwards. Let $\mathbf{F}(x, y) = (x^3y - y^2 + x, x^2y - 3x + 5y)$.

The endpoints of \mathcal{C}_1 are $(-2, 0)$ and $(2, 0)$. Let \mathcal{C}_2 be the straight line segment from $(-2, 0)$ to $(2, 0)$. Finally, let \mathcal{C} be the simple closed curve that runs from $(-2, 0)$ to $(2, 0)$ along \mathcal{C}_2 , and then from $(2, 0)$ to $(-2, 0)$ along \mathcal{C}_1 .

The curve \mathcal{C} enclosed the domain \mathcal{D} given by $0 \leq y \leq 4 - x^2$, which is under the parabola $y = 4 - x^2$ and above the x -axis. The outward unit normal on the boundary of \mathcal{D} is the unit normal along \mathcal{C} with the specified direction of travel.

We can compute the arc length integral of $\mathbf{F} \cdot \mathbf{N}$ by first integrating over \mathcal{C}_1 , and then continuing to integrate over \mathcal{C}_2 :

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds = \int_{\mathcal{C}_1} \mathbf{F} \cdot \mathbf{N} ds + \int_{\mathcal{C}_2} \mathbf{F} \cdot \mathbf{N} ds .$$

Then by the Divergence Theorem,

$$\int_{\mathcal{C}_1} \mathbf{F} \cdot \mathbf{N} ds = \int_{\mathcal{D}} \operatorname{div} \mathbf{F} d^2 \mathbf{x} - \int_{\mathcal{C}_2} \mathbf{F} \cdot \mathbf{N} ds .$$

It is easy to evaluate both of the integrals on the right: First, $\operatorname{div} \mathbf{F}(x, y) = 3x^2y + 6 + x^2$. The region \mathcal{D} is given by

$$0 \leq y \leq 4 - x^2 \quad \text{and} \quad -2 < x < 2 .$$

Hence,

$$\begin{aligned} \int_{\mathcal{D}} \operatorname{div} \mathbf{F} d^2 \mathbf{x} &= \int_{-2}^2 \left(\int_0^{4-x^2} [3x^2y + 6 + x^2] dy \right) dx \\ &= \int_{-2}^2 \left[48 - 14x^2 + \frac{1}{2}x^4 \right] dx = \frac{1856}{15} . \end{aligned}$$

Even more simply, we parameterize \mathcal{C}_2 by $\mathbf{x}(t) = (-2 + 4t, 0)$, $t \in (0, 1)$. Since $y = 0$ all along \mathcal{C}_2 , $\mathbf{F}(\mathbf{x}(t))$ is much simpler than \mathbf{F} in general:

$$\mathbf{F}(\mathbf{x}(t)) = (t - 2, 6 - 3t) .$$

Then $-(\mathbf{x}'(t))^\perp = (0, -4)$. Hence

$$\int_{\mathcal{C}_2} \mathbf{F} \cdot \mathbf{N} ds = 4 \int_0^1 (3t - 6) dt = -18 .$$

The proof of Theorem 96 rests on two lemmas that are of interest in their own right.

Lemma 30. Let \mathcal{C} be a simple closed curve in the plane. Let \mathcal{D} be the region in the plane enclosed by \mathcal{C} . Let \mathbf{F} be a continuously differentiable Lipschitz vector field on \mathbb{R}^2 , and let Φ be the flow generated by \mathbf{F} . For $t \in \mathbb{R}$, let \mathcal{D}_t be image of \mathcal{D} under the flow transformation Φ^t :

$$\mathcal{D}_t = \{ \Phi^t(\mathbf{u}) : \mathbf{u} \in \mathcal{D} \} . \tag{9.25}$$

$$\frac{d}{dt} \operatorname{area}(\mathcal{D}_t) \Big|_{t=0} = \int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds . \tag{9.26}$$

Before going into the proof, it will be good to consider what Lemma 30 says. If $\mathbf{F} \cdot \mathbf{N} > 0$ everywhere along \mathcal{C} , the vector field is “pushing” outward everywhere along \mathcal{C} , and area is getting “swept out” of \mathcal{D} to cover a strictly large domain \mathcal{D}_t . Evidently, in this case $\text{area}(\mathcal{D}_t)$ is an increasing function of t , and the rate of increase of $\text{area}(\mathcal{D}_t)$, as we have defined \mathcal{D}_t , corresponds to positive outward flux.

Proof of Lemma 30. Notice that

$$\mathcal{D}_t = (\mathcal{D}_t \cap \mathcal{D}) \cup (\mathcal{D}_t \cap \mathcal{D}^c) \quad \text{and} \quad \mathcal{D} = (\mathcal{D} \cap \mathcal{D}_t) \cup (\mathcal{D} \cap \mathcal{D}_t^c).$$

Next, $\mathbf{x} \in \mathcal{D}_t \cap \mathcal{D}^c$ means that $\mathbf{x} \notin \mathcal{D}$ but it is the image under Φ^t of a point that was originally in \mathcal{D} . Thus, $\mathcal{D}_t \cap \mathcal{D}^c$ consists of points that are carried outside \mathcal{D} from the inside by the flow in time t . Likewise, $\mathbf{x} \in \mathcal{D} \cap \mathcal{D}_t^c$ means that $\mathbf{x} \in \mathcal{D}$, but \mathbf{x} is the image under Φ^t of a point \mathbf{u} that was outside \mathcal{D} . Thus, $\mathcal{D} \cap \mathcal{D}_t^c$ consists of points that are carried inside of \mathcal{D} from the outside by the flow in time t . Therefore, differentiating both sides of (9.27) in t , Of course, $\mathcal{D}_t \cap \mathcal{D}$ is the part of \mathcal{D} that is left in \mathcal{D} by Φ^t . Evidently the three regions $\mathcal{D}_t \cap \mathcal{D}$, $\mathcal{D}_t \cap \mathcal{D}^c$ and $\mathcal{D}_t^c \cap \mathcal{D}$ are mutually disjoint. Thus,

$$\text{area}(\mathcal{D}_t) = \text{area}(\mathcal{D}_t \cap \mathcal{D}) + \text{area}(\mathcal{D}_t \cap \mathcal{D}^c) \quad \text{and} \quad \text{area}(\mathcal{D}_t) - \text{area}(\mathcal{D}) = \text{area}(\mathcal{D}_t \cap \mathcal{D}^c) - \text{area}(\mathcal{D} \cap \mathcal{D}_t^c).$$

Therefore,

$$\text{area}(\mathcal{D}_t) - \text{area}(\mathcal{D}) = \text{area}(\mathcal{D}_t \cap \mathcal{D}^c) - \text{area}(\mathcal{D} \cap \mathcal{D}_t^c). \quad (9.27)$$

Differentiating both sides of (9.27) in t ,

$$\frac{d}{dt} \text{area}(\mathcal{D}_t) \Big|_{t=0} = \frac{d}{dt} \text{area}(\mathcal{D}_t \cap \mathcal{D}^c) \Big|_{t=0} - \frac{d}{dt} \text{area}(\mathcal{D} \cap \mathcal{D}_t^c) \Big|_{t=0}.$$

The right hand side is the net flux out of \mathcal{D} under the flow, which is given by the integral $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds$ where \mathbf{N} is the outward unit normal, and this proves (9.26). \square

We now compute a formula for the left hand side of (9.27) be applying the change of variables formula for integrals in \mathbb{R}^2 .

The point is that the transformation Φ^t maps \mathcal{D} onto \mathcal{D}_t by definition, so we can use it to parameterize \mathcal{D}_t by \mathcal{D} , which is exactly what is done in (9.25). That is, define $\mathbf{x}^t(\mathbf{u}) = \Phi^t(\mathbf{u})$. Then $\mathbf{u} \in \mathcal{D}$ if and only if $\mathbf{x} \in \mathcal{D}_t$. Therefore, for each t , we define the transformation $\mathbf{x}^t(u, v) = \Phi^{-t}(u, v)$

By the change of variables formula,

$$\text{area}(\mathcal{D}_t) = \int_{\mathcal{D}_t} 1 d^2 \mathbf{x} = \int_{\mathcal{D}} |\det[D\mathbf{x}^t(\mathbf{u})]| d^2 \mathbf{u}. \quad (9.28)$$

Lemma 31. *Let \mathbf{F} be a continuously differentiable Lipschitz vector field on \mathbb{R}^2 , and for $t \in \mathbb{R}$, let Φ^t be the corresponding flow transformation. Define a transformation $\mathbf{x}^t(u, v)$ on \mathbb{R}^2 by $\mathbf{x}^t(u, v) = \Phi^t(u, v)$, and let $[D\mathbf{x}^t(u, v)]$ be the Jacobian matrix of this transformation. Then for all sufficiently small t , $\det([D\mathbf{x}^t(u, v)]) > 0$, so that*

$$\frac{d}{dt} |\det([D\mathbf{x}^t(u, v)])| \Big|_{t=0} = \frac{d}{dt} \det([D\mathbf{x}^t(u, v)]) \Big|_{t=0},$$

so that the absolute value on the determinant in (9.28) is redundant when we differentiate at $t = 0$, and moreover,

$$\frac{d}{dt} \det([D\mathbf{x}^t(u, v)]) \Big|_{t=0} = \operatorname{div}\mathbf{F}(u, v) . \quad (9.29)$$

Granted the validity of Lemma 31, it is now easy to prove the Divergence Theorem:

Proof of Theorem 96. Differentiating both sides of (9.28) in t at $t = 0$, and taking the derivative under the integral sign on the right (which can be justified, though we shall not do it here), and then applying (9.29),

$$\frac{d}{dt} \operatorname{area}(\mathcal{D}_t) \Big|_{t=0} = \int_{\mathcal{D}} \frac{d}{dt} \det([D\mathbf{x}^t(u, v)]) \Big|_{t=0} d^2\mathbf{u} = \int_{\mathcal{D}} \operatorname{div}\mathbf{F}(u, v) d^2\mathbf{u} .$$

Applying Lemma 30, we obtain (9.38). \square

Proof of Lemma 31. The Jacobian matrix of the transformation $\mathbf{x}^t(u, v)$ is

$$[D\mathbf{x}^t(\mathbf{u})] = \left[\frac{\partial \mathbf{x}^t(u, v)}{\partial u}, \frac{\partial \mathbf{x}^t(u, v)}{\partial v} \right] .$$

Since $\mathbf{x}^0(u, v) = (u, v)$, the Jacobian matrix at $t = 0$ is the identity matrix:

$$[D\mathbf{x}^0(\mathbf{u})] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} . \quad (9.30)$$

Therefore, $\det([D\mathbf{x}^0(\mathbf{u})]) = 1$, and then by continuity, $\det([D\mathbf{x}^t(\mathbf{u})]) > 0$ for all sufficiently small t , proving the first part of the lemma. Now recall that

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc = (a, c)^\perp \cdot (b, d) .$$

Hence, for $t \neq 0$, the Jacobian determinant is

$$\det[D\mathbf{x}^t(\mathbf{u})] = \det \left[\frac{\partial \mathbf{x}^t(u, v)}{\partial u}, \frac{\partial \mathbf{x}^t(u, v)}{\partial v} \right] = \left(\frac{\partial \mathbf{x}^t(u, v)}{\partial u} \right)^\perp \cdot \frac{\partial \mathbf{x}^t(u, v)}{\partial v} .$$

Differentiating the dot product, we get two terms. These are easy to evaluate at $t = 0$ if we use the fact that

$$\frac{\partial \mathbf{x}^0(u, v)}{\partial u} = \mathbf{e}_1 \quad \text{and} \quad \frac{\partial \mathbf{x}^0(u, v)}{\partial v} = \mathbf{e}_2 ,$$

which follows from (9.30). We find

$$\begin{aligned} \frac{d}{dt} \left(\left(\frac{\partial \mathbf{x}^t(u, v)}{\partial u} \right)^\perp \cdot \frac{\partial \mathbf{x}^t(u, v)}{\partial v} \right) &= \left(\frac{\partial^2 \mathbf{x}^t(u, v)}{\partial t \partial u} \right)^\perp \cdot \mathbf{e}_2 + \mathbf{e}_1^\perp \cdot \frac{\partial^2 \mathbf{x}^t(u, v)}{\partial t \partial v} \\ &= \frac{\partial^2 \mathbf{x}^t(u, v)}{\partial u \partial t} \cdot \mathbf{e}_1 + \frac{\partial^2 \mathbf{x}^t(u, v)}{\partial v \partial t} \cdot \mathbf{e}_2 , \end{aligned}$$

where we have used Clairault's Theorem to interchange the order of the temporal and spatial derivatives. Now since for fixed u and v , $\mathbf{x}^t(u, v)$ traces out a flow curve of \mathbf{F} as t varies,

$$\frac{\partial}{\partial t} \mathbf{x}^t(u, v) = \mathbf{F}(\mathbf{x}^t(u, v)) .$$

Evaluating this at $t = 0$, and recalling once more that $\mathbf{x}^t(u, v) = (u, v)$,

$$\begin{aligned}\frac{d}{dt} \left(\left(\frac{\partial \mathbf{x}^t(u, v)}{\partial u} \right)^\perp \cdot \frac{\partial \mathbf{x}^t(u, v)}{\partial v} \right) \Big|_{t=0} &= \frac{\partial}{\partial u} (\mathbf{F}(u, v) \cdot \mathbf{e}_1) \frac{\partial}{\partial v} (\mathbf{F}(u, v) \cdot \mathbf{e}_2) \\ &= \operatorname{div} \mathbf{F}(u, v) .\end{aligned}$$

□

9.2.2 The Divergence Theorem in \mathbb{R}^3

A direct analog of the Divergence Theorem is true in every dimension, and the proof is very much like the proof in the planar case. We discuss this next in \mathbb{R}^3 .

While the boundary of a domain \mathcal{D} in \mathbb{R}^2 that has no “holes” in it is a simple closed curve, the boundary of a domain \mathcal{D} in \mathbb{R}^3 that has no holes in it is a simple closed surface \mathcal{S} . Think of the unit ball in \mathbb{R}^3 ; its boundary is the unit sphere in \mathbb{R}^3 .

It is now easy to prove the three dimensional version of the Divergence Theorem. Let \mathcal{S} be a simple closed surface that encloses a region $\mathcal{V} \subset \mathbb{R}^3$.

We define a time-dependent region \mathcal{V}_t in analogy with (9.25):

$$\mathcal{V}_t = \{ \Phi^t(\mathbf{u}) : \mathbf{u} \in \mathcal{D} \} . \quad (9.31)$$

Notice that this definition of \mathcal{V}_t is different from the one in (9.17). There, the third component w of (u, v, w) represented time which varied over an interval $(0, t)$. Here the time t is fixed, and (u, v, w) denotes a point in \mathcal{V} . Also notice that we use a negative time t in the flow in (9.31).

The same reasoning that led from (9.25) to (9.27) leads to

$$\text{area}(\mathcal{V}_t) - \text{area}(\mathcal{V}) = \text{area}(\mathcal{V}_t \cap \mathcal{V}^c) - \text{area}(\mathcal{V} \cap \mathcal{V}_t^c) . \quad (9.32)$$

and then the same reason that lead from (9.27) to (9.26) leads from (9.32) to

$$\frac{d}{dt} \text{vol}(\mathcal{V}_t) \Big|_{t=0} = \int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS . \quad (9.33)$$

We now compute the volume on the left hand side using the change of variables formula: For any fixed t , define the transformation $\mathbf{x}^t(\mathbf{u})$ that send $\mathbf{u} = (u, v, w)$ to the point $\Phi^t(u, v, w)$. Let us write

$$\mathbf{x}^t(u, v, w) = (x(t, u, v, w), y(t, u, v, w), z(t, u, v, w)) .$$

The Jacobian matrix of this transformation is

$$[D\mathbf{x}^t(u, v, w)] = \left[\frac{\partial \mathbf{x}^t}{\partial u}, \frac{\partial \mathbf{x}^t}{\partial v}, \frac{\partial \mathbf{x}^t}{\partial w} \right] . \quad (9.34)$$

Since $\Phi^0(\mathbf{u}) = \mathbf{u}$, $\mathbf{x}^0(u, v, w) = (u, v, w)$, and so the Jacobian matrix is simply the identity matrix at $t = 0$:

$$[D\mathbf{x}^0(u, v, w)] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \quad (9.35)$$

Therefore $\det[D\mathbf{x}^0(u, v, w)] = 1$, and by continuity, $\det[D\mathbf{x}^t(u, v, w)] > 0$ for all sufficiently small t .

