Honors Math A Notes

Robert Friedman

December 10, 2010

Contents

Chapter 1

Vectors

We begin with a discussion of vectors in $\mathbb{R}^2$. There are three general ideas behind the definition. One is the need to model forces coming from physics. A second motivation is to translate the questions of Euclidean geometry into an algebraic framework. Finally there is the (algebraic) problem of solving systems of two linear equations in two unknowns.

For a physicist, a vector $\mathbf{v}$ is an arrow with magnitude and direction, corresponding typically to a force acting upon a particle. In physics, the magnitude of the vector $\mathbf{v}$ is often denoted by v but we shall denote it by $\|\mathbf{v}\|$. We could think of this description as saying that a vector is a directed line segment in the plane (directed in that it has a starting point and an end point, indicated by an arrowhead). Two such arrows are declared to be equal if they have the same magnitude and direction; mathematically, we can speak of directed line segments and formulate a notion of when two such are equivalent. If we need to record as well the starting point of the vector, we speak of a *located* vector; if the starting point is some point P in the plane, we speak of $\mathbf{v}$ as *located at* P. Finally, there is the *zero vector*, whose starting point is equal to its endpoint. (Its magnitude is zero and its direction is undefined.)

It is easier mathematically to keep track of vectors by always locating them at the origin O (given a choice of origin for the plane). In this way we may identify a vector with its directed endpoint (the one with the arrowhead) and so we can identify all vectors with the set of points of the plane. Now if we have introduced Cartesian coordinates, the plane may be further identified with the set of all ordered pairs of real numbers (x, y), in other words with $\mathbb{R}^2$. The upshot is that a vector $\mathbf{v}$ in the plane is the same thing as an ordered pair (x, y). We call the entries of the vector $\mathbf{v} = (x, y)$ its

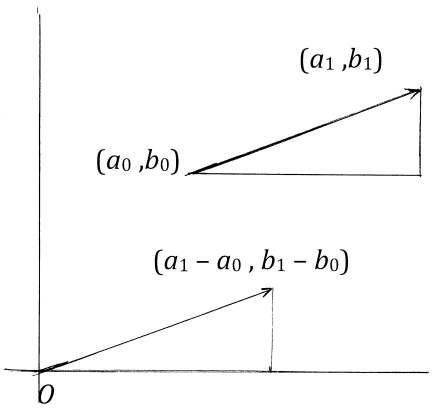

Figure 1.1: Locating a vector at the origin

components, so that x is the first component and y the second (or sometimes the x- and y-components).

Thus, a vector located at (a_0, b_0) with directed endpoint (a_1, b_1) may be identified with a point of $\mathbb{R}^2$ corresponding to a vector located at $O = (0, 0)$. What is the corresponding point of $\mathbb{R}^2$? We claim that it is $(a_1 - a_0, b_1 - b_0)$. This follows by examining the congruent triangles depicted in Figure 1.1. Note that the zero vector is then $\mathbf{0} = (0, 0)$.

Next, vectors can be added. Physically this corresponds to a superposition of two forces corresponding to vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, and experiment shows that in many cases the combined force is given by the *parallelogram law* (Figure 1.2): we imagine that $\mathbf{v}_1$ and $\mathbf{v}_2$ are located at the same point P. Then, unless $\mathbf{v}_1$ and $\mathbf{v}_2$ lie on a line, they define a parallelogram, three of whose vertices are P and the endpoints of $\mathbf{v}_1$ and $\mathbf{v}_2$. The vector sum of $\mathbf{v}_1$ and $\mathbf{v}_2$ is then the directed line segment starting at P and ending at the fourth vertex of the parallelogram. There are special rules in case $\mathbf{v}_1$ and $\mathbf{v}_2$ lie on a line, or one or both of them are the zero vector, but it is simplest to give an algebraic formula for the sum which works in all cases.

To find the formula for the parallelogram law, we seek to find an algebraic formula for the addition of twovectors $\mathbf{v}_1 = (x_1, y_1)$ and $\mathbf{v}_2 = (x_2, y_2)$

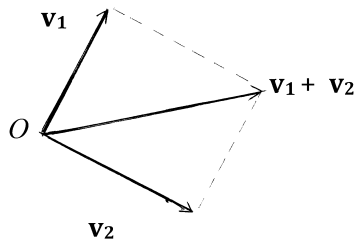

Figure 1.2: The parallelogram law

corresponding to the parallelogram law. By the above discussion on located vectors, the diagonal $\mathbf{v}$ of the parallelogram drawn in Figure 1.2 is equal to $(x_1 + x_2, y_1 + y_2) = \mathbf{v}_1 + \mathbf{v}_2$. Indeed, the vector located at (x_1, y_1) whose endpoint is $(x_1 + x_2, y_1 + y_2)$ is congruent as a directed line segment to the vector located at O with endpoint

$$(x_1 + x_2 - x_1, y_1 + y_2 - y_1) = (x_2, y_2) = \mathbf{v}_2,$$

and similarly for the vector located at (x_2, y_2) whose endpoint is (x_1+x_2, y_1+y_2). In this way the geometry of the parallelogram law has been translated into a very straightforward algebraic operation.

As should be familiar, two vectors $\mathbf{v}_1 = (x_1, y_1)$ and $\mathbf{v}_2 = (x_2, y_2)$ lie on the same line through O if either $\mathbf{v}_1 = \mathbf{0}$ or there is a real number t such that $x_2 = tx_1$ and $y_2 = ty_1$. (This include both the case where $x_1 \neq 0$, so that the slope of the line through O and $\mathbf{v}_1$ is y_1/x_1, and the case where $x_1 = 0$ and $\mathbf{v}_1$ lies on the y-axis.) We may write the condition that $x_2 = tx_1$ and $y_2 = ty_1$ more briefly by saying that $\mathbf{v}_2 = t\mathbf{v}_1$. This gives a new operation on vectors and real numbers called *scalar multiplication*. Unlike usual multiplication however where we combine two numbers and get a third, this operation combines a *real number* t (called a *scalar* in physics because multiplying by

t changes the scale) and a *vector* $\mathbf{v}$ to produce a new vector, written $t\mathbf{v}$. Thus unlike the usual algebraic operations where we combine two apples to produce an apple, this operation combines an apple and an orange to produce an orange. As we shall see below, if $t > 0$, $t\mathbf{v}$ points in the same direction as $\mathbf{v}$ and its magnitude is $t\|\mathbf{v}\|$, whereas if $t < 0$, then $t\mathbf{v}$ points in the *opposite* direction as $\mathbf{v}$ and its magnitude is $|t|\|\mathbf{v}\|$. Of course, if $t = 0$, then $t\mathbf{v} = \mathbf{0}$ does not have a direction and its magnitude is 0.

We record the following properties of vector addition:

1. For all $\mathbf{v}, \mathbf{w}, \mathbf{u} \in \mathbb{R}^2$, $(\mathbf{v} + \mathbf{w}) + \mathbf{u} = \mathbf{v} + (\mathbf{w} + \mathbf{u})$.