This means that we have

$$\text{vol}(\mathcal{V}_t) = \int_{\mathcal{V}} \det[D\mathbf{x}^t(u, v, w)] d^3\mathbf{u},$$

where no absolute value sign on the determinant is needed. Therefore, taking the derivative under the integral sign,

$$\frac{d}{dt} \text{vol}(\mathcal{V}_t) \Big|_{t=0} = \int_{\mathcal{V}} \frac{d}{dt} \det[D\mathbf{x}^t(u, v, w)] \Big|_{t=0} d^3\mathbf{u}. \quad (9.36)$$

By (9.34),

$$\det[D\mathbf{x}^t(u, v, w)] = \frac{\partial \mathbf{x}^t}{\partial u} \times \frac{\partial \mathbf{x}^t}{\partial v} \cdot \frac{\partial \mathbf{x}^t}{\partial w}.$$

Differentiating in t , we get three terms. These can be written simply evaluating at $t = 0$ since by (9.35)

$$\frac{\partial \mathbf{x}^0}{\partial u} = \mathbf{e}_1, \quad \frac{\partial \mathbf{x}^0}{\partial v} = \mathbf{e}_2 \quad \text{and} \quad \frac{\partial \mathbf{x}^0}{\partial w} = \mathbf{e}_3.$$

Therefore,

$$\begin{aligned} \frac{d}{dt} \det[D\mathbf{x}^t(u, v, w)] \Big|_{t=0} &= \left(\frac{\partial \mathbf{x}^0}{\partial t \partial u} \times \mathbf{e}_2 \cdot \mathbf{e}_3 \right) + \left(\mathbf{e}_1 \times \frac{\partial \mathbf{x}^0}{\partial t \partial v} \cdot \mathbf{e}_3 \right) + \left(\mathbf{e}_1 \times \mathbf{e}_2 \cdot \frac{\partial \mathbf{x}^0}{\partial t \partial w} \right) \\ &= \left(\mathbf{e}_2 \times \mathbf{e}_3 \cdot \frac{\partial \mathbf{x}^0}{\partial u \partial t} \right) + \left(\mathbf{e}_3 \times \mathbf{e}_1 \cdot \frac{\partial \mathbf{x}^0}{\partial v \partial t} \right) + \left(\mathbf{e}_1 \times \mathbf{e}_2 \cdot \frac{\partial \mathbf{x}^0}{\partial w \partial t} \right) \\ &= \left(\mathbf{e}_1 \cdot \frac{\partial \mathbf{x}^0}{\partial u \partial t} \right) + \left(\mathbf{e}_2 \cdot \frac{\partial \mathbf{x}^0}{\partial v \partial t} \right) + \left(\mathbf{e}_3 \cdot \frac{\partial \mathbf{x}^0}{\partial w \partial t} \right), \end{aligned}$$

where we have used symmetries of the triple product and also have used Clairault's Theorem to change the order of the spatial and time derivatives. Now since $\mathbf{x}^t(u, v, w) = \Phi^t(u, v, w)$, holding u, v, w fixed and differentiating in t we have

$$\frac{\partial \mathbf{x}^t(u, v, w)}{\partial t} = \mathbf{F}(\mathbf{x}^t(u, v, w)).$$

Therefore, evaluating at $t = 0$, and recalling that $\mathbf{x}^0(\mathbf{u}) = \mathbf{u}$,

$$\frac{d}{dt} \det[D\mathbf{x}^t(u, v, w)] \Big|_{t=0} = \text{div}\mathbf{F}(u, v, w).$$

Then, (9.36) becomes

$$\frac{d}{dt} \text{vol}(\mathcal{V}_t) \Big|_{t=0} = \int_{\mathcal{V}} \text{div}\mathbf{F}(u, v, w) d^3\mathbf{u}. \quad (9.37)$$

Combining this with (9.33) we have proved:

Theorem 97 (The Divergence Theorem in \mathbb{R}^3). *Let \mathcal{S} be a simple closed surface in \mathbb{R}^3 , and let \mathcal{V} be the domain in \mathbb{R}^3 enclosed by \mathcal{S} . Let \mathbf{N} be the outward normal on \mathcal{S} . Let \mathbf{F} be a continuously differentiable Lipschitz vector field on \mathbb{R}^3 . Then*

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS = \int_{\mathcal{V}} \text{div}\mathbf{F}(\mathbf{x}) d^3\mathbf{x}. \quad (9.38)$$

This approach to computing flux via divergence extends easily to higher dimensions. This is taken up in the exercises.

Example 144. Let \mathcal{V} be the region in \mathbb{R}^3 that lies inside the sphere $x^2 + y^2 + z^2 = 4$, and above the graph of $z = 1/\sqrt{x^2 + y^2}$. Let \mathcal{S} be its boundary, equipped with the outward normal \mathbf{N} .

Let $\mathbf{F}(x, y, z) = (x, x, z)$. Compute $\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS$ where \mathbf{N} is the outward unit normal vector. We compute $\operatorname{div}(\mathbf{F}) = 2$, and hence by the Divergence Theorem,

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS = 2 \int_{\mathcal{V}} dV .$$

Using cylindrical coordinates, the limits of integration are

$$0 \leq \theta \leq 2\pi \quad 1/r \leq z \leq \sqrt{4 - r^2} \quad \text{and} \quad \sqrt{2 - \sqrt{3}} \leq r \leq \sqrt{2 + \sqrt{3}} .$$

Thus,

$$\begin{aligned} \int_{\mathcal{V}} dV &= 2\pi \int_{\sqrt{2-\sqrt{3}}}^{\sqrt{2+\sqrt{3}}} \left(\int_{1/r}^{\sqrt{4-r^2}} dz \right) r dr \\ &= 2\pi \int_{\sqrt{2-\sqrt{3}}}^{\sqrt{2+\sqrt{3}}} \left(r\sqrt{4-r^2} - 1 \right) dr \\ &= 2\pi \frac{2^{3/2}}{3} . \end{aligned}$$

The final answer is

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS = \pi \frac{2^{7/2}}{3} .$$

There are two pieces to the boundary, and this is much, much simpler than parameterizing both of them and doing the direct calculation.

In analogy with what we have seen in the plane, the Divergence Theorem in \mathbb{R}^3 is not only useful for computing the flux across closed surfaces \mathcal{S} , but also open surfaces

Example 145 (Trading one surface in on another). There is a better way to compute the flux integral in Example 141. Let \mathcal{S}_1 be surface introduced there, the part of the paraboloid $z = 1 - x^2 - y^2$ above the x, y plane with the upward unit normal \mathbf{N} . Let \mathcal{S}_2 be the unit disk in the x, y plane with the downward unit normal. Then let $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. Note that \mathcal{S} is the boundary of the region

$$\mathcal{V} = \{(x, y, z) : 0 \leq z \leq 1 - x^2 - y^2\} ,$$

oriented with the outward unit normal. Therefore, by the Divergence Theorem,

$$\operatorname{flux}(\mathbf{F}, \mathcal{S}_1) + \operatorname{flux}(\mathbf{F}, \mathcal{S}_2) = \operatorname{flux}(\mathbf{F}, \mathcal{S}) = \int_{\mathcal{V}} \operatorname{div} \mathbf{F} d^3 \mathbf{x} .$$

Hence if we compute $\operatorname{flux}(\mathbf{F}, \mathcal{S}_2)$ and $\int_{\mathcal{V}} \operatorname{div} \mathbf{F} d^3 \mathbf{x}$ we will have determined $\operatorname{flux}(\mathbf{F}, \mathcal{S}_1)$.

On \mathcal{S}^2 , things are easy: $\mathbf{N} = (0, 0, -1)$ everywhere on \mathcal{S}_2 , and $dS = d^2 \mathbf{x}$, the usual planar area element. Also since $z = 0$ everywhere on \mathcal{S}_2 , we have a relatively simple expression for \mathbf{F} on \mathcal{S}_2 :

$$\mathbf{F}(x, y, 0) = (y^2, -2xy, x^2y) .$$

Hence on \mathcal{S}_2 , $\mathbf{F} \cdot \mathbf{N} dS = -x^2 y d^2 \mathbf{x}$. Doing the integral in polar coordinates, we immediately find $\int_{\mathcal{S}_2} \mathbf{F} \cdot \mathbf{N} dS = 0$.

Next, we compute $\operatorname{div}\mathbf{F}(x, y, z) = 2yz - 2x$. Notice how much simpler this is than \mathbf{F} itself. Also, since \mathcal{V} is invariant under the transformation sending (x, y, z) to $(-x, -y, z)$ (which is a rotation by π) and since $2yz - 2x$ changes sign under this transformation, it is immediate that $\int_{\mathcal{V}} \operatorname{div}\mathbf{F} d^3\mathbf{x} = 0$, though it is also not hard to do the integral using cylindrical coordinates.

9.3 Line integrals, force fields and work

Let \mathbf{F} be a continuous vector field on \mathbb{R}^n . In this section, we think of \mathbf{F} as representing a *force field*; that is \mathbf{F} gives the force that acts on a point particle located at \mathbf{x} . For instance, if some electric charges are distributed in \mathbb{R}^3 , they will produce an electric field $\mathbf{E}(\mathbf{x})$, and then any point particle at \mathbf{x} with an electrical charge q will be acted upon by a force $\mathbf{F}(\mathbf{x}) = q\mathbf{E}(\mathbf{x})$. Let $\mathbf{x}(t)$, $a \leq t \leq b$, be a continuously differentiable parameterized curve in \mathbb{R}^n . Suppose we move the point particle along the path $\mathbf{x}(t)$. We ask: How much work is done on the point particle as it moves along the curve from $\mathbf{x}_0 := \mathbf{x}(a)$ to $\mathbf{x}_1 := \mathbf{x}(b)$? Let $h > 0$ be a small time step. As the particle moves from $\mathbf{x}(t)$ to $\mathbf{x}(t+h)$, the work $\Delta W(t)$ done is approximately given by the dot product of the displacement of the particle and the force acting time t :

$$\Delta W(t) \approx \mathbf{F}(\mathbf{x}(t)) \cdot (\mathbf{x}(t+h) - \mathbf{x}(t)) .$$

This is not exact since the force \mathbf{F} is not constant, but if the segment is very short, the variation in the force is a small percentage of the force itself. In this same small step limit, there is one more useful approximation to make:

$$\mathbf{F}(\mathbf{x}(t)) \cdot (\mathbf{x}(t+h) - \mathbf{x}(t)) = \mathbf{F}(\mathbf{x}(t)) \cdot \frac{(\mathbf{x}(t+h) - \mathbf{x}(t))}{h} h \approx \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t)h .$$

Thus, if we divide the path into many such small segments, and then add up all of the contributions from all of the segments, and take the limit as the length of the segments tends to zero, we obtain an integral giving the exact value of the work that gets done: This is

$$\int_a^b \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt . \quad (9.39)$$

Such integrals are frequently called *line integrals*.

Example 146 (Computing a line integral). *Let $\mathbf{F}(x, y, z) = (z, x, y)$. Let $\mathbf{x}(t) = (\cos t \sin t, t)$ for $t \in (0, 2\pi)$, so that $\mathbf{x}(t)$ traces out a helix. To compute the corresponding line integral, we work out*

$$\mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) = -t \sin t + \cos^2 t + \sin t .$$

If \mathbf{F} is a force field, and $\mathbf{x}(t)$ is the path of a point particle, the work done on the particle as it moves along the path is

$$\int_0^{2\pi} [-t \sin t + \cos^2 t + \sin t] dt = 3\pi .$$

For computational purposes, it is best to represent the line integral in terms of some explicit parameterization of the curve. But the work done on the particle as it moves along the curve \mathcal{C} ,

with a specified direction of travel, has a well defined meaning that is independent of any particular parameterization.

Therefore, let \mathcal{C} be an smooth oriented closed curve. for present purposes we do not require the curve \mathcal{C} to be simple. That is, we do not require that it be free of self-intersections. Suppose that $\mathbf{x}(t)$ is a continuously differentiable parameterization of \mathcal{C} . Then the preferred unit tangent vector specifying the orientation is

$$\mathbf{T}(t) = \pm \frac{1}{\|\mathbf{x}'(t)\|} \mathbf{x}'(t)$$

where the plus sign is valid if the parameterization is consistent with the orientation, and the minus sign otherwise. The element of arc length along the curve, ds , is given by $ds = \|\mathbf{x}'(t)\|dt$. Therefore,

$$\mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{T}(\mathbf{x}(t)) ds = \pm \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt , \quad (9.40)$$

with the plus sign for a consistent parameterization, and a minus sign otherwise. This gives us the geometric form of the line integral:

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds .$$

Here, if this is to be interpreted as a computing of work done by a force field on a particle as it moves along \mathcal{C} , we regard \mathcal{C} as oriented by the direction of motion so that \mathbf{T} points in this direction. However, one can then use *any* parameterization of \mathcal{C} to compute the work, provided one uses (9.40) to take the sign into account.

Example 147 (Computing another line integral). *The main difference between this example and the previous example, is the that time we will only be given the curve \mathcal{C} , and the direction of motion along it, but not the position at each time t . But still we can calculate the work done by the force field \mathbf{F} on the particle.*

As before, let $\mathbf{F}(x, y, z) = (z, x, y)$. Let \mathcal{C} be the curve that runs along the parabola $y = 1 - x^2$ in the x, y pane from $(1, 0, 0)$ to $(-1, 0, 0)$. We parameterize the path by $\mathbf{x}(t) = (t, 1 - t^2, 0)$, $t \in (-1, 1)$. This traces out \mathcal{C} , but does so backwards, starting at $(-1, 0, 0)$ and ending at $(1, 0, 0)$. Hence, this is certainly not the actual trajectory of the particle parameterized by time. Nonetheless, we can use it to compute the work done on the particle by the force field \mathbf{F} as it moves along its actual trajectory. Taking into account the minus sign required in (9.40) due to the “backwards” parameterization,

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = - \int_{-1}^1 (0, t, 1 - t^2) \cdot (1, -2t, 0) dt = - \int_{-1}^1 2t^2 dt = -\frac{4}{3} .$$

9.3.1 Conservative vector fields

There is a particularly nice kind of line integral: One in which the vector field $\mathbf{F}(\mathbf{x})$ is the gradient of some function $\varphi(\mathbf{x})$. By the chain rule of Chapter 3,

$$\frac{d}{dt} \varphi(\mathbf{x}(t)) = \nabla \varphi(\mathbf{x}(t)) \cdot \mathbf{x}'(t) .$$

Therefore, if \mathcal{C} is the path running along $\mathbf{x}(t)$ for, say, $a \leq t \leq b$, the fundamental Theorem of Calculus gives us

$$\varphi(\mathbf{x}(b)) - \varphi(\mathbf{x}(a)) = \int_a^b \nabla \varphi(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt = \int_{\mathcal{C}} \nabla \varphi \cdot \mathbf{T} ds .$$

Notice that the left hand side depends only on the initial and final points along the curve \mathcal{C} . Therefore

$$\int_{\mathcal{C}} \nabla \varphi \cdot \mathbf{T} ds$$

only depends on the curve \mathcal{C} through its starting point and end point: Let \mathcal{C} be *any* smooth, oriented curve starting at \mathbf{p} and ending at \mathbf{q} . Then

$$\int_{\mathcal{C}} \nabla \varphi \cdot \mathbf{T} ds = \varphi(\mathbf{q}) - \varphi(\mathbf{p}) . \quad (9.41)$$

Definition 103 (Conservative vector field). *A vector field \mathbf{F} on an open, path-wise connected set $U \subset \mathbb{R}^3$ is conservative in case for any pair of points $\mathbf{p}, \mathbf{q} \in U$, and for any two smooth oriented curves $\mathcal{C}_1, \mathcal{C}_2$ starting at \mathbf{p} and ending at \mathbf{q} and staying within U ,*

$$\int_{\mathcal{C}_1} \mathbf{F} \cdot \mathbf{T} ds = \int_{\mathcal{C}_2} \mathbf{F} \cdot \mathbf{T} ds . \quad (9.42)$$

By (9.41), every gradient vector is conservative. We show next that every conservative vector field is a gradient vector field.