2. For all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$, $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$.

3. For all $\mathbf{v} \in \mathbb{R}^2$, $\mathbf{v} + \mathbf{0} = \mathbf{v}$.

4. For all $\mathbf{v} \in \mathbb{R}^2$, $\mathbf{v} + (-1)\mathbf{v} = \mathbf{0}$.

We denote the vector $(-1)\mathbf{v}$ by $-\mathbf{v}$ and $\mathbf{v} + (-\mathbf{w})$ by $\mathbf{v} - \mathbf{w}$.

Next there are some properties which relate scalar multiplication to vector addition:

1. For all $s, t \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^2$, $s(t\mathbf{v}) = (st)\mathbf{v}$.

2. For all $s, t \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^2$, $(s + t)\mathbf{v} = s\mathbf{v} + t\mathbf{v}$.

3. For all $t \in \mathbb{R}$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$, $t(\mathbf{v} + \mathbf{w}) = t\mathbf{v} + t\mathbf{w}$.

4. For all $\mathbf{v} \in \mathbb{R}^2$, $1 \cdot \mathbf{v} = \mathbf{v}$.

Here the first group of properties is the natural analogue of the formal properties of addition of real numbers, and the second group reflects the fact that multiplication of real numbers is associative, distributes over addition, and has a multiplicative identity. Of course these properties are not exactly the analogues of the associativity or distributivity laws since scalar multiplication combines a scalar and a vector to produce a vector. Note that there is no sense in which we multiply two vectors here and so there is no possible sense in which we would expect to find a multiplicative inverse. The verification of all of these properties is a completely mechanical exercise, which consists in writing out what they mean in terms of the components and then reducing each one to a standard property of real numbers. We will use all of these properties (and some others which can easily be derived from them or checked directly) without comment.

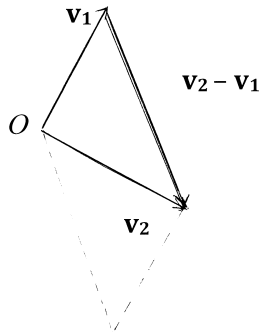

Figure 1.3: Vector subtraction and triangles

Note that the equation $\mathbf{v} + (-\mathbf{v}) = \mathbf{0} = (0, 0)$ is a special case of the parallelogram law for a *very degenerate* parallelogram. More generally, the equation $t\mathbf{v} + s\mathbf{v} = (t + s)\mathbf{v}$ is a special case of the parallelogram law. We write $\mathbf{v}_1 - \mathbf{v}_2$ for the vector $\mathbf{v}_1 + (-\mathbf{v}_2)$. It is easy to see from the parallelogram law that the vector $\mathbf{v}_1 - \mathbf{v}_2$ has the property that the endpoint of the vector corresponding to $\mathbf{v}_1 - \mathbf{v}_2$ but located at $\mathbf{v}_2$ is $\mathbf{v}_1$, so that we can use vector methods to describe the triangle with endpoints O, $\mathbf{v}_1$, $\mathbf{v}_2$ (Figure 1.3).

One thing that vectors enable us to do very easily is to describe lines in the plane. For example, if $\mathbf{v}$ is a nonzero vector in the plane, then we have seen that the line L through O and $\mathbf{v}$ is exactly the set

$$\{t\mathbf{v} : t \in \mathbb{R}\}.$$

Here L has O as an origin, and we may think of the point $t\mathbf{v}$ as a variable point on L corresponding to the choice of t. For example $t = 1$ corresponds to the point $\mathbf{v}$, $t = 2$ to the point $2\mathbf{v}$ where the line segment through O and $2\mathbf{v}$ is a line segment lying on L with magnitude $2\|\mathbf{v}\|$, and so on. Thus via the function $f\colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = t\mathbf{v}$, the real number t becomes a coordinate on L, in other words a point of L is specified by the number t.

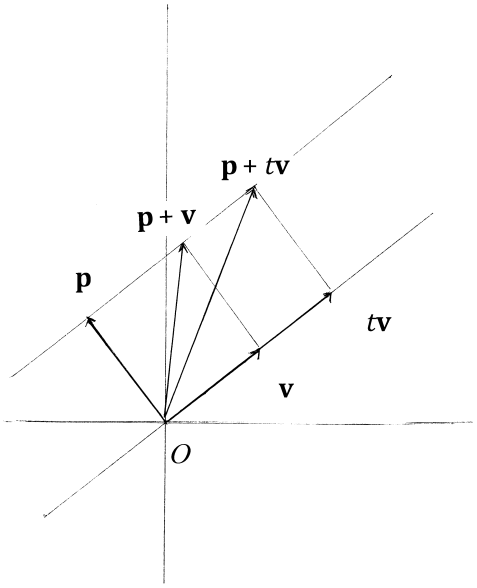

Figure 1.4: A line through the point $\mathbf{p}$

The point $\mathbf{v}$ corresponds to $t = 1$, i.e. to one unit in the "positive" direction. Changing the choice of the point $\mathbf{v}$ would change the scale of the line and also (if we used a negative multiple of $\mathbf{v}$) the choice of a positive direction for the line.

We can use this discussion to describe lines L that do not necessarily pass through O (and this also applies to lines passing through O for which we want to choose some other origin). Let L be such a line, and fix one point $\mathbf{p} \in L$, which will then correspond to our choice of origin. Choose a point $\mathbf{q} \in L$ different from $\mathbf{p}$, so that the vector located at $\mathbf{p}$ with endpoint $\mathbf{q}$ can then be taken to be a unit of measurement on L. Let $\mathbf{v} = \mathbf{q} - \mathbf{p}$. Then the line parallel to L but passing through O is given by taking all scalar multiples of $\mathbf{v}$. If we start with an arbitrary point $t\mathbf{v}$ on this line through O and add to it the vector $\mathbf{p}$, the parallelogram law says that the result will be the point on L pictured in Figure 1.4.

Thus L is described as

$$\{\mathbf{p} + t\mathbf{v} : t \in \mathbb{R}\}.$$

We can describe the same thing by writing down the x and y coordinates as

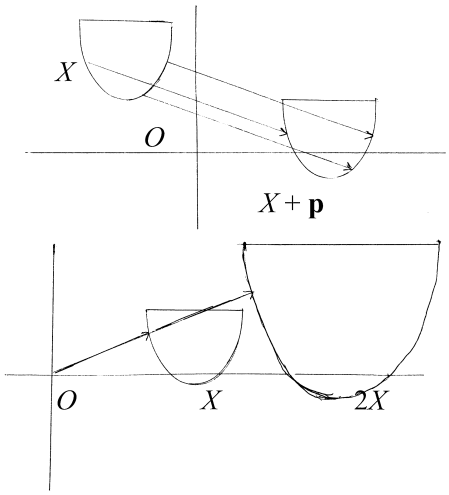

Figure 1.5: Vector addition and scalar multiplication as transformations

well: suppose that $\mathbf{p} = (x_0, y_0)$ and that $\mathbf{v} = (a, b)$. Then

$$L = \{(x, y) : \text{ there exists a } t \in \mathbb{R} \text{ such that } x = x_0 + ta, y = y_0 + tb\}.$$

In this way we can determine the slope and y-intercept of L. Note that this description of L is again as a *parametrized* line: different choices of a point of L correspond to different choices of the parameter t. Moreover $t = 0$ corresponds to $\mathbf{p}$ and $t = 1$ to $\mathbf{q}$.