Theorem 98. *Let U be an open path-wise connected subset of \mathbb{R}^3 . Then a continuous vector field \mathbf{F} defined on U is conservative if and only if there is a continuously differentiable function φ defined on U such that $\mathbf{F}(\mathbf{x}) = \nabla \varphi(\mathbf{x})$ for all $\mathbf{x} \in U$.*

Proof. We have already seen that all gradient vector fields are conservative. Now let \mathbf{F} be conservative. Pick any base point $\mathbf{x}_0 \in U$. We define a function $\varphi(\mathbf{x})$ on U as follows: Let $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}$ be an smooth curve in U that starts at \mathbf{x}_0 and ends at \mathbf{x} . Such a curve exists since U is path-wise connected. Then we put:

$$\varphi(\mathbf{x}) = \int_{\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}} \mathbf{F} \cdot \mathbf{T} ds . \quad (9.43)$$

The function φ is well-defined because the value of the integral does not depend on the choice of $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}$.

Let $\mathbf{F}(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$. We now prove that

$$\nabla \varphi(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x})) \quad (9.44)$$

for each $\mathbf{x} \in U$. Fix an $\mathbf{x} \in U$. Since U is open, $\mathbf{x} + h\mathbf{e}_1 \in U$ for all sufficiently small h . Pick any smooth oriented curve $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}$ starting at \mathbf{x}_0 and ending at \mathbf{x} . Define $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}+h\mathbf{e}_1}$ to be the continuation of $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}$ to goes from \mathbf{x} to $\mathbf{x} + h\mathbf{e}_1$ on the straight line segment connection these point which is parameterize by $\mathbf{x}(t) = \mathbf{x}_0 + t h\mathbf{e}_1$, $t \in (0, 1)$. Then

$$\varphi(\mathbf{x} + h\mathbf{e}_1) - \varphi(\mathbf{x}) = \int_{\mathcal{C}_{\mathbf{x}_0, \mathbf{x}+h\mathbf{e}_1}} \mathbf{F} \cdot \mathbf{T} ds - \int_{\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}} \mathbf{F} \cdot \mathbf{T} ds = \int_0^1 \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt ,$$

since the integral on the right is the final part of the line integral along $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}+h\mathbf{e}_1}$ that is not on $\mathcal{C}_{\mathbf{x}_0, \mathbf{x}}$.

We now compute $\mathbf{x}'(t) = h\mathbf{e}_1$, and so $\mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) = hP(\mathbf{x}_0 + t h\mathbf{e}_1)$, and thus

$$\frac{\varphi(\mathbf{x} + h\mathbf{e}_1) - \varphi(\mathbf{x})}{h} = \int_0^1 P(\mathbf{x}_0 + t h\mathbf{e}_1) dt .$$

Taking the limit $h \rightarrow 0$ on both sides, we obtain $\frac{\partial \varphi(x, y, z)}{\partial x} = P(x, y, z)$. This proves the equality of the first components on both sides of (9.44). The proof of equality for the other components is very much the same. \square

Definition 104 (Potential function). *Let \mathbf{F} be a conservative vector field defined on an open pathwise connected set U . A continuously differentiable function φ on U such that $\mathbf{F}(\mathbf{x}) = \nabla \varphi(\mathbf{x})$ for all $\mathbf{x} \in U$ is called a potential function for \mathbf{F} . By the previous theorem, every conservative vector field has at least one potential function. Let φ_1 and φ_2 be two potential functions for \mathbf{F} . Then*

$$\nabla(\varphi_1 - \varphi_2)(\mathbf{x}) = \nabla\varphi_1(\mathbf{x}) - \nabla\varphi_2(\mathbf{x}) = \mathbf{F}(x) = \mathbf{F}(\mathbf{x}) = \mathbf{0}.$$

Therefore, $\varphi_1 - \varphi_2$ is constant. Hence the potential function of \mathbf{F} is unique up to an additive constant

It is clear that if \mathcal{C} is a closed curve; i.e, a curve starting and ending at the same \mathbf{p} in \mathbb{R}^3 , then for every conservative vector field \mathbf{F} , $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = 0$ if φ is a potential function for \mathbf{F} ,

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = \varphi(\mathbf{p}) - \varphi(\mathbf{p}) = 0.$$

Conversely, suppose that $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = 0$ for every closed curve. Consider any pair of points \mathbf{p}, \mathbf{q} and any two curves \mathcal{C}_1 and \mathcal{C}_2 from \mathbf{p} to \mathbf{q} . Define a closed curve \mathcal{C} by following \mathcal{C}_1 from \mathbf{p} to \mathbf{q} , and the following \mathcal{C}_2 backwards from \mathbf{q} to \mathbf{p} . Then

$$0 = \int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = \int_{\mathcal{C}_1} \mathbf{F} \cdot \mathbf{T} ds - \int_{\mathcal{C}_2} \mathbf{F} \cdot \mathbf{T} ds,$$

and thus \mathbf{F} is conservative.

Going forward, it will be useful to use a standard notation that emphasizes when a line integral is taken over a closed curve \mathcal{C} : We write

$$\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds$$

for such a line integral, and call it *circulation integral*.

One way to show that a vector field \mathbf{F} is conservative is to find a potential function for it. Another way would be to show that $\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = 0$ for all closed curve \mathcal{C} . This may sound impractical upon first consideration, but there is a simple way to do it, and this brings us to the notion of the *curl* of a vector field \mathbf{F} , as we explain next.

9.3.2 Curl, circulation and Stokes' Theorem

Let $\mathbf{F}(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$ be a vector field on \mathbb{R}^3 . Then \mathbf{F} is a continuously differentiable function from \mathbb{R}^3 to \mathbb{R}^3 , and its derivative is given by the Jacobian matrix

$$\begin{bmatrix} \frac{\partial P}{\partial x} & \frac{\partial P}{\partial y} & \frac{\partial P}{\partial z} \\ \frac{\partial Q}{\partial x} & \frac{\partial Q}{\partial y} & \frac{\partial Q}{\partial z} \\ \frac{\partial R}{\partial x} & \frac{\partial R}{\partial y} & \frac{\partial R}{\partial z} \end{bmatrix}. \quad (9.45)$$

The divergence of \mathbf{F} is the trace of this matrix; i.e., the sum of its diagonal entries. The antisymmetric part of this matrix holds the key to efficient calculation of circulation integrals, as we now explain.

Let M be any $n \times n$ matrix. Then its transpose M^T is also an $n \times n$ matrix, and we can form linear combinations of M and M^T . Note that

$$M = \frac{1}{2} (M - M^T) + \frac{1}{2} (M + M^T) .$$

The matrix $A_M := \frac{1}{2} (M - M^T)$ is called the *antisymmetric part of M* , and the matrix $S_M := \frac{1}{2} (M + M^T)$ *symmetric part of M* . Note that

$$A_M^T = \frac{1}{2} (M - M^T)^T = \frac{1}{2} (MT - M^{TT}) = \frac{1}{2} (M^T - M) = -A_M ,$$

so that A_M is an antisymmetric matrix. A similar calculation shows that S_M is symmetric.

Every antisymmetric 3×3 matrix A has the form

$$A = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix} ,$$

for some number a , b and c , and with this pattern of signs, if we define the vector $\mathbf{a} = (a, b, c)$, applying the matrix A to any vector $\mathbf{x} \in \mathbb{R}^3$ gives the same result as computing the cross product $\mathbf{a} \times \mathbf{x}$. That is, the matrix A is the matrix representative of the linear transformation sending \mathbf{x} to $\mathbf{a} \times \mathbf{x}$. In this spirit, we can call \mathbf{a} the vector representative of the antisymmetric matrix A .

In this terminology, the *curl* of a matrix is the vector \mathbf{F} representative of the antisymmetric part of the Jacobian of \mathbf{F} , multiplied by 2. To write this out, we first note from (9.45) that twice the antisymmetric part of $[D\mathbf{F}]$ is

$$2A_{[D\mathbf{F}]} = \begin{bmatrix} 0 & \frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} & \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x} \\ \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} & 0 & \frac{\partial Q}{\partial z} - \frac{\partial R}{\partial y} \\ \frac{\partial R}{\partial x} - \frac{\partial P}{\partial z} & \frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z} & 0 \end{bmatrix} . \quad (9.46)$$

All of the information contained in this matrix is contained in its vector representative:

$$\left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) . \quad (9.47)$$

Definition 105 (Curl of a vector field on \mathbb{R}^3). *Let $\mathbf{F}(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$ be a continuously differentiable vector field on \mathbb{R}^3 . The curl of \mathbf{F} , $\text{curl}\mathbf{F}$ is the vector field given by (9.47). That is, it is twice the vector representative of the Jacobian matrix $[D\mathbf{F}]$ of \mathbf{F} .*

The next theorem explains the geometric meaning of the curl.

Theorem 99. Let \mathbf{x}_0 be any vector in \mathbb{R}^3 , and let \mathbf{v} and \mathbf{w} be any two linearly independent vectors in \mathbb{R}^3 . Pick a number $h > 0$, and define $\mathcal{C}(h)$ to be the oriented triangle that starts at \mathbf{x}_0 , and goes along the straight line segment from this point to $\mathbf{x}_0 + h\mathbf{v}$, and then along the straight line segment from $\mathbf{x}_0 + h\mathbf{v}$ to $\mathbf{x}_0 + h\mathbf{w}$, and finally along the straight from there back to \mathbf{x}_0 . Let $\mathcal{D}(h)$ be the domain planar domain enclosed by the triangle in the plane through its three vertices. Then for any continuously differentiable vector field \mathbf{F} ,

$$\lim_{h \rightarrow 0} \frac{1}{\text{area}(\mathcal{D}(h))} \int_{\mathcal{C}(h)} \mathbf{F} \cdot \mathbf{T} ds = \text{curl} \mathbf{F}(\mathbf{x}_0) \cdot \frac{\mathbf{v} \times \mathbf{w}}{\|\mathbf{v} \times \mathbf{w}\|}. \quad (9.48)$$

Proof. The first step is to evaluate the denominator, which is easily done using the cross product:

$$\text{area}(\mathcal{D}(h)) = \frac{h^2}{2} \|\mathbf{v} \times \mathbf{w}\|. \quad (9.49)$$

We turn to the numerator, which is more complicated. The next step is to parameterize $\mathcal{C}(h)$. Define $\mathbf{x}(t)$ as follows:

$$\mathbf{x}(t) = \begin{cases} \mathbf{x}_0 + t h \mathbf{v} & 0 \leq t \leq 1 \\ \mathbf{x}_0 + h \mathbf{v} + (t-1)h(\mathbf{w} - \mathbf{v}) & 1 \leq t \leq 2 \\ \mathbf{x}_0 + (3-t)h \mathbf{w} & 2 \leq t \leq 3 \end{cases}$$

$$\oint_{\mathcal{C}(h)} \mathbf{F} \cdot \mathbf{T} ds = \int_0^3 \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt.$$

We now make a first order Taylor approximation to \mathbf{F} :

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0) + [D\mathbf{F}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) + \mathbf{r}(\mathbf{x})\|\mathbf{x} - \mathbf{x}_0\| \quad (9.50)$$

where the remainder term $\mathbf{r}(\mathbf{x})\|\mathbf{x} - \mathbf{x}_0\|$ satisfies $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \mathbf{r}(\mathbf{x}) = 0$. To simplify notation, define $M = [D\mathbf{F}(\mathbf{x}_0)]$ and $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{x}_0$. Then since $\mathbf{z}'(t) = \mathbf{x}'(t)$,

$$\oint_{\mathcal{C}(h)} \mathbf{F} \cdot \mathbf{T} ds = \int_0^3 \mathbf{F}(\mathbf{x}_0) \cdot \mathbf{z}'(t) dt + \int_0^3 \mathbf{z}'(t) \cdot M \mathbf{z}(t) dt + \int_0^3 \|\mathbf{z}(t)\| \mathbf{r}(\mathbf{x}_0 + \mathbf{z}(t)) \cdot \mathbf{z}'(t) dt.$$

Since $\mathbf{F}(\mathbf{x}_0)$ is independent of t ,

$$\int_0^3 \mathbf{F}(\mathbf{x}_0) \cdot \mathbf{z}'(t) dt = \mathbf{F}(\mathbf{x}_0) \cdot \left(\int_0^3 \mathbf{z}'(t) dt \right) = \mathbf{F}(\mathbf{x}_0) \cdot \mathbf{0},$$

and so the constant term makes no contribution.

The remainder term makes no contribution in the limit since for all t both $\|\mathbf{z}(t)\|$ and $\|\mathbf{z}'(t)\|$ are bounded by $\max\{\|\mathbf{v}\|, \|\mathbf{w}\|\}h$ and so by The Cauchy-Schwarz inequality,

$$\left| \int_0^3 \|\mathbf{z}(t)\| \mathbf{r}(\mathbf{x}_0 + \mathbf{z}(t)) \cdot \mathbf{z}'(t) dt \right| \leq h^2 (\max\{\|\mathbf{v}\|, \|\mathbf{w}\|\})^2 \int_0^3 \|\mathbf{r}(\mathbf{x}_0 + \mathbf{z}(t))\| dt.$$

Therefore, using (9.49),

$$\frac{1}{\text{area}(\mathcal{D}(h))} \left| \int_0^3 \|\mathbf{z}(t)\| \mathbf{r}(\mathbf{x}_0 + \mathbf{z}(t)) \cdot \mathbf{z}'(t) dt \right| \leq \frac{2(\max\{\|\mathbf{v}\|, \|\mathbf{w}\|\})^2}{\|\mathbf{v} \times \mathbf{w}\|} \int_0^3 \|\mathbf{r}(\mathbf{x}_0 + \mathbf{z}(t))\| dt,$$

and since $\mathbf{z}(t)$ converges to \mathbf{x}_0 as h converges to zero, $\lim_{h \rightarrow 0} \int_0^3 \|\mathbf{r}(\mathbf{x}_0 + \mathbf{z}(t))\| dt = 0$.