It is also worth noting that if we substitute in $\mathbf{q} - \mathbf{p}$ for $\mathbf{v}$ and use some of the basic properties of vector addition and scalar multiplication, then we can rewrite the description of L as

$$L = \{t\mathbf{q} + (1 - t)\mathbf{p} : t \in \mathbb{R}\}.$$

Also, the part of the line which is the line segment joining $\mathbf{p}$ to $\mathbf{q}$ is given by $t\mathbf{q} + (1 - t)\mathbf{p}$, $0 \leq t \leq 1$, with the value $t = 0$ corresponding to $\mathbf{p}$ and the value $t = 1$ corresponding to $\mathbf{q}$.

We can give the following geometric interpretation of vector addition and scalar multiplication: vector addition by a fixed vector $\mathbf{p}$ displaces a set X by a parallel motion. In other words, given a set X in $\mathbb{R}^2$, define the set

$X + \mathbf{p}$ to be the set $\{\mathbf{v} + \mathbf{p} : \mathbf{v} \in X\}$. Then $X + \mathbf{p}$ is a parallel copy of X but displaced over by $\mathbf{p}$ (Figure 1.5). Likewise, given $t \in \mathbb{R}$ with $t \neq 0$, then $tX = \{t\mathbf{v} : \mathbf{v} \in X\}$ is similar to X (in the sense of Euclidean geometry but it has been scaled by the scale factor $|t|$ (and it has usually been displaced as well).

Vectors can also be used to give different coordinate systems on $\mathbb{R}^2$. For simplicity we shall just consider the case where we fix the origin O but want to allow different axes, possibly not at right angles, as well as different scales on the axes. There are many reasons why we might want to change coordinate systems. For example, the original choice was arbitrary and we might want to formulate properties that remain unchanged under a rotation of axes (for example, the laws of physics should not depend on our choice). Also, some objects might simplify under a change of coordinates (for example eigenvalues, where we are led to study axes which are not necessarily perpendicular).

To see how to do this, imagine drawing two distinct lines L and L' passing through O, not necessarily perpendicular, and marking off vectors $\mathbf{v}$ and $\mathbf{w}$ on L and L' respectively. Now given any point $\mathbf{p} = (x, y) \in \mathbb{R}^2$, we can draw lines through $\mathbf{p}$ parallel to L and L'. The line through $\mathbf{p}$ parallel to L' meets L in a unique point $t\mathbf{v}$, for some $t \in \mathbb{R}$, and likewise the line through $\mathbf{p}$ parallel t to L meets L' in a unique point $s\mathbf{w}$. From Figure 1.6 it is clear that $\mathbf{p} = t\mathbf{v} + s\mathbf{w}$.

Again referring to Figure 1.6, we see that we can cover the plane with a (not necessarily rectangular) grid whereby we can describe the location of all points (x, y) by their "coordinates" t, s with respect to the grid. So we can describe $\mathbf{p}$ by its Cartesian coordinates x, y with respect to the standard grid or by some other pair (t, s) of real numbers; to say that $\mathbf{p}$ has coordinates t, s with respect to the new grid is just another way of saying that $\mathbf{p} = t\mathbf{v} + s\mathbf{w}$. Geometrically this means that we locate $\mathbf{p}$ by going out t units along L, where unit is based on the scale given by $\mathbf{v}$, and then from there we go s units along a line parallel to L', where unit here refers to the scale given by $\mathbf{w}$. We can also think of this process in terms of a function $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined by $F(t, s) = t\mathbf{v} + s\mathbf{w}$.

The above has another interpretation in terms of linear equations in two unknowns. Setting $\mathbf{v} = (a, b)$ and $\mathbf{w} = (c, d)$, to solve the equation $\mathbf{p} = t\mathbf{v} + s\mathbf{w}$, where $\mathbf{p} = (x, y)$, is to solve the system

$$\begin{aligned} at + cs &= x; \\ bt + ds &= y. \end{aligned}$$

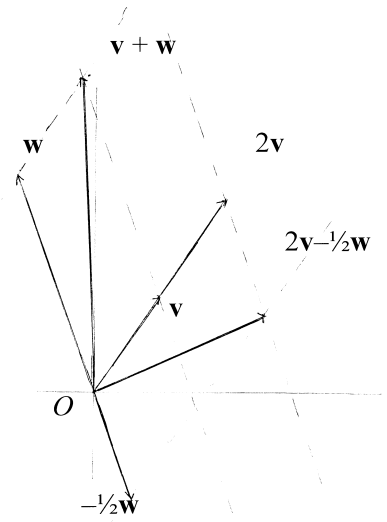

Figure 1.6: Parallelogram coordinates

Up to a choice of letters, this is the usual system of two linear equations in two unknowns, and our geometric reasoning indicates that, for every choice of (x, y), such a system has a unique solution (t, s), provided that the vectors $\mathbf{v}$ and $\mathbf{w}$ really lie on different lines through O. We say that $\mathbf{v}$ and $\mathbf{w}$ are *linearly independent* if they do not both lie on the same line through the origin and *linearly dependent* if they do lie on the same line through the origin, and call an expression of the form $t\mathbf{v} + s\mathbf{w}$ a *linear combination* of $\mathbf{v}$ and $\mathbf{w}$. Thus we have argued geometrically that every vector in $\mathbb{R}^2$ is a linear combination of $\mathbf{v}$ and $\mathbf{w}$ exactly when $\mathbf{v}$ and $\mathbf{w}$ are linearly independent, and in this case it is a linear combination in exactly one way, i.e. the coefficients t, s above are unique. Similarly the geometry suggests that, if $\mathbf{v}$ and $\mathbf{w}$ lie on the *same* line L through O, but are not both $(0, 0)$, then the system has a solution exactly when $\mathbf{p}$ also lies on L (but the solution is definitely not unique). In other words, if $\mathbf{v}$ and $\mathbf{w}$ are linearly dependent but not both the zero vector, then the set of linear combinations of $\mathbf{v}$ and $\mathbf{w}$ is just the line through the origin which contains them both. In this way, a purely algebraic problem has been restated in a geometric way.