We are left with evaluating $\int_0^3 \cdot \mathbf{z}'(t) \cdot M\mathbf{z}(t) dt$. Notice that

$$\frac{d}{dt}(\mathbf{z}(t) \cdot M\mathbf{z}(t)) = \mathbf{z}'(t) \cdot M\mathbf{z}(t) + \mathbf{z}(t) \cdot M\mathbf{z}'(t) = 2\mathbf{z}'(t) \cdot S_M\mathbf{z}(t) ,$$

where S_M is the symmetric part of M , so that

$$\int_0^3 \mathbf{z}'(t) \cdot S_M\mathbf{z}(t) dt = 2 \int_0^3 \frac{d}{dt}(\mathbf{z}(t) \cdot M\mathbf{z}(t)) dt = 2(\mathbf{z}(3) \cdot M\mathbf{z}(3) - \mathbf{z}(0) \cdot M\mathbf{z}(0)) = 0 .$$

Therefore, the symmetric part of M makes no contribution to our integral, and

$$\oint_{C(h)} \mathbf{F} \cdot \mathbf{T} ds = \int_0^3 \mathbf{z}'(t) \cdot A_M\mathbf{z}(t) dt + \text{negligible remainder} . \quad (9.51)$$

Now let \mathbf{a} be the vector representative of the antisymmetric matrix A_M . Then the integrand on the right in (9.51) is

$$\mathbf{z}'(t) \cdot \mathbf{a} \times \mathbf{z}(t) = \mathbf{a} \times \mathbf{z}(t) \cdot \mathbf{z}'(t) = \mathbf{a} \cdot \mathbf{z}(t) \times \mathbf{z}'(t) .$$

Notice that for $t \in (0, 1)$ and $t \in (2, 3)$, $\mathbf{z}(t)$ and $\mathbf{z}'(t)$ are both proportional to one another, and so $\mathbf{z}(t) \times \mathbf{z}'(t) = 0$ except when $t \in (1, 2)$. For $t \in (1, 2)$,

$$\mathbf{z}(t) \times \mathbf{z}'(t) = h^2(\mathbf{v} + (t-1)(\mathbf{w} - \mathbf{v})) \times (\mathbf{w} - \mathbf{v}) = h^2\mathbf{v} \times \mathbf{w} .$$

$$\int_0^3 \mathbf{z}'(t) \cdot A_M\mathbf{z}(t) dt = \mathbf{a} \cdot \int_1^2 h^2\mathbf{v} \times \mathbf{w} dt = h^2\mathbf{a} \cdot \mathbf{v} \times \mathbf{w} . \quad (9.52)$$

Using (9.49) once more we have

$$\lim_{h \rightarrow 0} \frac{1}{\text{area}(\mathcal{D}(h))} \int_{C(h)} \mathbf{F} \cdot \mathbf{T} ds = \frac{2\mathbf{a} \cdot \mathbf{v} \times \mathbf{w}}{\|\mathbf{v} \times \mathbf{w}\|} ,$$

which is the same as (9.48). \square

Now suppose that \mathbf{F} is a conservative vector field on \mathbb{R}^3 . Then if \mathcal{C} runs around any triangle, $\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = 0$. The by considering triangles though a point \mathbf{x}_0 with a given normal \mathbf{N} , which can be any unit vector, the previous theorem says that $\text{curl } \mathbf{F}(\mathbf{x}_0) \cdot \mathbf{N} = 0$. Since this is true for all unit vectors \mathbf{N} , and all \mathbf{x}_0 , $\text{curl } \mathbf{F} = \mathbf{0}$.

This can be seen another way: \mathbf{F} is conservative if and only if $\mathbf{F} = \nabla\varphi$ for some potential function φ . Suppose that φ is twice continuously differentiable, so that $\text{curl } \nabla\varphi$ is well-defined and continuous. Using the formula (9.47), we compute

$$\text{curl } \nabla\varphi = \left(\frac{\partial^2\varphi}{\partial y\partial z} - \frac{\partial^2\varphi}{\partial z\partial y}, \frac{\partial^2\varphi}{\partial z\partial x} - \frac{\partial^2\varphi}{\partial x\partial z}, \frac{\partial^2\varphi}{\partial x\partial y} - \frac{\partial^2\varphi}{\partial y\partial x} \right) .$$

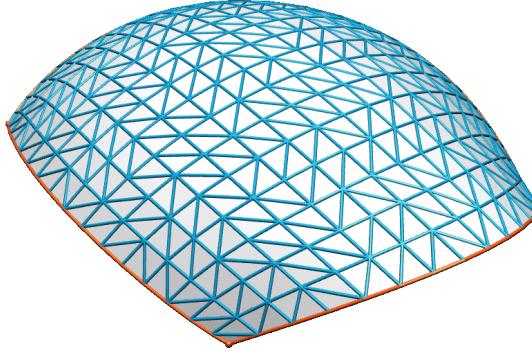
Each entry on the right is zero by Clairault's Theorem. Hence, a continuously differentiable conservative vector field \mathbf{F} satisfies $\text{curl } \mathbf{F} = \mathbf{0}$. The converse is also true. The proof of this, and much else, rests on the next theorem.

Theorem 100. Let \mathcal{S} be an smooth oriented surface in \mathbb{R}^3 bounded by a smooth simple curve \mathcal{C} which is necessarily closed. Give the \mathcal{C} the orientation induced by that of \mathcal{S} : We choose the unit tangent vector \mathbf{T} to \mathcal{C} so that at each point of \mathcal{C} , $\mathbf{T} \times \mathbf{N}$ points outward, away from \mathcal{S} .

Let \mathbf{F} be a twice continuously differentiable vector field defined on a neighborhood of \mathcal{S} . Then

$$\int_{\mathcal{S}} \operatorname{curl} \mathbf{F} dS = \oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds .$$

Proof. Subdivide the surfaces \mathcal{S} into many small approximately triangular tiles: This is a “triangulation” of \mathcal{S} . The triangles with an edge on the boundary inherit an orientation from \mathcal{C} . Each edge in the triangulation that is not on the boundary \mathcal{C} is internal and is part of the boundary to two neighboring triangles. Here is a picture of a triangulation:



If we give each triangle tile the orientation it inherits from \mathcal{S} , each of the internal edges is crossed twice, and in opposite directions. Therefore,

$$\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = \sum_{\text{tiles}} \oint_{\text{boundary of tile}} \mathbf{F} \cdot \mathbf{T} ds ,$$

since all of the integrations along internal edges cancel out in pairs, leaving only the edges along \mathcal{C} .

By the previous theorem, for any small tile with a vector at \mathbf{x}_0 , and with \mathbf{N} denoting the normal to \mathcal{S} at \mathbf{x}_0 ,

$$\oint_{\text{boundary of tile}} \mathbf{F} \cdot \mathbf{T} ds \approx (\operatorname{curl} \mathbf{F}(\mathbf{x}_0) \cdot \mathbf{N}) (\text{area of the tile}) ,$$

and the error in this approximation goes to zero percentage-wise as the tile diameter goes to zero.

However,

$$\sum_{\text{tiles}} (\operatorname{curl} \mathbf{F}(\mathbf{x}_0) \cdot \mathbf{N}) (\text{area of the tile})$$

is a Riemann sum for $\int_{\mathcal{S}} \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS$. Therefore, taking the diameter of the tiles to zero, we obtain the stated equality. \square

Example 148 (Verifying Stokes' Theorem in an example). Let \mathcal{C} be the unit circle in the plane $x + y + z = 1$ that is centered at $(0, 0, 1)$. Let \mathcal{S} be the disk in this plane that is bounded by \mathcal{C} . At each point of \mathcal{S} there are two unit normal vectors, $\pm 3^{-1/2}(1, 1, 1)$. Orient \mathcal{S} by choosing \mathbf{N} to point upward; i.e.,

$$\mathbf{N} = \frac{1}{\sqrt{3}}(1, 1, 1) .$$

We give \mathcal{C} the induced orientation. Let $\mathbf{F}(x, y, z) = (xy, 1, xy)$.

We will compute both $\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds$ and $\int_S \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS$. Starting with the line integral, the first step is to parameterize \mathcal{C} . First we seek any parameterization. Later, we will adjust it if necessary to make it consistent with the orientation.

Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be an right handed orthonormal basis of \mathbb{R}^3 with $\mathbf{u}_3 = \mathbf{N}$. Define

$$\mathbf{x}(t) = (0, 0, 1) + \cos t \mathbf{u}_1 + \sin t \mathbf{u}_2, \quad t \in (0, 2\pi).$$

Then $\mathbf{x}'(t) = -\sin t \mathbf{u}_1 + \cos t \mathbf{u}_2$, which is a unit vector. The unit tangent at $\mathbf{x}(t)$ is therefore $\pm(-\sin t \mathbf{u}_1 + \cos t \mathbf{u}_2)$. Since the normal vector in $\mathbf{N} = \mathbf{u}_3$, the condition specifying the orientation is that

$$\pm(-\sin t \mathbf{u}_1 + \cos t \mathbf{u}_2) \times \mathbf{u}_3 = \pm(\cos t \mathbf{u}_1 \sin t \mathbf{u}_2)$$

points outward from S . (In doing the calculation, we have used the fact that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right-handed orthonormal basis.) Choosing the plus sign, this vector points to the outside of the disk. Hence, our parameterization is consistent with the orientation. (Had we chosen a left handed orthonormal basis, this would not be the case, but then we could correct it by replacing $\mathbf{x}(t)$ with $\mathbf{x}(2\pi - t)$, which traces out the same curve backwards, thus reversing the direction motion.)

For explicit computation, we now choose such a basis. Notice that $2^{-1/2}(1, -1, 0)$ is a unit vector that is orthogonal to \mathbf{u}_3 . Hence we get the basis we seek by choosing $\mathbf{u}_1 = 2^{-1/2}(1, -1, 0)$ and then

$$\mathbf{u}_2 = \mathbf{u}_3 \times \mathbf{u}_3 = 6^{-1/2}(1, 1, -2).$$

Therefore

$$\mathbf{x}(t) = (2^{-1/2} \cos t + 6^{-1/2} \sin t, -2^{-1/2} \cos t + 6^{-1/2} \sin t, 1 - (2/3)^{1/2} \sin t).$$

Evaluating,

$$\mathbf{F}(\mathbf{x}(t)) = \frac{1}{6}(\sin^2 t - 3 \cos^2 t, 6, \sin^2 t - 3 \cos^2 t).$$

Next,

$$\mathbf{x}'(t) = (-2^{-1/2} \sin t + 6^{-1/2} \cos t, 2^{-1/2} \sin t + 6^{-1/2} \cos t, (2/3)^{1/2} \cos t).$$

Since our parameterization is consistent with the orientation,

$$\begin{aligned} \mathbf{F} \cdot \mathbf{T} ds &= \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt = (\sin^2 t - 3 \cos^2 t)(-2^{-1/2} \sin t + 6^{-1/2} \cos t) dt \\ &\quad + (2^{-1/2} \sin t + 6^{-1/2} \cos t) dt \\ &\quad + (\sin^2 t - 3 \cos^2 t)(2/3)^{1/2} \cos t dt. \end{aligned}$$

Then since $\int_0^{2\pi} \sin^m t \cos^n t dt = 0$ for any non-negative integers m and n with $m + n$ odd, we see that

$$\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = \int_0^{2\pi} \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt = 0.$$

Next we compute $\int_S \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS$. We first compute

$$\operatorname{curl} \mathbf{F} = (x, -y, -x).$$

Since $\mathbf{N} = 3^{-1/2}(1, 1, 1)$ everywhere on \mathcal{S} , $\mathbf{F} \cdot \mathbf{N} = -y$ everywhere on \mathcal{S} . Therefore,

$$\int_{\mathcal{S}} \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS = - \int_{\mathcal{S}} y dS .$$

Since the surface \mathcal{S} is symmetric under the transformation sending y to $-y$, $\int_{\mathcal{S}} y dS = 0$. Therefore,

$$\int_{\mathcal{S}} \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS = 0 .$$

Example 149. Let \mathcal{C} be the curve that runs from $(1, 0, 0)$ to $(0, 1, 0)$, and from there to $(0, 0, 1)$, and from there back to $(1, 0, 0)$. Let $\mathbf{F} = (xy, 1, xy)$. Compute the total circulation

$$\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds .$$

When asked to compute a circulation integral, or even a line integral, unless the answer is obvious on grounds of symmetry, say, the first step is to compute the curl of the vector field. From the previous example we have

$$\operatorname{curl}(\mathbf{F}) = (x, -y, -z) .$$

This is quite simple, so it will be good to use Stokes' Theorem. This is especially direct since \mathcal{C} bounds a triangular surface \mathcal{S} in the plane passing through the vertices $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$

The triangle \mathcal{S} lies in the plane given by $x + y + z = 1$, and for this plane the unit normal is

$$\mathbf{N} = \pm \frac{1}{\sqrt{3}}(1, 1, 1) .$$

Therefore, $\operatorname{curl}(\mathbf{F}) \cdot \mathbf{N} = -y$, and so

$$\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = -y \int_{\mathcal{S}} dS .$$

To compute this integral, we parameterize \mathcal{S} . The projection of \mathcal{S} onto the x, y plane is the triangle with vertices $((1, 0), (0, 1)$ and $(0, 0)$. This is the domain

$$\{(x, y) : 0 \leq x \leq 1 - y \text{ and } 0 \leq y \leq 1\} .$$

The equation for the plane containing the triangle is $z = 1 - x - y$, and using this to eliminate z , we obtain

$$\mathbf{X}(x, y) = (x, y, 1 - x - y) , \quad 0 \leq x \leq 1 - y \text{ and } 0 \leq y \leq 1 .$$

Then $\mathbf{X}_x(x, y) = (1, 0, -1)$, $\mathbf{X}_y = (0, 1, -1)$, and $\mathbf{X}_x \times \mathbf{X}_y(x, y) = (1, 1, 1)$. Hence $dS = \sqrt{3} dx dy$ and

$$-\int_{\mathcal{S}} y dS = \int_0^1 \left(\int_0^{1-y} y dx \right) dy = \int_0^1 (y - y^2) dy = \frac{1}{6} .$$

Stokes' Theorem can also be used to efficiently evaluate line integrals along complicated curves \mathcal{C} that are not closed by finding a simple curve \mathcal{C}' such that \mathcal{C} followed by \mathcal{C}' is closed.

Example 150. Let \mathcal{C} be the contour that runs from $(0, 0, 0)$ to $(1, 0, 0)$, and from there to $(1, 0, 1)$, and from there to $(0, 0, 1)$. Let $\mathbf{F} = (x, x, z)$. We compute the line integral

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds .$$

To do this efficiently, let \mathcal{C}' be the segment from $(0, 0, 1)$ to $(0, 0, 0)$. Let $\mathcal{C} + \mathcal{C}'$ denote \mathcal{C} followed by \mathcal{C}' . This is a closed curve, enclosing a square in the x, z plane. The unit normal to this surface is $\pm(0, -1, 0)$, which is orthogonal to $\text{curl}(\mathbf{F}) = (0, 0, 1)$.

Hence

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds + \int_{\mathcal{C}'} \mathbf{F} \cdot \mathbf{T} ds = \int_{\mathcal{C} + \mathcal{C}'} \mathbf{F} \cdot \mathbf{T} ds = 0 ,$$

and so

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = - \int_{\mathcal{C}'} \mathbf{F} \cdot \mathbf{T} ds .$$

We parameterize \mathcal{C}' by $\mathbf{x}(t) = (0, 0, 1-t)$ for $0 \leq t \leq 1$. Then

$$\int_{\mathcal{C}'} \mathbf{F} \cdot \mathbf{T} ds = \int_0^1 (1-t)(-1) dt = -\frac{1}{2} \quad \text{and hence} \quad \int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = \frac{1}{2} .$$

Example 151. Let \mathcal{C} be the curve that is the intersection of the sphere $x^2 + y^2 + z^2 = 4$ and the plane $x + y + z = 1$, oriented so that it runs counterclockwise when viewed from above. Let $\mathbf{F}(x, y, z) = (xy, 1, xz)$. Let us compute $\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds$.