We also have the length of a vector $\mathbf{v}$: if $\mathbf{v} = (x, y)$, then its length is $\sqrt{x^2 + y^2}$. We write this quantity as $\|\mathbf{v}\|$. We have the following properties

of length:

1. For all $\mathbf{v} \in \mathbb{R}^2$, $\|\mathbf{v}\| \geq 0$, and $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = (0, 0)$.
2. For all $\mathbf{v} \in \mathbb{R}^2$ and $t \in \mathbb{R}$, $\|t\mathbf{v}\| = |t|\|\mathbf{v}\|$
3. For all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$, $\|\mathbf{v}_1 + \mathbf{v}_2\| \leq \|\mathbf{v}_1\| + \|\mathbf{v}_2\|$.

The only property that is not an easy calculation is the last one; it is called the *triangle inequality*, and essentially says that the sums of the lengths of any two sides of a triangle is greater than the length of the remaining side. We shall discuss this in more detail below.

Before we finish our discussion of $\mathbb{R}^2$, let us also work out angles from the viewpoint of vectors. Let $\mathbf{v}_1 = (x_1, y_1)$ and $\mathbf{v}_2 = (x_2, y_2)$ be two vectors. Let θ be the angle between them, where we take the angle that is less than π. Then we claim that

$$\cos\theta = \frac{x_1x_2 + y_1y_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}.$$

Here the quantity $x_1x_2 + y_1y_2$ in the numerator is important enough to be given a special name: it is denoted $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ or $\mathbf{v}_1 \cdot \mathbf{v}_2$ and is called the *scalar* or *dot* product of $\mathbf{v}_1$ and $\mathbf{v}_2$. Note that this is again an unusual kind of product, since the product (in this sense) of two vectors is not a vector but rather a real number (a scalar, hence the name).

To verify this formula, let us recall the law of cosines: in a triangle with sides of lengths a, b, c, if θ is the angle between the side of length a and that of length b, then $a^2 + b^2 - 2ab\cos\theta = c^2$. Applying this to the triangle with vertices given by the endpoints of $\mathbf{v}_1$ and $\mathbf{v}_2$ gives

$$\|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2 - 2\|\mathbf{v}_1\|\|\mathbf{v}_2\|\cos\theta = \|\mathbf{v}_1 - \mathbf{v}_2\|^2.$$

Rewrite this as

$$\begin{aligned} 2\|\mathbf{v}_1\|\|\mathbf{v}_2\|\cos\theta &= x_1^2 + y_1^2 + x_2^2 + y_2^2 - \left[(x_1 - x_2)^2 + (y_1 - y_2)^2\right] \\ &= 2(x_1x_2 + y_1y_2). \end{aligned}$$

Solving, we get

$$\cos\theta = \frac{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}.$$

In particular we see that $\mathbf{v}_1$ and $\mathbf{v}_2$ are perpendicular if and only if $\cos\theta = 0$ if and only if $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$. For lines in $\mathbb{R}^2$, this is the familiar statement that two lines are perpendicular if the product of their slopes is

-1 (or one is vertical and the other is horizontal). Moreover from the fact that $|\cos\theta| \leq 1$, we derive the Cauchy-Schwarz inequality:

$$|\langle \mathbf{v}_1, \mathbf{v}_2 \rangle| \leq \|\mathbf{v}_1\| \|\mathbf{v}_2\|,$$

which can also be derived by squaring both sides and doing some algebraic manipulations (which we shall do in more generality later).

We note the following formal properties of the inner product:

1. For all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$, $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \langle \mathbf{v}_2, \mathbf{v}_1 \rangle$;
2. For all $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w} \in \mathbb{R}^2$, $\langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{w} \rangle = \langle \mathbf{v}_1, \mathbf{w} \rangle + \langle \mathbf{v}_2, \mathbf{w} \rangle$;
3. For all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$ and $t \in \mathbb{R}$, $\langle t\mathbf{v}_1, \mathbf{v}_2 \rangle = \langle \mathbf{v}_1, t\mathbf{v}_2 \rangle = t\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$;
4. For all $\mathbf{v} \in \mathbb{R}^2$, $\langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|^2$. Thus $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ if and only if, for all $\mathbf{w} \in \mathbb{R}^2$, $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ if and only if $\mathbf{v} = 0$.

Using these properties, it is easy to give a proof of the triangle inequality:

$$\begin{aligned}
\|\mathbf{v}_1 + \mathbf{v}_2\|^2 &= \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2 \rangle \\
&= \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + \langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_1 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\
&= \|\mathbf{v}_1\|^2 + 2\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \|\mathbf{v}_2\|^2 \\
&\leq \|\mathbf{v}_1\|^2 + 2|\langle \mathbf{v}_1, \mathbf{v}_2 \rangle| + \|\mathbf{v}_2\|^2 \\
&\leq \|\mathbf{v}_1\|^2 + 2\|\mathbf{v}_1\| \|\mathbf{v}_2\| + \|\mathbf{v}_2\|^2 = \|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2.
\end{aligned}$$

Vectors in space can be described similarly. The choice of an origin and three mutually perpendicular axes identifies space with $\mathbb{R}^3$, the set of ordered triples of real numbers. After locating a vector at the origin, we can identify a vector $\mathbf{v}$ with a point (x, y, z) of $\mathbb{R}^3$. Experiment again suggests that the correct way to add two vectors is via the parallelogram law. It translates into the following rule: if $\mathbf{v}_1 = (x_1, y_1, z_1)$ and $\mathbf{v}_2 = (x_2, y_2, z_2)$, then

$$\mathbf{v}_1 + \mathbf{v}_2 = (x_1 + x_2, y_1 + y_2, z_1 + z_2).$$

Scalar multiplication $t\mathbf{v} = t(x, y, z) = (tx, ty, tz)$ is defined similarly and has a similar geometric significance. An application of the Pythagorean theorem shows that we should take the length $\|\mathbf{v}\|$ of the vector (x, y, z) to be $\sqrt{x^2 + y^2 + z^2}$, and we define the inner product of two vectors $\mathbf{v}_1 = (x_1, y_1, z_1)$ and $\mathbf{v}_2 = (x_2, y_2, z_2)$ by the formula

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = x_1x_2 + y_1y_2 + z_1z_2$$

as before. A somewhat more involved argument agains shows that the angle θ between $\mathbf{v}_1$ and $\mathbf{v}_2$ is given by the same formula $\cos\theta = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle / \|\mathbf{v}_1\| \|\mathbf{v}_2\|$.

One new feature of $\mathbb{R}^3$ is that, in addition to lines, we also have planes. We first describe the situation for lines and planes through the origin. As in $\mathbb{R}^2$, a line through the origin in $\mathbb{R}^3$ containing the nonzero vector $\mathbf{v}$ is just the set $\{t\mathbf{v} : t \in \mathbb{R}\}$ of scalar multiples of $\mathbf{v}$. If instead we consider the line L passing through a vector $\mathbf{p}$ and parallel to the line through the origin containing $\mathbf{v}$, then L is given in parametric form by

$$L = \{\mathbf{p} + t\mathbf{v} : t \in \mathbb{R}\}.$$

To describe a plane P in $\mathbb{R}^3$ through the origin, first choose a nonzero vector $\mathbf{v} \in P$. Then the plane contains the line through any two points in it, and in particular it contains $t\mathbf{v}$ for all $t \in \mathbb{R}$. By assumption, the plane is not a line, so there is a $\mathbf{w} \in P$ not of the form $t\mathbf{v}$ for any $t \in \mathbb{R}$. In this case, we say that $\mathbf{v}$ and $\mathbf{w}$ are *linearly independent* just as for vectors in $\mathbb{R}^2$. Likewise P will contain all vectors of the form $s\mathbf{w}, s \in \mathbb{R}$, in othe words the line through $\mathbf{w}$ and the origin. Finally P will contain the parallelogram spanned by any two line segments in it which meet at a common point, in other words it will contain $\mathbf{v} + \mathbf{w}$, and in fact it will contain $t\mathbf{v} + s\mathbf{w}$ for all $t, s \in \mathbb{R}$. Such an expression is again called a *linear combination* of $\mathbf{v}$ and $\mathbf{w}$. Geometric reasoning shows that in fact

$$P = \{t\mathbf{v} + s\mathbf{w} : t, s \in \mathbb{R}\}.$$

This is a parametric description for P, using two real parameters s and t. We say that P is the set of *linear combinations* of $\mathbf{v}$ and $\mathbf{w}$ or the plane *spanned by* $\mathbf{v}$ and $\mathbf{w}$.

In general there is no best choice of $\mathbf{v}$ and $\mathbf{w}$. However, if $\mathbf{v}$ and $\mathbf{w}$ both have length one and are perpendicular, one can check that $\|t\mathbf{v} + s\mathbf{w}\| = \sqrt{t^2 + s^2}$, which means that we can measure distance in the plane P by measuring distance in the plane with coordinates (t, s); this would not be true for a more general choice of $\mathbf{v}$ and $\mathbf{w}$.

Finally, the plane Q containing a point $\mathbf{p}$ and parallel to the plane through the origin spanned by $\mathbf{v}$ and $\mathbf{w}$ is given in parametric form by:

$$Q = \{\mathbf{p} + t\mathbf{v} + s\mathbf{w} : t, s \in \mathbb{R}\}.$$

Just as in the plane, if we choose a nonzero vector $\mathbf{v}$, a vector $\mathbf{w}$ not a scalar multiple of $\mathbf{v}$ (in other words, not on the line through the origin and $\mathbf{v}$) and a vector $\mathbf{u}$ which is not a linear combination of $\mathbf{v}$ and $\mathbf{w}$ (in

other words, not on the plane spanned by $\mathbf{v}$ and $\mathbf{w}$), then every vector (x, y, z) in $\mathbb{R}^3$ can be uniquely written as $t\mathbf{v} + s\mathbf{w} + r\mathbf{u}$, in other words as a linear combination of $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{u}$. This fact has an interpretation in terms of solving systems of three linear equations in three unknowns, which is analogous the the case of a system of two equations in two unknowns given in our discussion of vectors in $\mathbb{R}^2$.

One final point here: As we have already observed, a plane P through the origin in $\mathbb{R}^3$ is closed under the vector operations. In other words, for all $\mathbf{a}, \mathbf{b} \in P$, $\mathbf{a} + \mathbf{b} \in P$, and for all $\mathbf{a} \in P$ and $t \in \mathbb{R}$, $t\mathbf{a} \in P$. Of course, we could also see this directly from the point of view of linear combinations: if P is spanned by $\mathbf{v}$ and $\mathbf{w}$ and $\mathbf{a} = t_1\mathbf{v} + s_1\mathbf{w}$, $\mathbf{b} = t_2\mathbf{v} + s_2\mathbf{w}$, then

$$\mathbf{a} + \mathbf{b} = t_1\mathbf{v} + s_1\mathbf{w} + t_2\mathbf{v} + s_2\mathbf{w} = (t_1 + t_2)\mathbf{v} + (s_1 + s_2)\mathbf{w} \in P.$$

Likewise a line through the origin in $\mathbb{R}^3$ or in $\mathbb{R}^2$ is closed under the vector operations. We will call any subset V of $\mathbb{R}^3$ which is closed under the vector operations a *vector subspace* of $\mathbb{R}^3$ (and similarly for $\mathbb{R}^2$). In fact, it is easy to show, assuming the above discussion, that V is a vector subspace of $\mathbb{R}^3$ if and only if V is one of the following:

1. $V = \{(0, 0, 0)\}$;
2. V is a line through the origin;
3. V is a plane through the origin;
4. $V = \mathbb{R}^3$.

Chapter 2

Vectors II

2.1 Vectors in n-dimensional space

We begin with a definition of vectors in n-space. Of course, we cannot visualize n-dimensional space if $n > 3$, but by analogy with the cases $n = 2, 2$ we shall identify n-dimensional space with $\mathbb{R}^n$, i.e. with the set of ordered n-tuples $(x_1, \ldots, x_n)$ of real numbers. As for vectors, we could still think of vectors as arrows with magnitude and direction, but, again based on our experience in $\mathbb{R}^2$ and $\mathbb{R}^3$, we shall just identify a vector with a point of $\mathbb{R}^n$, in other words with an ordered n-tuple $(x_1, \ldots, x_n)$ of real numbers. The *origin* O is the vector $\mathbf{0} = (0, \ldots, 0)$, also called the *zero vector*. (We will try to write $\mathbf{0}$ when we think of the zero vector as a vector, but write it as either $\mathbf{0}$ or O when we think of it as a point.) Given $\mathbf{v} = (x_1, \ldots, x_n)$, the number x_i is the i^{th} *coordinate* or *component* of $\mathbf{v}$. We define vector addition and scalar multiplication in the obvious way: if $\mathbf{v} = (x_1, \ldots, x_n)$ and $\mathbf{w} = (y_1, \ldots, y_n)$ are two vectors in $\mathbb{R}^n$ and $t \in \mathbb{R}$, then we set $\mathbf{v}+\mathbf{w} = (x_1+y_1, \ldots, x_n+y_n)$ and $t\mathbf{v} = (tx_1, \ldots, tx_n)$. These definitions have the following formal properties:

1. For all $\mathbf{v}, \mathbf{w}, \mathbf{u} \in \mathbb{R}^n$, $(\mathbf{v} + \mathbf{w}) + \mathbf{u} = \mathbf{v} + (\mathbf{w} + \mathbf{u})$.
2. For all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$.
3. For all $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v} + \mathbf{0} = \mathbf{v}$.
4. For all $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v} + (-1)\mathbf{v} = \mathbf{0}$.

We denote the vector $(-1)\mathbf{v}$ by $-\mathbf{v}$ and $\mathbf{v} + (-\mathbf{w})$ by $\mathbf{v} - \mathbf{w}$.