We first compute that $\text{curl} \mathbf{F} = (0, -z, -x)$. This is fairly simple. Also note that \mathcal{C} is the boundary of \mathcal{S} , where \mathcal{S} is the part of the plane $x + y + z = 1$ inside the sphere $x^2 + y^2 + z^2 = 4$, and if we choose the upward unit normal on \mathcal{S} , the orientations of \mathcal{S} and \mathcal{C} are consistent, and by Stokes' Theorem,

$$\oint_{\mathcal{C}} \mathbf{F} \cdot \mathbf{T} ds = \int_{\mathcal{S}} \text{curl} \mathbf{F} \cdot \mathbf{N} dS .$$

Since $\mathbf{N} = 3^{-1/2}(1, 1, 1)$ everywhere on \mathcal{S} , $\text{curl} \mathbf{F} \cdot \mathbf{N} = -3^{-1/2}(x + z)$ everywhere on \mathcal{S} , and since $x + z = 1 - y$ everywhere on \mathcal{S} and we are left with computing

$$-\frac{1}{\sqrt{3}} \int_{\mathcal{S}} (1-y) dS .$$

The surface \mathcal{S} is a disk in the plane $x + y + z = 1$. The center of the disk is at $\frac{1}{3}(1, 1, 1)$. To find radius of the disk, note that on the intersection of this plane and the sphere $x^2 + y^2 + z^2 = 4$, we have

$$x^2 + y^2 + z^2 - \frac{2}{3}(x + y + z) = 4 - \frac{2}{3} = \frac{10}{3} .$$

Completing squares,

$$\left(x - \frac{1}{3}\right)^2 + \left(y - \frac{1}{3}\right)^2 + \left(z - \frac{1}{3}\right)^2 = \frac{11}{3} .$$

Hence \mathcal{S} is the disk in the plane $x + y + z = 1$ with radius $\sqrt{11/3}$ and center $(1/3, 1/3, 1/3)$. Our integrand $1 - y$ can be written as $2/3 - (y - 1/3)$. and by symmetry

$$\int_{\mathcal{S}} (y - 1/3) dS = 0 .$$

Therefore,

$$-\frac{1}{\sqrt{3}} \int_S (1-y) dS = -\frac{2}{3\sqrt{3}} \int_S 1 dS .$$

This last integral is the area of S , which by elementary geometry is $11\pi/3$. Finally we have

$$\oint_C \mathbf{F} \cdot \mathbf{T} ds = -\frac{22\pi}{9\sqrt{3}} .$$

The direct parameterization leads to calculations that are more cumbersome.

9.3.3 Curl and conservative vector fields

Let \mathcal{C} be any simple closed curve in \mathbb{R}^3 . Let U be an open set in \mathbb{R}^3 containing \mathcal{C} . Suppose that there exists a smooth simple surface S such that \mathcal{C} is the boundary of S .

If \mathbf{F} is any vector field with $\operatorname{curl} \mathbf{F} = \mathbf{0}$, and we give S an orientation that is consistent with that of \mathcal{C} , we have

$$0 = \int_S \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS = \int_C \mathbf{F} \cdot \mathbf{T} ds .$$

Definition 106 (Simply connected domain). An open set $U \subset \mathbb{R}^3$ is simply connected in case whenever \mathcal{C} is any smooth simply closed curve in U , there is a smooth oriented surface S in U such that \mathcal{C} is the boundary of S .

The set \mathbb{R}^3 itself is simply connected. This is not an entirely simple matter. Consider a smooth simple closed curve \mathcal{C} in \mathbb{R}^3 , and suppose that \mathcal{C} is the image of $\mathbf{x}(t)$ for $t \in [0, b]$ so that $\mathbf{x}(0) = \mathbf{x}(b)$. Let $\mathbf{p} = \mathbf{x}(0)$, and let $\mathbf{q} = \mathbf{x}(b/2)$. Then $\mathbf{x}(t)$, $t \in (0, b)$, and $\mathbf{x}(b-t)$, $t \in (0, b/2)$ are two curves running from \mathbf{p} to \mathbf{q} along the two “halves” of \mathcal{C} . We construct a surface by connecting $\mathbf{x}(t)$ and $\mathbf{x}(b-t)$ with a straight line segment for each t , connecting up the two halves of \mathcal{C} . This straight line segment is parameterized by $(1-v)\mathbf{x}(t) + v\mathbf{x}(b-t)$. Replacing t by u , we have

$$\mathbf{X}(u, v) = (1-v)\mathbf{x}(u) + c\mathbf{x}(b-u) , \quad (u, v) \in (0, b) \times (0, 1) .$$

This gives us what we seek provided that

$$\mathbf{X}_u \times \mathbf{X}_v(u, v) \neq \mathbf{0}$$

at any $(u, v) \in (0, b) \times (0, 1)$. If this is the case, we can define

$$\mathbf{N}(u, v) = \frac{1}{\|\mathbf{X}_u \times \mathbf{X}_v(u, v)\|} \mathbf{X}_u \times \mathbf{X}_v(u, v) ,$$

which is continuous since $\mathbf{x}(t)$ is continuously differentiable, and so this gives us an oriented surface whose boundary is clearly \mathcal{C} . The subtlety is that $\mathbf{X}_u \times \mathbf{X}_v(u, v) = \mathbf{0}$ may be true for some parameter values. But one can try something other than a straight line segment, or a different division of \mathcal{C} into two halves. Some small twist of the procedure will work.

It is interesting to apply this procedure when \mathcal{C} is the boundary of the Möbius band. Since the Möbius band is not orientable, we might hope that it will yield be another surface S whose boundary is also \mathcal{C} , but which is orientable. This is the case.

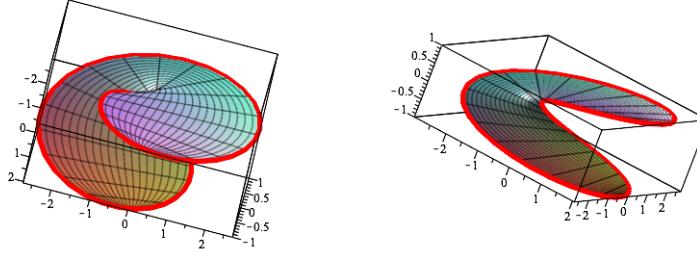
Example 152 (Orientable surface woth the same boundary as the Möbius band). Recall that the boundary of the Möbius band is the curve $\mathbf{x}(t)$ parameterized by

$$\mathbf{x}(u) = (\cos u(2 + \sin(u/2)), \sin u(2 + \sin(u/2)), \cos(u/2)) , \quad u \in [0, 4\pi] . \quad (9.53)$$

We get the surface we seek by forming the straight line segment between $\mathbf{x}(u)$ and $\mathbf{x}(4\pi - u)$ for $u \in [0, 2\pi]$, as described above. Hence we define

$$\mathbf{X}(u, v) = (1 - v)\mathbf{x}(u) + v\mathbf{x}(4\pi - u) , \quad ((u, v) \in [0, 2\pi] \times [0, 1]) .$$

The image of this function in \mathbb{R}^3 is the surface we seek. Clearly, the boundary consists of the points of the form $\mathbf{X}(u, 0)$ and $\mathbf{X}(u, 1)$ for $u \in [0, 2\pi]$, which are the two parts of \mathcal{C} correpsonding to $u \in [0, 2\pi]$ and $u \in [2\pi, 4\pi]$ in (9.53). Here is a plot showing the surface and its bounding curve from two pespectives:



Given a simple smooth closed curve \mathcal{C} in \mathbb{R}^3 , we can find an oriented surface \mathcal{S} that has \mathcal{C} as its boundary, but \mathcal{S} may have self intersections. However, we do not require that \mathcal{S} be free of self intersections to apply Stokes' Theorem; we only require that \mathcal{S} can be oriented. Then when we triagulate \mathcal{S} , we can apply Stokes' theorem all to of the triangles in the triangulation.

The next example concerns a more complicated curve in \mathbb{R}^3 , the *trefoil knot*. It is still a simple closed curve however, since it does not intersect itself.

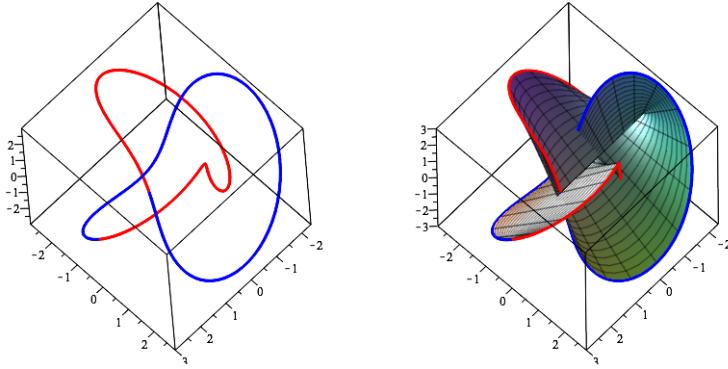
Example 153 (An oriented surface spanning the trefoil knot). The trefoil knot may be parameterized by

$$\mathbf{x}(t) = ((2 + \cos(3t/2)) \cos t, (2 + \cos(3t/2)) \sin t, 3 \sin(3t/2)) , \quad t \in [0, 4\pi] .$$

Our standard parameterization procedure is to join the two curves $\mathbf{x}(u)$ and $\mathbf{x}(4\pi - u)$ for $u \in (0, 2\pi)$ be straight line segments running from $\mathbf{x}(u)$ and $\mathbf{x}(4\pi - u)$. We define

$$\mathbf{X}(u, v) = (1 - v)\mathbf{x}(u) + v\mathbf{x}(4\pi - u) , \quad (u, v) \in (0, 2\pi) \times (0, 1) .$$

The next plots show the trefoil knot itself, and the spanning surface that we have just specified.



The part of the knot in red is the first part corresponding to $t \in (0, 2\pi)$, and the part in blue is the second part, corresponding to $t \in (2\pi, 4\pi)$. The lines on the surface are lines of constant u and v . They divide the surface into “tiles” that can be further subdivided into triangles.

The surface has self intersections, and at the self intersections there is no single unit normal vector \mathbf{N} . However, it is evident that the parameterization is continuously differentiable at the points of intersection, and the self intersections are one dimensional, and do not account for any surface area at all. Therefore we can ignore these points, and for any continuously differentiable vector field \mathbf{F} , we can and still compute $\int_S \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS$, and Stokes’ Theorem is still valid. In particular, if $\operatorname{curl} \mathbf{F} = \mathbf{0}$ everywhere, then

$$\int_C \mathbf{F} \cdot \mathbf{T} ds = \int_S \operatorname{curl} \mathbf{F} \cdot \mathbf{N} dS = 0 .$$

For an example of an open set in \mathbb{R}^3 that is not simply connected, let U be \mathbb{R}^3 with the z -axis removed. Let \mathcal{C} be the unit circle in the x, y plane, which lies in U . There is no surface S in U having \mathcal{C} as its boundary since any surface having \mathcal{C} as its boundary must intersect the z -axis somewhere.

Because U is not simply connected, there exist continuously differentiable vector fields \mathbf{F} defined on U such that $\operatorname{curl} \mathbf{F} = \mathbf{0}$ everywhere on U , but such that \mathbf{F} is not conservative, and not a gradient vector field.

Example 154 (A vector field \mathbf{F} with $\operatorname{curl} \mathbf{F} = \mathbf{0}$ that is not conservative). Let

$$\mathbf{F}(x, y, z) = \frac{1}{x^2 + y^2} (-y, x, 0) .$$

Then

$$\begin{aligned} \operatorname{curl} \mathbf{F}(x, y, z) &= \left(0, 0, \frac{\partial}{\partial x} \frac{x}{x^2 + y^2} + \frac{\partial}{\partial y} \frac{y}{x^2 + y^2} \right) \\ &= \left(0, 0, \frac{y^2 - x^2}{(x^2 + y^2)^2} + \frac{x^2 - y^2}{(x^2 + y^2)^2} \right) = \mathbf{0} . \end{aligned}$$

Now, if \mathcal{C} is the circle of radius $r > 0$ in the x, y plane with its usual orientation, we may parameterize it consistently by $\mathbf{x}(t) = r(\cos t, \sin t, 0)$. Then $\mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) = (-\sin t, \cos t) \cdot (-\sin t, \cos t) = 1$, and hence

$$\oint_C \mathbf{F} \cdot \mathbf{T} ds = \int_0^{2\pi} \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt = 2\pi ,$$

independent of $r > 0$.

This is not zero, despite the fact that \mathbf{F} is continuously differentiable everywhere on U and $\operatorname{curl}\mathbf{F} = \mathbf{0}$ everywhere on U . Stokes' Theorem is not violated because there is no surface (oriented or not) in U that has C as its boundary. Any surface with C as its boundary must intersect the line along which \mathbf{F} is singular. Even though it may do this in a single point, the fact that $\|\mathbf{F}\|$ is arbitrarily large in a neighborhood of this single point allows the single points to make a difference. The hypothesis in Theorem 99 that \mathbf{F} is continuously differentiable on the triangle is essential.

If an open set $U \subset \mathbb{R}^3$ is simply connected, and if \mathbf{F} is a continuously differentiable vector field on U such that $\operatorname{curl}\mathbf{F} = \mathbf{0}$ everywhere on U , then \mathbf{F} is conservative on U . To see this, let C be any smooth, simple oriented curve in U , and let S be a smooth simple surface whose boundary is C . Give S the orientation that is consistent with that of C . Then by Stokes' Theorem

$$\int_C \mathbf{F} \cdot \mathbf{T} ds = \int_S \operatorname{curl}\mathbf{F} \cdot \mathbf{N} dS = 0 .$$

Therefore, \mathbf{F} is conservative.

That is, for vector fields \mathbf{F} on a simply connected set $U \subset \mathbb{R}^3$, we can test whether \mathbf{F} is conservative or not by computing $\operatorname{curl}\mathbf{F}$: The vector field \mathbf{F} on U is conservative if and only if $\operatorname{curl}\mathbf{F} = \mathbf{0}$. It is not hard to see that \mathbb{R}^3 itself is simply connected. Therefore, if we want to know whether a vector field \mathbf{F} defined on all of \mathbb{R}^3 is a conservative vector field, and the gradient of some potential function φ , we first compute the curl of \mathbf{F} . If this is zero, then \mathbf{F} is a gradient vector field, and otherwise it is not.

Example 155. Consider the two vector fields

$$\mathbf{F} = (y + z^2, x + z^2, 2zx + 2zy) \quad \text{and} \quad \mathbf{G} = (y + z^2, x + z^2, 2x + 2y) .$$

One of the vector fields \mathbf{F} and \mathbf{G} is equal to $\nabla\varphi$ for some potential function φ . Which one is it? Find such a potential function for the conservative vector field.

To do this, we first compute the curls. We find

$$\operatorname{curl}(\mathbf{F}) = \mathbf{0} \quad \text{and} \quad \operatorname{curl}(\mathbf{G}) = (2 - 2z, 2z - 2, 0) .$$

A vector field on \mathbb{R}^3 is a gradient if and only if its curl is zero at every point in \mathbb{R}^3 . Hence \mathbf{F} is a gradient. To find the potential function, pick $\mathbf{0}$ as a base point. For any $\mathbf{x} \in \mathbb{R}^3$, define

$$\mathbf{x}(t) = t\mathbf{x} , \quad t \in (0, 1) .$$

Then any potential function φ satisfies

$$\varphi(\mathbf{x}) = \varphi(\mathbf{0}) + \int_{C_{\mathbf{0}, \mathbf{x}}} \mathbf{F} \cdot \mathbf{T} ds .$$

Then

$$\mathbf{F}(\mathbf{x}(t)) = (ty + t^2z^2, tx + t^2z^2, t^2(2zx + 2zy))$$

and $\mathbf{x}'(t) = (x, y, z)$. Hence

$$\begin{aligned}\varphi(\mathbf{x}) - \varphi(\mathbf{0}) &= \int_0^1 [(ty + t^2z^2)x + (tx + t^2z^2)y + t^2(2zx + 2zy)z] dt \\ &= \frac{1}{6}[(3y + 4z^2)x + (3x + 4z^2)y + (2xz + 2xy)z \\ &= xy + z^2(x + y) .\end{aligned}$$

Since an arbitrary constant may be subtracted from φ , we may set $\varphi(\mathbf{0}) = 0$, and then we have

$$\varphi(x, y, z) = xy + z^2(x + y) .$$

It is now easy to check that $\nabla\varphi = \mathbf{F}$.