Next there are some properties which relate scalar multiplication to vector addition:

1. For all $s, t \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$, $s(t\mathbf{v}) = (st)\mathbf{v}$.

2. For all $s, t \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$, $(s+t)\mathbf{v} = s\mathbf{v} + t\mathbf{v}$.

3. For all $t \in \mathbb{R}$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $t(\mathbf{v} + \mathbf{w}) = t\mathbf{v} + t\mathbf{w}$.

4. For all $\mathbf{v} \in \mathbb{R}^n$, $1 \cdot \mathbf{v} = \mathbf{v}$.

These properties can be checked as in the case of $\mathbb{R}^2$ or $\mathbb{R}^3$.

We also have inner product and length in $\mathbb{R}^n$. The inner product of two vectors $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{y} = (y_1, \ldots, y_k)$ in $\mathbb{R}^k$ is given by the formula $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{k} x_i y_i$ and the length $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. We will discuss the properties of these later. However let us note that inner product is defined via algebraic operations with the components but that length is not a strictly algebraic operation in this sense since it involves taking a square root.

2.2 Abstract vector spaces

Any set V for which there is defined an operation of addition and scalar multiplication satisfying the eight properties of the previous section is called a *vector space*. Here, by the operation of addition we mean a function $F\colon V \times V \to V$, where instead of writing $F(\mathbf{v}, \mathbf{w})$ for the value of F on the pair $(\mathbf{v}, \mathbf{w})$, we simply write $\mathbf{v} + \mathbf{w}$. Scalar multiplication looks a little different: it is given by a function $G\colon \mathbb{R} \times V \to V$, where we denote the value $G(t, \mathbf{v})$ by $t\mathbf{v}$. With this notation, the eight properties then look like:

1. For all $\mathbf{v}, \mathbf{w}, \mathbf{u} \in V$, $(\mathbf{v} + \mathbf{w}) + \mathbf{u} = \mathbf{v} + (\mathbf{w} + \mathbf{u})$. (Vector addition is associative.)

2. For all $\mathbf{v}, \mathbf{w} \in V$, $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$. (Vector addition is commutative.)

3. There exists an element $\mathbf{0} \in V$ such that, for all $\mathbf{v} \in V$, $\mathbf{v} + \mathbf{0} = \mathbf{v}$. (Existence of an additive identity.)

4. For all $\mathbf{v} \in V$, $\mathbf{v} + (-1)\mathbf{v} = \mathbf{0}$. (Existence of additive inverses.)

5. For all $s, t \in \mathbb{R}$ and $\mathbf{v} \in V$, $s(t\mathbf{v}) = (st)\mathbf{v}$. (An analogue of the associative law for multiplication.)

6. For all $s, t \in \mathbb{R}$ and $\mathbf{v} \in V$, $(s+t)\mathbf{v} = s\mathbf{v} + t\mathbf{v}$. (Scalar multiplication distributes over scalar addition.)

7. For all $t \in \mathbb{R}$ and $\mathbf{v}, \mathbf{w} \in V$, $t(\mathbf{v}+\mathbf{w}) = t\mathbf{v}+t\mathbf{w}$. (Scalar multiplication distributes over vector addition.)

8. For all $\mathbf{v} \in \mathbb{R}^n$, $1 \cdot \mathbf{v} = \mathbf{v}$. (A kind of identity law for scalar multiplication.)

As before, we denote the vector $(-1)\mathbf{v}$ by $-\mathbf{v}$ and $\mathbf{v} + (-\mathbf{w})$ by $\mathbf{v} - \mathbf{w}$.

For example, a line or plane through the origin in $\mathbb{R}^3$ is a vector space under the operations of $\mathbb{R}^3$; the main point to check is closure under addition and scalar multiplication. We say that a line or plane through the origin in $\mathbb{R}^3$ is a *vector subspace* of $\mathbb{R}^3$. More generally, we can make the following definition:

Definition 2.1. Let V be a vector space. A subset $W \subseteq V$ which is closed under the operations of vector addition and scalar multiplication (i.e. such that, for all $\mathbf{v}, \mathbf{w} \in W$, $\mathbf{v} + \mathbf{w} \in W$, and for all $t \in \mathbb{R}, \mathbf{v} \in W$, $t\mathbf{v} \in W$) and which satisfies the eight properties above for the inherited operations of addition and scalar multiplication, is a *vector subspace* of V.

For example, $\{\mathbf{0}\}$ and V are always vector subspaces of V. In this course, our main interest will be in vector subspaces of $\mathbb{R}^n$ which we will describe from a different point of view in later in this chapter.

It is worth noting that the conditions in Definition 2.1 are highly redundant. If W is any subset of V which is closed under operations of vector addition and scalar multiplication *and which is nonempty*, or equivalently such that $\mathbf{0} \in W$, then vectors in W automatically satisfy all of the eight properties since they are satisfied by the vectors in V. We leave this as an exercise.

We continue with more examples of abstract vector spaces. The set of all real-valued functions whose domain is $\mathbb{R}$ (or an interval $[a, b]$, or any set X) is a vector space under the natural operations (pointwise addition and multiplication by constant functions i.e. scalars), as is the space of continuous functions with domain $\mathbb{R}$ or an interval. Here the pointwise sum of the real-valued functions f and g is the function $f + g$ defined by $(f+g)(x) = f(x)+g(x)$ for all x in the domain of f and g, and similarly for the function tf. It is easy to check from this that the set of all real-valued functions on $\mathbb{R}$, or on any set X is a vector space, in other words that the eight properties of vector addition and scalar multiplication hold. Note that, in case $X = \{1, \ldots, n\}$ is the set of the first n positive integers, then a function $f\colon \{1, \ldots, n\} \to \mathbb{R}$ can be identified with the n-tuple $(f(1), \ldots, f(n))$ of real numbers, and conversely an n-tuple $(x_1, \ldots, x_n)$ of real numbers defines

a function $f\colon \{1,\dots,n\} \to \mathbb{R}$ via the rule $f(i) = x_i$. So in this case we are just giving a slightly different description of $\mathbb{R}^n$.