9.3.4 Vector potentials

Theorem 101. Let $\mathbf{F}(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$ be a twice continuously differentiable vector field. Then

$$\operatorname{div}(\operatorname{curl}(\mathbf{F}(\mathbf{x}))) = 0 .$$

Proof.

$$\begin{aligned}\operatorname{div}(\operatorname{curl}\mathbf{F}(\mathbf{x})) &= \operatorname{div}\left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) \\ &+ \frac{\partial}{\partial x}\left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) + \frac{\partial}{\partial y}\left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) + \frac{\partial}{\partial z}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) \\ &= \left(\frac{\partial^2 P}{\partial y \partial z} - \frac{\partial^2 P}{\partial z \partial y}\right) + \left(\frac{\partial^2 Q}{\partial z \partial x} - \frac{\partial^2 Q}{\partial x \partial z}\right) + \left(\frac{\partial^2 R}{\partial x \partial y} - \frac{\partial^2 R}{\partial y \partial x}\right) = 0 ,\end{aligned}$$

where the final equality is valid on account of Clairault's Theorem. \square

Vector fields \mathbf{G} that satisfy $\operatorname{div}(\mathbf{G}(\mathbf{x})) = 0$ are called *divergence free vector fields*. Theorem 101 says that every vector field that is a curl is divergence free. There is an important converse: If \mathbf{G} is divergence free on a simply connected domain $U \subset \mathbb{R}^3$, such as \mathbb{R}^3 itself, then there is a twice continuously differentiable vector field $\mathbf{A}(x)$ on U such that $\mathbf{G}(\mathbf{x}) = \operatorname{curl}\mathbf{A}(\mathbf{x})$ for all $\mathbf{x} \in U$.

The first thing to notice is that if \mathbf{G} is divergence free in U and $\mathbf{G}(\mathbf{x}) = \operatorname{curl}\mathbf{A}(\mathbf{x})$ for some twice-continuously differentiable vector field \mathbf{A} , then we also have

$$\mathbf{G}(\mathbf{x}) = \operatorname{curl}(\mathbf{A}(\mathbf{x}) + \nabla\varphi(\mathbf{x}))$$

for any twice-continuously differentiable function φ on U . Hence if there is one such vector field \mathbf{A} , there are infinitely many others. We might hope, therefore, that it is possible to make a simple choice for \mathbf{A} . This turns out to be the case.

Let $\mathbf{G} = (P, Q, R)$. Consider a twice-continuously differentiable vector field $\mathbf{A}(\mathbf{x}) = (F(\mathbf{x}), 0, H(\mathbf{x}))$. Then

$$\operatorname{curl}\mathbf{A}(\mathbf{x}) = \left(\frac{\partial H}{\partial y}(\mathbf{x}), \frac{\partial F}{\partial z}(\mathbf{x}) - \frac{\partial H}{\partial x}(\mathbf{x}), -\frac{\partial F}{\partial y}(\mathbf{x})\right) .$$

Then $\mathbf{G}(\mathbf{x}) = \operatorname{curl}\mathbf{A}(\mathbf{x})$ becomes

$$P(\mathbf{x}) = \frac{\partial H}{\partial y}(\mathbf{x}), \quad Q(\mathbf{x}) = \frac{\partial F}{\partial z}(\mathbf{x}) - \frac{\partial H}{\partial x}(\mathbf{x}) \quad \text{and} \quad R(\mathbf{x}) = -\frac{\partial F}{\partial y}(\mathbf{x}).$$

Using the first and third of the equations and the Fundamental Theorem of Calculus, we conclude (in the case $U = \mathbb{R}^3$) that

$$\begin{aligned} H(x, y, z) &= \int_0^y P(x, t, z) dt + \alpha(x, z) \\ F(x, y, z) &= - \int_0^y R(x, t, z) dt + \beta(x, z) \end{aligned}$$

for twice-continuously differentiable functions $\alpha(x, z)$ and $\beta(x, z)$. We then compute the middle term in $\operatorname{curl}\mathbf{A}$:

$$\frac{\partial F}{\partial z}(\mathbf{x}) - \frac{\partial H}{\partial x}(\mathbf{x}) = - \int_0^y \left[\frac{\partial R}{\partial z}(x, t, z) + \frac{\partial P}{\partial x}(x, t, z) \right] dt + \frac{\partial \alpha}{\partial z}(x, z) - \frac{\partial \beta}{\partial x}(x, z).$$

However, since $\operatorname{div}\mathbf{G} = 0$,

$$\frac{\partial R}{\partial z}(x, t, z) + \frac{\partial P}{\partial x}(x, t, z) = -\frac{\partial Q}{\partial y}(x, t, z),$$

and so

$$\frac{\partial F}{\partial z}(\mathbf{x}) - \frac{\partial H}{\partial x}(\mathbf{x}) = Q(x, y, z) - Q(x, 0, z) + \frac{\partial \alpha}{\partial z}(x, z) - \frac{\partial \beta}{\partial x}(x, z). \quad (9.54)$$

Now choose

$$\alpha(x, z) = \int_0^z Q(x, 0, t) dt \quad \text{and} \quad \beta(x, z) = 0.$$

with this choice, (9.54) reduces to

$$\frac{\partial F}{\partial z}(\mathbf{x}) - \frac{\partial H}{\partial x}(\mathbf{x}) = Q(x, y, z),$$

and thus we have $\operatorname{curl}\mathbf{A}(\mathbf{x}) = \mathbf{G}(\mathbf{x})$, as we sought.

Definition 107. Let \mathbf{G} be a continuously differentiable vector field on an open set $U \subset \mathbb{R}^3$. Any twice-continuously differentiable vector field $\mathbf{A}(\mathbf{x})$ on U such that $\operatorname{curl}\mathbf{A}(\mathbf{x}) = \mathbf{G}(\mathbf{x})$ for all $\mathbf{x} \in U$ is called a vector potential for \mathbf{G} on U .

We have proved the following theorem:

Theorem 102. Let $\mathbf{G} = (P, Q, R)$ be a continuously differentiable divergence free vector field on \mathbb{R}^3 . Let \mathbf{A} be the vector field on \mathbb{R}^3 defined by $\mathbf{A}(\mathbf{x}) = (F(\mathbf{x}), 0, H(\mathbf{x}))$ where

$$\begin{aligned} F(x, y, z) &= - \int_0^y R(x, t, z) dt + \int_0^z Q(x, 0, t) dt \\ H(x, y, z) &= \int_0^y P(x, t, z) dt. \end{aligned} \quad (9.55)$$

Then \mathbf{A} is a vector potential for \mathbf{G} .

Example 156 (Computing a vector potential). Let $\mathbf{G} = (-xy - 2y, x, yz)$. Then $\operatorname{div}\mathbf{G} = -y + y = 0$, so that $\mathbf{G} = \mathbf{0}$. Writing $\mathbf{G} = (P, Q, R)$, we have $P(x, y, z) = xy - 2y$, $Q(x, y, z) = x$ and $R(x, yz) = yz$. Then from (9.55) we get

$$F(x, y, z) = -\frac{1}{2}y^2z + xz \quad \text{and} \quad H(x, y, z) = -\frac{1}{2}xy^2 - y^2 .$$

Hence

$$\mathbf{A}(x, y, z) = (xz - y^2z/2, 0, -y^2 - xy^2/2) .$$

As you can easily check, $\operatorname{curl}\mathbf{A} = \mathbf{G}$.

9.4 The Laplace operator and Poisson's Equation

9.4.1 The basic problem of electrostatics

The two basic equations of electrostatics that describe the electric field \mathbf{E} that is produced by a static electric charge density $\varrho(\mathbf{x})$ are

$$\operatorname{div}\mathbf{E}(\mathbf{x}) = \varrho(\mathbf{x}) \quad \text{and} \quad \operatorname{curl}\mathbf{E}(\mathbf{x}) = \mathbf{0} . \quad (9.56)$$

Let $\varrho(\mathbf{x})$ be a given charge density. It can take on both positive and negative values since both positive and negative charges exist in nature. Let us assume that the charge is well-localized so that $\varrho(\mathbf{x}) = 0$ for all \mathbf{x} such that $\|\mathbf{x}\| > R$ for some finite R . Let us also suppose that ϱ is continuously differentiable. As we shall now explain, the equations (9.58) specify the vector field \mathbf{E} : We can even use what we have learned to derive a formula for $\mathbf{E}(\mathbf{x})$ in terms of the given charge density ϱ .

The first thing to notice is that since $\operatorname{curl}\mathbf{E}(\mathbf{x}) = \mathbf{0}$, \mathbf{E} is a conservative vector field, and hence, supposing that it is continuously differentiable, it has the form $\mathbf{E}(\mathbf{x}) = \nabla\varphi(\mathbf{x})$ for some twice continuously differentiable function $\varphi(\mathbf{x})$. Now inserting $\mathbf{E}(\mathbf{x}) = \nabla\varphi(\mathbf{x})$ into the first equation in (9.58), we obtain

$$\operatorname{div}\nabla\varphi(\mathbf{x}) = 0 .$$

Writing this out more explicitly,

$$\operatorname{div}\nabla\varphi(\mathbf{x}) = \frac{\partial^2\varphi(\mathbf{x})}{\partial x^2} + \frac{\partial^2\varphi(\mathbf{x})}{\partial y^2} + \frac{\partial^2\varphi(\mathbf{x})}{\partial z^2} . \quad (9.57)$$

Definition 108 (Laplacian). Let $\varphi(\mathbf{x})$ be a twice continuously differentiable function on \mathbb{R}^3 . The Laplacian of φ is the function given by (9.57), and is it denoted by $\Delta\varphi(\mathbf{x})$. That is

$$\Delta\varphi(\mathbf{x}) := \operatorname{div}\nabla\varphi(\mathbf{x}) .$$

The Laplace operator is the transformation from the twice continuously differentiable functions on \mathbb{R}^3 to the continuous functions on \mathbb{R}^3 given by sending φ to $\Delta\varphi$. The operator itself is often written as

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} ,$$

which is then applied to φ to produce $\Delta\varphi$.

We can now re-write (9.58) in terms of φ as

$$\Delta\varphi(\mathbf{x}) = \varrho(\mathbf{x}) \quad \text{and} \quad \mathbf{E}(\mathbf{x}) = \nabla\varphi(\mathbf{x}) . \quad (9.58)$$

The equation $\Delta\varphi(\mathbf{x}) = \varrho(\mathbf{x})$, thought of as an equation for the unknown potential φ in terms of the given electric charge distribution ϱ , is *Poisson's* equation.

We will show that for a continuous charge distribution $\varrho(x)$ that is identically zero outside of a ball of some radius R , there is a *unique* solution of Poisson's equation with the property that

$$\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0 .$$

The gradient of this potential function is then the electric field \mathbf{E} produced by the static charge density ϱ .

First, suppose that $\varphi(\mathbf{x})$ and $\psi(\mathbf{x})$ are two solutions of Poisson's equations for the given charge density ϱ . That is,

$$\Delta\varphi(\mathbf{x}) = \varrho(\mathbf{x}) \quad \text{and} \quad \Delta\psi(\mathbf{x}) = \varrho(\mathbf{x}) . \quad (9.59)$$

Let $\phi = \varphi - \psi$ be their difference. Then, since differentiation is a linear operation,

$$\Delta\phi(\mathbf{x}) = \Delta\varphi(\mathbf{x}) - \Delta\psi(\mathbf{x}) = \varrho(\mathbf{x}) - \varrho(\mathbf{x}) = 0 .$$

Definition 109. A harmonic functions on \mathbb{R}^3 is any twice-continuously differentiable function ϕ on \mathbb{R}^3 such that $\Delta\phi(\mathbf{x}) = 0$ for all \mathbf{x} .

9.4.2 Harmonic functions

We have just seen that the difference $\phi := \varphi - \psi$ of two solutions of Poisson's equation for the charge density ϱ is a harmonic function. Harmonic functions have the *mean value property*, as we now explain. Let $S(\mathbf{x}, r)$ denote the sphere of radius r in \mathbb{R}^3 with the center at \mathbf{x} . likewise, let $B(\mathbf{x}, r)$ denote the ball of radius r in \mathbb{R}^3 centered at \mathbf{x} . Then $S(\mathbf{x}, r)$ is the boundary of $B(\mathbf{x}, r)$. The average of a continuous function ϕ over $S(\mathbf{x}, r)$ is given by $\frac{1}{4\pi r^2} \int_{S(\mathbf{x}, r)} \phi(\mathbf{y}) dS(\mathbf{y})$. The average of a continuous function ϕ over $B(\mathbf{x}, r)$ is given by $\frac{3}{4\pi r^3} \int_{B(\mathbf{x}, r)} \phi(\mathbf{y}) dV(\mathbf{y})$.

Theorem 103 (Mean Value Theorem). *Let ϕ be a harmonic function on \mathbb{R}^3 . Then for all $\mathbf{x} \in \mathbb{R}^3$ and all $r > 0$,*

$$\phi(\mathbf{x}) = \frac{1}{4\pi r^2} \int_{S(\mathbf{x}, r)} \phi(\mathbf{y}) dS(\mathbf{y}) = \frac{3}{4\pi r^3} \int_{B(\mathbf{x}, r)} \phi(\mathbf{y}) dV(\mathbf{y}) . \quad (9.60)$$

That is the average, or mean, value of ϕ over any sphere or ball centered at \mathbf{x} is equal to the value of ϕ at \mathbf{x} .

Proof of Theorem 103. Let ϕ be harmonic. Fix any \mathbf{x} , and then for each $r > 0$ define

$$f(r) = \frac{1}{4\pi r^2} \int_{S(\mathbf{x}, r)} \phi(\mathbf{y}) dS(\mathbf{y}) .$$

Note that we have the alternate formula $f(r) = \frac{1}{4\pi} \int_{S(\mathbf{0},1)} \phi(\mathbf{x} + r\mathbf{y}) dS(\mathbf{y})$ which moves the r from the domain of integration to the integrand. It is now an easy matter to differentiate in r :

$$f'(r) = \frac{1}{4\pi} \int_{S(\mathbf{0},1)} \nabla \phi(\mathbf{x} + r\mathbf{y}) \cdot \mathbf{y} dS(\mathbf{y}).$$

Note that the outward unit normal $\mathbf{N}(\mathbf{y})$ at $\mathbf{y} \in S(\mathbf{0},1)$ is simply \mathbf{y} itself. Hence

$$f'(r) = \frac{1}{4\pi} \int_{S(\mathbf{0},1)} \nabla \phi(\mathbf{x} + r\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS(\mathbf{y}) = \frac{1}{4\pi r^2} \int_{S(\mathbf{x},r)} \nabla \phi(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS(\mathbf{y}).$$

Then by the Divergence Theorem

$$\int_{S(\mathbf{x},r)} \nabla \phi(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS(\mathbf{y}) = \int_{B(\mathbf{x},r)} \operatorname{div} \nabla \phi(\mathbf{y}) dV = \int_{B(\mathbf{x},r)} \Delta \phi(\mathbf{y}) dV = 0.$$

Therefore, $f(r)$ is constant, and since ϕ is continuous, $\phi(\mathbf{x}) = \lim_{r \rightarrow 0} f(r)$. This proves that $f(r) = \phi(\mathbf{x})$ for all \mathbf{x} .