Returning to the case of real-valued functions whose domain is $\mathbb{R}$ or an interval, to say that the subset of all *continuous* functions from $\mathbb{R}$ or an interval to $\mathbb{R}$, under the induced operations, is a vector space, is to say that that the sum of two continuous functions is again continuous and every scalar multiple of a continuous function by a constant is continuous. The space of all continuous functions (on an fixed interval, say) is then a vector subspace of the space of all real-valued functions. Similarly the space of all differentiable functions on an open interval is a vector subspace of the space of all real-valued functions on that interval. The content of this statement is that the sum of two differentiable functions is again differentiable and likewise for scalar multiples; in fact, if f and g are differentiable, we know a formula for the derivative of $f + g$, and also for tf if t is a constant. Since every differentiable function is continuous, the space of all differentiable functions on an interval is a vector subspace of the space of all continuous functions on that interval. Likewise the set of all polynomials with real coefficients is a vector space, or the set of all polynomials of degree at most d, and both of these are subspaces of the space of all differentiable functions. (What about the set of all polynomials of degree exactly d?) These vector spaces (except for the space of polynomials of degree at most d) are very large, much larger than $\mathbb{R}^n$ for any n, but they are important because we often want to solve equations for an unknown function. A class of such kinds of equations are differential equations, which are equations for an unknown function f in terms of its derivatives and are of basic importance in most applications of mathematics. It turns out that every abstract vector space arises as a vector subspace of some space of functions. For example, $\mathbb{R}^n$ is identified with the space of functions from the set $\{1,\dots,n\}$ to $\mathbb{R}$.

2.3 Matrices

For another very important example of a vector space, think first of a vector as a set of numbers displayed as a row. There are many other ways to display a set of numbers, for example as a column (in which case we think of the vector as a *column vector*.) Of course, this is just another way of writing an element of $\mathbb{R}^n$. Of other ways to display a set of real numbers, the most

important is as a rectangular array

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kn} \end{pmatrix}.$$

Such an array is called a $k \times n$ *matrix*. The above matrix consists of k rows and n columns. We refer to the number a_{ij} as the $(i,j)^{\text{th}}$ *entry*. This means that a_{ij} is the number in the i^{th} row and j^{th} column. In particular a vector $(x_1, \dots, x_n)$ is also a matrix, in this case a $1 \times n$ matrix. We will call such a matrix a *row vector*. We can also think of a vector as an $n \times 1$ matrix, which we shall refer to as a *column vector*. (We will often have to think of vectors as column vectors because of our conventions on the way we write functions.)

Matrices arise in nature in the following way: suppose that we are given a system of linear equations in n unknowns to solve:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1; \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{k1}x_1 & + & a_{k2}x_2 & + & \cdots & + & a_{kn}x_n & = & b_k. \end{array}$$

Then the system is specified by the matrix A as above. If we use vector shorthand $\mathbf{x}$ for the vector $(x_1, \dots, x_n)$ and $\mathbf{b}$ for $(b_1, \dots, b_k)$, then it would be nice (and suggestive) to write the system above as a single vector/matrix equation

$$A \cdot \mathbf{x} = \mathbf{b}.$$

To do so, we introduce multiplication of matrices. Suppose that A is a $k \times n$ matrix and that B is a $n \times m$ matrix. Then $A \cdot B$ will be a $k \times m$ matrix. To give a formula for $A \cdot B$, suppose that the $(i,\ell)^{\text{th}}$ entry of A is $a_{i\ell}$ and that the $(\ell,j)^{\text{th}}$ entry of B is $b_{\ell j}$. Then the $(i,j)^{\text{th}}$ entry of $A \cdot B$ is $\sum_{\ell=1}^{n} a_{i\ell} b_{\ell j}$. If the rows of A are given by row vectors $\mathbf{r}_1, \dots, \mathbf{r}_k$, where $\mathbf{r}_i \in \mathbb{R}^n$, so that $\mathbf{r}_i = (a_{i1}, \dots, a_{in})$ and the columns of B are given by column vectors $\mathbf{c}_1, \dots, \mathbf{c}_m$, where the $\mathbf{c}_i \in \mathbb{R}^n$ as well, so that $\mathbf{c}_j = (b_{1j}, \dots, b_{nj})$, then the $(i,j)^{\text{th}}$ entry of $A \cdot B$ is given by the inner product $\langle \mathbf{r}_i, \mathbf{c}_j \rangle$. In particular, if A is a $k \times n$ matrix and $\mathbf{x}$ is a vector in $\mathbb{R}^n$, viewed as a column vector, i.e. an $n \times 1$ matrix, then $A \cdot \mathbf{x}$ is a $k \times 1$ matrix and so a column vector again. Thus matrix multiplication gives a shorthand for the system of linear equations above, which will be useful in many different ways.

We can also think of a matrix A as defining, via matrix multiplication, a function from $\mathbb{R}^n$ to $\mathbb{R}^k$: given $A = (a_{ij})$, it defines the function $F \colon \mathbb{R}^n \to \mathbb{R}^k$

given by

$$F(x_1, \ldots, x_n) = (a_{11}x_1 + \cdots + a_{1n}x_n, \ldots, a_{k1}x_1 + \cdots + a_{kn}x_n).$$

Of course these are very special functions from $\mathbb{R}^n$ to $\mathbb{R}^k$. For example, if $n = k = 1$, then A is a 1×1 matrix, in other words a single real number a, and the corresponding function from $\mathbb{R}$ to $\mathbb{R}$ is $f(x) = ax$, in other words the functions we obtain this way are just the linear functions without a constant term.

Note that we can add two $k \times n$ matrices and multiply a $k \times n$ matrix times a scalar t to get another $k \times n$ matrix. Thus the set of all $k \times n$ matrices, denoted $\mathbb{M}_{k,n}$, is naturally a vector space, and it is really the same as $\mathbb{R}^{kn}$. (Sometimes we denote this by $\mathbb{M}_{k,n}(\mathbb{R})$, to denote the fact that we are considering matrices with entries in $\mathbb{R}$. We would define $\mathbb{M}_{k,n}(\mathbb{C})$ or $\mathbb{M}_{k,n}(\mathbb{Q})$ or even $\mathbb{M}_{k,n}(\mathbb{Z})$ in the same way.) The set $\mathbb{M}_{n,n}$ is just denoted $\mathbb{M}_n$ and such a matrix is called a *square matrix*. Note that $\mathbb{M}_{k,n}$ is yet another example of a space of real-valued functions, in this case the space of real-valued functions on the set $\{1, \ldots, k\} \times \{1, \ldots, n\}$ of ordered pairs (i, j) with $1 \leq i \leq k$ and $1 \leq j \leq n$.

Of course, there is something more we can do with matrices: we can (sometimes) multiply them. Matrix multiplication has some important properties which we list below.

1. If $A \in \mathbb{M}_{k,n}, B \in \mathbb{M}_{n,m}, C \in \mathbb{M}_{m,r}$, then $(A \cdot B) \cdot C = A \cdot (B \cdot C)$.
2. If $A \in \mathbb{M}_{k,n}$ and $B, C \in \mathbb{M}_{n,m}$, then $A \cdot (B + C) = A \cdot B + A \cdot C$.
3. If $A, B \in \mathbb{M}_{k,n}$ and $C \in \mathbb{M}_{n,m}$, then $(A + B) \cdot C = A \cdot C + B \cdot C$.
4. If $A \in \mathbb{M}_{k,n}, B \in \mathbb{M}_{n,m}$, and $t \in \mathbb{R}$, then $(tA) \cdot B = A \cdot (tB) = t(A \cdot B)$.