Now integrating in spherical coordinates,

$$\int_{B(\mathbf{x},r)} \phi(\mathbf{y}) dV(\mathbf{y}) = \int_0^r \left(\int_{S(\mathbf{x},s)} \phi(\mathbf{y}) dS(\mathbf{y}) \right) ds = 4\pi \int_0^r s^2 f(s) ds = \frac{4\pi}{3} r^3 \phi(\mathbf{x}).$$

Dividing by the volume of $B(\mathbf{x},r)$ we obtain the second formula. \square

Corollary 7. *Let ϕ be a harmonic function on \mathbb{R}^3 such that $\lim_{\|\mathbf{x}\| \rightarrow \infty} \phi(\mathbf{x}) = 0$. Then $\varphi(\mathbf{x}) = 0$ for all \mathbf{x} .*

Proof. By Theorem 103, $\phi(\mathbf{x}) = \lim_{r \rightarrow \infty} \frac{1}{4\pi r^2} \int_{S(\mathbf{0},1)} \phi(\mathbf{x} + r\mathbf{y}) dS(\mathbf{y}) = 0$. \square

Now let φ and ψ be such that for some charge density ρ , $\Delta \varphi(\mathbf{x}) = \rho(\mathbf{x})$ and $\Delta \psi(\mathbf{x}) = \rho(\mathbf{x})$, and suppose that

$$\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0 \quad \text{and} \quad \lim_{\|\mathbf{x}\| \rightarrow \infty} \psi(\mathbf{x}) = 0.$$

Then defining $\phi = \varphi - \psi$, ϕ is harmonic and $\lim_{\|\mathbf{x}\| \rightarrow \infty} \phi(\mathbf{x}) = 0$. Then Corollary 7 says that $\varphi(\mathbf{x}) = \psi(\mathbf{x})$ for all \mathbf{x} . That is, there is *at most one solution* φ to Poisson's equation $\Delta \varphi = \rho$ with the property that $\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0$. Of course, we may add a constant to any solution of $\Delta \varphi = \rho$ to obtain another solution, so without the condition that $\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0$, there would be no uniqueness. But with it, there is.

We now turn to the existence of solutions.

Lemma 32. *Let $G(\mathbf{x})$ be the function defined by*

$$G(\mathbf{x}) := \frac{1}{\|\mathbf{x}\|}. \tag{9.61}$$

Then for all $\mathbf{x} \neq \mathbf{0}$, $\Delta G(\mathbf{x}) = 0$ and

$$\nabla G(\mathbf{x}) := -\frac{1}{\|\mathbf{x}\|^3} \mathbf{x}. \tag{9.62}$$

Proof. We write $G(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{x})^{-1/2}$, and then by the chain rule,

$$\nabla G(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} \cdot \mathbf{x})^{-3/2}2\mathbf{x} = -\frac{1}{\|\mathbf{x}\|^3}\mathbf{x} .$$

Then

$$\operatorname{div}(\nabla G(\mathbf{x})) = \operatorname{div}((\mathbf{x} \cdot \mathbf{x})^{-3/2}\mathbf{x}) = 3(\mathbf{x} \cdot \mathbf{x})^{-5/2}\mathbf{x} \cdot \mathbf{x} - 3(\mathbf{x} \cdot \mathbf{x})^{-3/2} = 0 .$$

□

Lemma 33. *Let D be a region bounded by the piecewise continuously differentiable surface \mathcal{S} . Let f and g be two twice-continuously differentiable functions on D . Then*

$$\int_D \nabla f(\mathbf{x}) \cdot \nabla g(\mathbf{x}) dV = - \int_D f(\mathbf{x}) \Delta g(\mathbf{x}) dV + \int_{\mathcal{S}} f(\mathbf{x}) \nabla g(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}) dS . \quad (9.63)$$

Proof. We compute

$$\operatorname{div}(f(\mathbf{x}) \nabla g(\mathbf{x})) = \nabla f(\mathbf{x}) \cdot \nabla g(\mathbf{x}) + f(\mathbf{x}) \Delta g(\mathbf{x}) .$$

Now integrate both sides over D , and use the Divergence Theorem to conclude

$$\int_D \operatorname{div}(f(\mathbf{x}) \nabla g(\mathbf{x})) dV = \int_{\mathcal{S}} f(\mathbf{x}) \nabla g(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}) dS .$$

□

Since (9.64) is symmetric in f and g , we have, under the same hypotheses on f and g , that

$$-\int_D f(\mathbf{x}) \Delta g(\mathbf{x}) dV + \int_{\mathcal{S}} f(\mathbf{x}) \nabla g(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}) dS = -\int_D g(\mathbf{x}) \Delta f(\mathbf{x}) dV + \int_{\mathcal{S}} g(\mathbf{x}) \nabla f(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}) dS . \quad (9.64)$$

We now apply this and Lemma 32 as follows. Fix $0 < r < R$, $\mathbf{x} \in \mathbb{R}^3$, and let $A(\mathbf{x}, r, R)$ denote the annular region

$$A(\mathbf{x}, r, R) = \{\mathbf{y} : r < \|\mathbf{y} - \mathbf{x}\| < R\} .$$

Then the boundary of $A(\mathbf{x}, r, R)$ consists of two pieces, $S(\mathbf{x}, r)$ and $S(\mathbf{x}, R)$, but the outward normal for $A(\mathbf{x}, r, R)$ on $S(\mathbf{x}, r)$ points *towards* \mathbf{x} ; it is the opposite of the usual “outward normal” on $S(\mathbf{x}, r)$.

Now let $g(\mathbf{y})$ be given by $g(\mathbf{y}) := G(\mathbf{x} - \mathbf{y})$ where G is defined in Lemma 32. Then by Lemma 32, $\Delta g(\mathbf{x}) = 0$ everywhere in $A(\mathbf{x}, r, R)$. Then with this choice of g and with $D_{\mathbf{x}, r, R} = A(\mathbf{x}, r, R)$ and $\mathcal{S}_{\mathbf{x}, r, R}$ denoting its boundary, (9.64) becomes

$$\int_{D_{\mathbf{x}, r, R}} g(\mathbf{y}) \Delta f(\mathbf{y}) dV = - \int_{\mathcal{S}_{\mathbf{x}, r, R}} f(\mathbf{y}) \nabla g(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS + \int_{\mathcal{S}_{\mathbf{x}, r, R}} g(\mathbf{y}) \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS . \quad (9.65)$$

We will take the limit $r \rightarrow 0$ and $R \rightarrow \infty$, and shall deduce the following:

Theorem 104. *Let f be a twice continuously differentiable function such that $\lim_{\|\mathbf{x}\| \rightarrow \infty} f(\mathbf{x}) = 0$ and such that $\int_{\mathbb{R}^3} |\Delta f(\mathbf{y})| dV < \infty$. Then*

$$f(\mathbf{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \Delta f(\mathbf{y}) dV .$$

Proof. By Lemma 32, $\nabla g(\mathbf{y}) = -\|\mathbf{y} - \mathbf{x}\|^{-3}(\mathbf{y} - \mathbf{x})$, and hence

$$\nabla g(\mathbf{y}) \cdot \mathbf{N} = \frac{1}{\|\mathbf{y} - \mathbf{x}\|^3}(\mathbf{y} - \mathbf{x}) \cdot \mathbf{N}(\mathbf{y}) = \pm \frac{1}{\|\mathbf{y} - \mathbf{x}\|^3}(\mathbf{y} - \mathbf{x}) \cdot \frac{1}{\|\mathbf{y} - \mathbf{x}\|}(\mathbf{y} - \mathbf{x}) = \pm \frac{1}{\|\mathbf{y} - \mathbf{x}\|^2},$$

where the $-$ sign is correct for $\mathbf{y} \in S(\mathbf{x}, R)$, and the $+$ sign is correct for $\mathbf{y} \in S(\mathbf{x}, r)$. Therefore,

$$\int_{S_{\mathbf{x},r,R}} f(\mathbf{y}) \nabla g(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dV = \frac{1}{r^2} \int_{S(\mathbf{x},r)} f(\mathbf{y}) dS - \frac{1}{R^2} \int_{S(\mathbf{x},R)} f(\mathbf{y}) dS.$$

Then since f is continuous, its average over $S(\mathbf{x}, r)$ tends to $f(\mathbf{x})$ as r tends to 0. That is,

$$\lim_{r \rightarrow 0} \frac{1}{4\pi r^2} \int_{S(\mathbf{x},r)} f(\mathbf{y}) dS = f(\mathbf{x}).$$

Likewise, since $\lim_{\|\mathbf{x}\| \rightarrow \infty} f(\mathbf{x}) = 0$, the average of f over $S(\mathbf{x}, R)$ tends to zero as R tends to infinity. That is,

$$\lim_{R \rightarrow \infty} \frac{1}{4\pi R^2} \int_{S(\mathbf{x},R)} f(\mathbf{y}) dS = 0.$$

Altogether, we conclude that

$$\lim_{r \rightarrow 0, R \rightarrow \infty} \left(- \int_{S_{\mathbf{x},r,R}} f(\mathbf{y}) \nabla g(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dV \right) = -4\pi f(\mathbf{x}). \quad (9.66)$$

Next, since $g(\mathbf{y}) = 1/r$ on $S(\mathbf{x}, r)$, and $g(\mathbf{y}) = 1/R$ on $S(\mathbf{x}, R)$, with \mathbf{N} now denoting the usual *outward* unit normals on $S(\mathbf{x}, r)$ and $S(\mathbf{x}, R)$ respectively,

$$\int_S g(\mathbf{y}) \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS = -\frac{1}{r} \int_{S(\mathbf{x},r)} \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS + \frac{1}{R} \int_{S(\mathbf{x},R)} \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS.$$

By the Divergence Theorem, $\int_{S(\mathbf{x},r)} \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS = \int_{B(\mathbf{x},r)} \Delta f(\mathbf{y}) dV$ and since Δf is continuous, $\lim_{r \rightarrow 0} \frac{3}{4\pi r^3} \int_{B(\mathbf{x},r)} \Delta f(\mathbf{y}) dV = \Delta f(\mathbf{x})$. Therefore, $\lim_{r \rightarrow 0} \frac{1}{r} \int_{S(\mathbf{x},r)} \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS = 0$. Similarly, by the Divergence Theorem,

$$\frac{1}{R} \int_{S(\mathbf{x},R)} \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dS = \frac{1}{R} \int_{B(\mathbf{x},R)} \Delta f(\mathbf{y}) dV.$$

Then under the assumption that $\int_{\mathbb{R}^3} |\Delta f(\mathbf{y})| dV < \infty$, $\lim_{R \rightarrow \infty} \int_{B(\mathbf{x},R)} |\Delta f(\mathbf{y})| dV = 0$. Altogether,

$$\lim_{r \rightarrow 0, R \rightarrow \infty} \left(- \int_{S_{\mathbf{x},r,R}} g(\mathbf{y}) \nabla f(\mathbf{y}) \cdot \mathbf{N}(\mathbf{y}) dV \right) = 0. \quad (9.67)$$

□

Theorem 104 gives us a candidate for the solution of Poisson's equation $\Delta \varphi(\mathbf{x}) = \varrho(\mathbf{x})$: Define

$$\varphi(x) := -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \varrho(\mathbf{y}) dV. \quad (9.68)$$

This is an “improper integral” in that the domain of integration is unbounded and, for some \mathbf{x} , the integrand is not continuous and bounded due to the singularity at $\mathbf{y} = \mathbf{x}$. However, recall that we are assuming that ϱ is continuous, and that for some R , $\varrho(\mathbf{y}) = 0$ whenever $\|\mathbf{y}\| > R$. Under these assumptions, the integral defining φ converges for all \mathbf{x} , and moreover:

Lemma 34. Suppose that ϱ is continuous, and that for some R , $\varrho(\mathbf{y}) = 0$ whenever $\|\mathbf{y}\| > R$. Then the integral in (9.68) converges for all $\mathbf{x} \in \mathbb{R}^3$, so that φ is a well-defined function on \mathbb{R}^3 . Moreover, φ satisfies $\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0$.

Proof. First suppose $\|\mathbf{x}\| \geq R + s$ for $s > 0$. Then by the triangle inequality, $\|\mathbf{x}\| = \|\mathbf{x} - \mathbf{y} + \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y}\|$, so that

$$\|\mathbf{x} - \mathbf{y}\| \geq \|\mathbf{x}\| - \|\mathbf{y}\| \geq (R + s) - R = s .$$

Hence for $\|\mathbf{x}\| > R + s$,

$$\left| \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \varrho(\mathbf{y}) dV \right| \leq \frac{1}{s} \int_{\mathbb{R}^3} |\varrho(\mathbf{y})| dV = \frac{1}{s} \int_{B(\mathbf{0}, R)} |\varrho(\mathbf{y})| dV .$$

Since ϱ is continuous and equals zero outside of $B(\mathbf{0}, R)$, there is a constant C so that $|\varrho(\mathbf{y})| \leq C$ for all \mathbf{y} . Then $\int_{B(\mathbf{0}, R)} |\varrho(\mathbf{y})| dV \leq \frac{4\pi}{3} R^3 C$. Hence, for $\|\mathbf{x}\| > R$,

$$\left| \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \varrho(\mathbf{y}) dV \right| \leq \frac{1}{\|\mathbf{x} - R\|} \frac{4\pi}{3} R^3 C .$$

This proves that for $\|\mathbf{x}\| > R$, the integral defining $\varphi(\mathbf{x})$ makes perfect sense and moreover, $\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0$. In fact, for any $p < 1$, $\lim_{\|\mathbf{x}\| \rightarrow \infty} \|\mathbf{x}\|^p \varphi(\mathbf{x}) = 0$.

It remains to show that for $\|\mathbf{x}\| \leq R$, the integral defining φ is convergent. If $\|\mathbf{x}\| \leq R$ and $\|\mathbf{y}\| \leq R$, then by the triangle inequality, $\|\mathbf{x} - \mathbf{y}\| \leq 2R$. Since $|\varrho(\mathbf{y})| \leq C$ for all \mathbf{y} , when $\|\mathbf{x}\| \leq R$,

$$\frac{1}{\|\mathbf{x} - \mathbf{y}\|} |\varrho(\mathbf{y})| \leq \begin{cases} C/\|\mathbf{x} - \mathbf{y}\| & \|\mathbf{x} - \mathbf{y}\| \leq 2R \\ 0 & \|\mathbf{x} - \mathbf{y}\| > 2R . \end{cases}$$

In particular, we are not actually integrating over all of \mathbb{R}^3 ; since the integrand is zero for $\|\mathbf{y} - \mathbf{x}\| > 2R$,

$$\int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} |\varrho(\mathbf{y})| dV = \int_{B(\mathbf{x}, 2R)} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} |\varrho(\mathbf{y})| dV \leq C \int_{B(\mathbf{x}, 2R)} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} dV ,$$

and it suffices to show that this last integral is actually convergent. This is easy:

$$\int_{A(\mathbf{x}, r, 2R)} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} dV(\mathbf{y}) = 4\pi \int_r^R \frac{1}{s} s^2 ds = C2\pi(R^2 - r^2) .$$

This shows that

$$\lim_{r \rightarrow 0} \int_{A(\mathbf{x}, r, R)} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} dV(\mathbf{y}) \leq C\pi R^2 < \infty .$$

Therefore, the integral in (9.68) makes perfect sense for all \mathbf{x} . \square

From here, it is easy to prove that φ is the solution to Poisson's equation that we seek. Making the change of variable $\mathbf{y} = \mathbf{x} - \mathbf{z}$, for which the Jacobian determinant is simply 1,

$$\int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \varrho(\mathbf{y}) dV(\mathbf{y}) = \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{z}\|} \varrho(\mathbf{x} - \mathbf{z}) dV(\mathbf{z}) .$$

Then differentiating under the integral sign, and undoing the change of variables,

$$\Delta \varphi(\mathbf{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{z}\|} \Delta \varrho(\mathbf{x} - \mathbf{z}) dV(\mathbf{z}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \Delta \varrho(\mathbf{y}) dV(\mathbf{y}) = \varrho(\mathbf{x}) ,$$

where in the last equality we have used Theorem 104. Altogether we have proved:

Theorem 105. Let ϱ be a twice continuously differentiable function on \mathbb{R}^3 such that for some R , $\varrho(\mathbf{y}) = 0$ if $\|\mathbf{y}\| > R$. Then there exists a unique twice continuously differentiable function φ on \mathbb{R}^3 such that

$$\Delta\varphi(\mathbf{x}) = \varrho(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{R}^3$$

and such that $\lim_{\|\mathbf{x}\| \rightarrow \infty} \varphi(\mathbf{x}) = 0$. Moreover, φ is given by (9.68).

Differentiation is a linear operation, and so we may regard the Laplace operator Δ as a linear transformation, acting not on a finite dimensional vector space, but on the space of twice continuously differentiable functions. This space may be regarded as a vector space because we have a natural notion of “vector addition” and “scalar multiplication” on it. More specifically, for numbers $a, b \in \mathbb{R}$ and twice continuously differentiable functions φ and ψ define $a\varphi + b\psi$ to be the function given by

$$(a\varphi + b\psi)(\mathbf{x}) = a\varphi(\mathbf{x}) + b\psi(\mathbf{x}).$$

There is no finite set of functions that spans this vector space, so it is infinite dimensional.

Theorem 105 can be viewed as specifying a transformation that is *inverse* to the Laplacian. While the Laplacian is a differential operator – its action consists of taking derivatives – the inverse transformation is an *integral operator* as we might expect on account of the Fundamental Theorem of Calculus. If we define the operation G on the space of twice continuously differentiable functions that are identically zero outside of $B(\mathbf{0}, R)$ for some R by

$$G\varphi(\mathbf{x}) := -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \varphi(\mathbf{y}) dV.$$

then by Theorem 104, when $\Delta\varphi$ is twice continuously differentiable and equals 0 outside $B(\mathbf{0}, R)$ for some R ,

$$G\Delta\varphi(\mathbf{x}) = \varphi(\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^3$.

9.4.3 The Hodge Decomposition of vector fields

Let $\mathbf{F}(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$ be a twice continuously differentiable vector field. Then $\operatorname{curl}\mathbf{F}(\mathbf{x})$ is a continuously differentiable vector field, and we may take the curl again. Since

$$\operatorname{curl}\mathbf{F}(\mathbf{x}) = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right),$$

the x -component of $\operatorname{curl}(\operatorname{curl}\mathbf{F})(\mathbf{x})$ is

$$\begin{aligned} \frac{\partial}{\partial y} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) - \frac{\partial}{\partial z} \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x} \right) &= - \left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) P(\mathbf{x}) + \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} Q(\mathbf{x}) + \frac{\partial}{\partial z} R(\mathbf{x}) \right) \\ &= -\Delta P(\mathbf{x}) + \frac{\partial}{\partial x} (\operatorname{div}\mathbf{F}(\mathbf{x})). \end{aligned}$$

Similar computations show that the y -component is $-\Delta Q(\mathbf{x}) + \frac{\partial}{\partial x} (\operatorname{div}\mathbf{F}(\mathbf{x}))$ and that the z -component is $-\Delta R(\mathbf{x}) + \frac{\partial}{\partial z} (\operatorname{div}\mathbf{F}(\mathbf{x}))$. Therefore, if we define

$$\Delta\mathbf{F}(\mathbf{x}) = (\Delta P(\mathbf{x}), \Delta Q(\mathbf{x}), \Delta R(\mathbf{x})),$$

we have proved the following:

Lemma 35 (Curl-curl identity). *Let $\mathbf{F}(\mathbf{x})$ be a twice continuously differentiable vector field. Then*

$$\operatorname{curl}(\operatorname{curl}\mathbf{F})(\mathbf{x}) = \nabla(\operatorname{div}\mathbf{F}(\mathbf{x})) - \Delta\mathbf{F}(\mathbf{x}) . \quad (9.69)$$

Now suppose that \mathbf{F} is a four times continuously differentiable vector field such that for some R , $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ whenever $\|\mathbf{x}\| > R$. Define $\varrho(\mathbf{x}) := \operatorname{div}\mathbf{F}(\mathbf{x})$. (We will be taking lots of derivatives soon.) Then ϱ is (at least) twice continuously differentiable and $\varrho(\mathbf{x}) = 0$ whenever $\|\mathbf{x}\| > R$. Define $\phi(\mathbf{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \varrho(\mathbf{y}) dV$ which, by the definition of ϱ , is the same as

$$\phi(\mathbf{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \operatorname{div}\mathbf{F}(\mathbf{y}) dV . \quad (9.70)$$

By Theorem 104, $\lim_{\|\mathbf{x}\| \rightarrow \infty} \phi(\mathbf{x}) = 0$, and $\Delta\phi(\mathbf{x}) = \operatorname{div}\nabla\phi(\mathbf{x}) = \varrho(\mathbf{x}) = \operatorname{div}\mathbf{F}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^3$. Therefore, if we define

$$\mathbf{G}(\mathbf{x}) := \mathbf{F}(\mathbf{x}) - \nabla\phi(\mathbf{x}) ,$$

$\operatorname{div}\mathbf{G}(\mathbf{x}) = \varrho(\mathbf{x}) - \varrho(\mathbf{x}) = 0$, and hence $\mathbf{G}(\mathbf{x})$ is divergence free and continuously differentiable. Then by Theorem 102, there is a vector field \mathbf{A} such that $\mathbf{G}(\mathbf{x}) = \operatorname{curl}\mathbf{A}(\mathbf{x})$ for all \mathbf{x} .

There is another way to find \mathbf{A} using the curl-curl identity: Notice that since the curl of a gradient is zero, $\operatorname{curl}\mathbf{G}(\mathbf{x}) = \operatorname{curl}\mathbf{F}(\mathbf{x})$ for all \mathbf{x} . Hence

$$\operatorname{curl}(\operatorname{curl}\mathbf{F})(\mathbf{x}) = \operatorname{curl}(\operatorname{curl}\mathbf{G})(\mathbf{x}) = -\Delta\mathbf{G}(\mathbf{x})$$

since $\mathbf{G}(\mathbf{x})$ is divergence free. Since $\mathbf{F}(\mathbf{x}) = 0$ whenever $\|\mathbf{x}\| > R$, each component of $\operatorname{curl}(\operatorname{curl}\mathbf{F})$ has this property, and is also twice continuously differentiable since \mathbf{F} itself is four times continuously differentiable. Then by Theorem 104,

$$\mathbf{G}(\mathbf{x}) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \operatorname{curl}(\operatorname{curl}\mathbf{F})(\mathbf{y}) dV .$$

Now define

$$\mathbf{A}(\mathbf{x}) := \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \operatorname{curl}\mathbf{F}(\mathbf{y}) dV(\mathbf{y}) . \quad (9.71)$$

Making the change of variables $\mathbf{z} := \mathbf{x} - \mathbf{y}$, whose Jacobian factor is 1, we can write this as

$$\mathbf{A}(\mathbf{x}) := \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{\|\mathbf{z}\|} \operatorname{curl}\mathbf{F}(\mathbf{x} - \mathbf{z}) dV(\mathbf{z}) .$$

Differentiating under the integral sign, and then undoing the change of variables, we see that $\operatorname{curl}\mathbf{A}(\mathbf{x}) = \mathbf{G}(\mathbf{x})$ for all \mathbf{x} . Therefore, \mathbf{F} can be decomposed as $\mathbf{F} = \nabla\phi + \operatorname{curl}\mathbf{A}$ where ϕ is given by where ϕ is given by (9.70) and \mathbf{A} is given by (9.71), and

$$\lim_{\|\mathbf{x}\| \rightarrow 0} \phi(\mathbf{x}) = 0 \quad \text{and} \quad \lim_{\|\mathbf{x}\| \rightarrow 0} \mathbf{A}(\mathbf{x}) = \mathbf{0} . \quad (9.72)$$

This decomposition is unique: Suppose that we also have $\mathbf{F} = \nabla\psi + \operatorname{curl}\mathbf{B}$ where ψ and \mathbf{B} satisfy the analogs of (9.72). Then

$$\nabla(\phi - \psi) = \operatorname{curl}(\mathbf{B} - \mathbf{A}) .$$

Taking the divergence of both sides, we obtain $\Delta(\phi - \psi) = 0$, so that $\phi - \psi$ is harmonic, and furthermore $\lim_{\|\mathbf{x}\| \rightarrow \infty} (\phi(\mathbf{x}) - \psi(\mathbf{x})) = 0$. Then by Corollary 7, $\phi(\mathbf{x}) - \psi(\mathbf{x}) = 0$ for all \mathbf{x} . Hence $\phi = \psi$, and then $\mathbf{B} = \mathbf{A}$.

Theorem 106. Let \mathbf{F} be a four times continuously differentiable vector field such that for some R , $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ whenever $\|\mathbf{x}\| > R$. Then there is a unique twice continuously differentiable function ϕ and a unique twice continuously differentiable vector field \mathbf{A} such that $\mathbf{F}(\mathbf{x}) = \nabla\phi(\mathbf{x}) + \operatorname{curl}\mathbf{A}(\mathbf{x})$ and such that (9.72) is satisfied. Moreover, ϕ is given by (9.70) and \mathbf{A} is given by (9.71)

The decomposition provided by Theorem 106 is called the *Hodge decomposition* of \mathbf{F} . Notice that if \mathbf{F} satisfies the hypotheses of Theorem 106, and $\operatorname{div}\mathbf{F}(\mathbf{x}) = 0$ for all \mathbf{x} and $\operatorname{curl}\mathbf{F}(\mathbf{x}) = \mathbf{0}$ for all \mathbf{x} , then by (9.70), $\phi(\mathbf{x}) = 0$ for all \mathbf{x} and by (9.71), $\mathbf{A}(\mathbf{x}) = \mathbf{0}$ for all \mathbf{x} . Then $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ for all \mathbf{x} : Under suitable smoothness and decay properties (such as (9.72)), a vector field \mathbf{F} is determined by its curl and its divergence.

Without some decay condition such as (9.72), there can be no uniqueness. There are plenty of functions $h(\mathbf{x})$ that are Harmonic on all of \mathbb{R}^3 ; e.g. $h(\mathbf{x}) = x$ or $h(\mathbf{x}) = xyz$. Then adding h to ϕ and subtracting ∇h from \mathbf{A} gives us a new pair $\psi = \phi - h$ and $\mathbf{B} = \mathbf{A} - \nabla h$ such that $\mathbf{F} = \nabla\psi + \operatorname{curl}\mathbf{B}$. However, ψ and \mathbf{B} do not satisfy the analog of (9.72).

9.5 Exercises

1. Let $\mathcal{D} \subset \mathbb{R}^2$ be the region that is to the left of the parabola $x = y(2 - y)$ and below the line $x - 2y + 4 = 0$. Let \mathcal{C} be its boundary given the outward normal orientation. Let $\mathbf{F}(x, y) = (-2xy, 4y + xy)$. Calculate the flux integral $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds$ both directly, and by making use of the Divergence Theorem.
2. Let \mathcal{C} be the oriented curve in the plane that starts at $(0, 0)$, and moves along straight line segments from this point to $(1, 2)$, then from this point to $(-1, 4)$, then from this point to $(-3, 2)$, and finally then from this point to $(-2, 0)$. Let $\mathbf{F}(x, y) = (x^3y + y^2x^2, x + y + x^2y + y^2x)$. Compute the flux integral $\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} ds$.

- 3: Let \mathcal{S} be the part of the surface in \mathbb{R}^3 given by $\sqrt{x^2 + y^2} = 8 - z$ that lies inside the cylinder $x^2 + y^2 = 4$. With $\mathbf{F} = (2yz - y^2, x^2z - 2x, x^2y)$, compute the flux

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS ,$$

where \mathbf{N} is taken to point outward from the z -axis.

- 4: Let \mathcal{V} be the region in \mathbb{R}^3 that lies inside the sphere $x^2 + y^2 + z^2 = 4$, and above the graph of $z = 1/\sqrt{x^2 + y^2}$, as in problem 8. Let $\mathbf{F} = (y + z^2, x + z^2, 2z(x + y))$ and let \mathbf{N} be the outward normal to \mathcal{S} , the boundary of \mathcal{V} . Compute the total flux

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS .$$

- 5: Consider the two vector fields

$$\mathbf{F} = (yz^2 - 2xy, xz^2 - x^2, 2xyz) \quad \text{and} \quad \mathbf{G} = (z^2, y, x) .$$

- (a) Compute the divergence of \mathbf{F} and \mathbf{G} .

(b) Let \mathcal{S} be the unit sphere, and \mathbf{N} its outward normal. Compute

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS \quad \text{and} \quad \int_{\mathcal{S}} \mathbf{G} \cdot \mathbf{N} dS .$$

Justify your answers to receive credit.

6: Consider the two vector fields

$$\mathbf{F} = (y + z^2, x + z^2, 2zx + 2zy) \quad \text{and} \quad \mathbf{G} = (y + z^2, x + z^2, 2x + 2y) .$$

(a) Compute the divergence \mathbf{F} and \mathbf{G} .

(b) Let \mathcal{V} be the intersection of the ball of radius 1 centered at the origin, and the ball of radius 1 centered at $(1, 0, 0)$. Let \mathcal{S} be the boundary of \mathcal{V} oriented with the outward unit normal \mathbf{N} . Compute

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS \quad \text{and} \quad \int_{\mathcal{S}} \mathbf{G} \cdot \mathbf{N} dS .$$

7: As in Exercise 3, let \mathcal{S} be the part of the surface in \mathbb{R}^3 given by $\sqrt{x^2 + y^2} = 8 - z$ that lies inside the cylinder $x^2 + y^2 = 4$. With $\mathbf{F} = (2yz - y^2, x^2z - 2x, x^2y)$, Use Stokes' Theorem evaluate the flux

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS ,$$

where \mathbf{N} is taken to point outward from the z -axis, by computing a line integral.

8: Consider the two vector fields

$$\mathbf{F} = (yz^2 - 2xy, xz^2 - x^2, 2xyz) \quad \text{and} \quad \mathbf{G} = (z^2, y, x) .$$

(a) Compute the curls of \mathbf{F} and \mathbf{G} .

(b) Let \mathcal{S} be the part of the centered sphere of radius 2 that lies above the plane $x + y + z = 1$, oriented with its unit normal \mathbf{N} pointing upwards. Compute

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS \quad \text{and} \quad \int_{\mathcal{S}} \mathbf{G} \cdot \mathbf{N} dS .$$

Justify your answers to receive credit.

(c) One of these vector fields is conservative. Identify the conservative vector field, and a potential function for it.

9: Consider the two vector fields

$$\mathbf{F} = (y + z^2, x + z^2, 2zx + 2zy) \quad \text{and} \quad \mathbf{G} = (y + z^2, x + z^2, 2x + 2y) .$$

(a) Compute the curl of \mathbf{F} and \mathbf{G} .

(b) Let \mathcal{S} be the part of the ellipsoidal surface $x^2 + \frac{1}{2}y^2 + \frac{1}{4}z^2 = 1$ above the plane $z = 1$, oriented so the unit normal \mathbf{N} points upwards. Compute

$$\int_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} dS \quad \text{and} \quad \int_{\mathcal{S}} \mathbf{G} \cdot \mathbf{N} dS .$$

(c) One of these vector fields is conservative. Identify the conservative vector field, and a potential function for it.