All of these properties can be worked out by brute force and a little patience from the definition, and it is a good exercise for instance to try (1) (the most difficult). Other properies of matrices follow from these, for example one can check that if A is a $k \times n$ matrix then $A \cdot O_{n,m} = O_{k,m}$ and $O_{r,k} \cdot A = O_{r,n}$, where $O_{n,m}$ denotes the $n \times m$ zero matrix. There is also the *identity matrix* I_n which has diagonal entries $a_{ii} = 1$ for $1 \leq i \leq n$ and has all other entries zero. It corresponds to the system

$$\begin{aligned} x_1 &= b_1; \\ &\vdots \\ x_n &= b_n, \end{aligned}$$

or in terms of functions, to the identity function $\mathrm{Id}_{\mathbb{R}^n}$. It is easy to check that, if $A \in \mathbb{M}_{k,n}$, then $I_k \cdot A = A \cdot I_n = A$. So in this sense I_n functions like a unit element for multiplication. Now the only reasonable time we can speak of closure for the operation of matrix multiplication is on the set $\mathbb{M}_n$ of square matrices. Here multiplication is an associative operation and addition of matrices distributes over it; moreover there is a multiplicative identity. However multiplication of matrices is not usually commutative (exercise) and we cannot in general cancel matrix multiplication: it is easy to find examples of matrices A and B in $\mathbb{M}_n$, for all $n \geq 2$, such that $A \cdot B = O$ but neither A nor B is O. In particular, for such matrices A and B, neither A nor B can have an inverse matrix. For example, an inverse matrix for A would be a matrix A^{-1} such that $A^{-1} \cdot A = A \cdot A^{-1} = I_n$. But in this case we would have $B = A^{-1} \cdot A \cdot B = A^{-1} \cdot O = O$, a contradiction. So it will be a special property of square matrices to have an inverse, which we will try to understand in more detail later. The existence of an inverse matrix is connected at least theoretically with the solution of the equation $A \cdot \mathbf{x} = \mathbf{b}$, since if A has an inverse matrix A^{-1}, then necessarily the vector $\mathbf{x}$ solving the equation is given by $\mathbf{x} = A^{-1} \cdot \mathbf{b}$. (This is just a theoretical answer since there are more efficient ways to solve a given system in general than actually finding A^{-1}.)

2.4 Linear independence and span

Let us now return to the problem of generalizing line and plane geometry to $\mathbb{R}^n$. For example, to describe a line L in $\mathbb{R}^n$ through O, it is natural to start with a nonzero vector $\mathbf{v}$ and then consider the set L of all scalar multiples $\{t\mathbf{v} : t \in \mathbb{R}\}$. Similarly, to describe a plane P, start with a nonzero vector $\mathbf{v}$ and then another $\mathbf{w}$ which does not lie on the line L through O and $\mathbf{v}$, or in other words is not a scalar multiple of $\mathbf{v}$. Then take P to be the set $\{t\mathbf{v} + s\mathbf{w} : t, s \in \mathbb{R}\}$. To generalize this to "dimension d" linear geometric objects in $\mathbb{R}^n$, start with d vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^n$. Then take V to be the set

$$V = \{t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d : t_1, \ldots, t_d \in \mathbb{R}\}.$$

Of course, by complete analogy with what we did with lines and planes, we should also make the requirement that $\mathbf{v}_1 \neq \mathbf{0}$, that $\mathbf{v}_2$ is not a scalar multiple of $\mathbf{v}_1$, that $\mathbf{v}_3$ is not a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$, and so on. However it will be useful to allow the $\mathbf{v}_i$ to be completely arbitrary at first, for example some $\mathbf{v}_i$ might be the zero vector or we might for example allow repeated use of the $\mathbf{v}_i$. In fact, let us be a little more precise: Consider a

finite sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^n$, of some length $r \geq 0$ (the case $r = 0$ is the *empty* sequence). A finite sequence is also called a *collection* or *list*. A sequence $\mathbf{v}_1, \ldots, \mathbf{v}_r$ defines a finite *set* of vectors, namely the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. But different sequences can define the same set, both because we can reorder the sequence $\mathbf{v}_1, \ldots, \mathbf{v}_r$ and because it might have repeated terms, i.e. we might have $\mathbf{v}_i = \mathbf{v}_j$, for some $i \neq j$. Throwing out repeated terms and/or reordering will not change the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, but it will change the sequence.

Let us introduce some terminology:

Definition 2.2. Let $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^n$ be a sequence of vectors. A *linear combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is a vector of the form $t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d$, where the t_i are real numbers. The *span* of $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is the set of all linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_d$. Thus

$$\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} = \{t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d : t_i \in \mathbb{R} \text{ for all } i\}.$$

By definition (or logic), $\operatorname{span} \emptyset = \{\mathbf{0}\}$.

We have the following properties of span:

Proposition 2.3. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^n$ be a sequence of vectors.*

1. *For all i, $\mathbf{v}_i \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$.*
2. *If $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ and $t \in \mathbb{R}$, then $t\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$.*
3. *If $\mathbf{v}$ and $\mathbf{w}$ lie in $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, then $\mathbf{v} + \mathbf{w} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$.*
4. *For every $\mathbf{v} \in \mathbb{R}^n$, $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} \subseteq \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\}$.*
5. *For $\mathbf{v} \in \mathbb{R}^n$, $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\} = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ if and only if $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$.*
6. *The set $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ depends only on the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, not on the sequence.*
7. *$\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is a vector subspace of $\mathbb{R}^n$. Moreover, if W is a vector subspace of $\mathbb{R}^n$ containing $\mathbf{v}_1, \ldots, \mathbf{v}_d$, then W contains $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$.*

Proof.

1. Take $t_j = 0, j \neq i$, and $t_i = 1$.
2. Given $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, there exist $t_i \in \mathbb{R}$ such that $\mathbf{v} = t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d$. Then (freely using and extending the basic properties of vector addition and scalar multiplication)

$$t\mathbf{v} = (tt_1)\mathbf{v}_1 + \cdots + (tt_d)\mathbf{v}_d \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}.$$

3. Write $\mathbf{v} = t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d$ and $\mathbf{w} = s_1\mathbf{w}_1 + \cdots + s_d\mathbf{w}_d$. Then

$$\mathbf{v} + \mathbf{w} = (s_1 + t_1)\mathbf{v}_1 + \cdots + (s_d + t_d)\mathbf{w}_d \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}.$$

4. We can write $t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d = t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d + 0 \cdot \mathbf{v}$.

5. We always have $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} \subseteq \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\}$. If

$$\mathbf{v} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\},$$

then every element of $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\}$ is of the form $t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d + t\mathbf{v}$ for some $t \in \mathbb{R}$. By (2) $t\mathbf{v} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, and since the sum of two elements in $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ lies in $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, we must have $t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d + t\mathbf{v} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. Thus

$$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\} \subseteq \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\},$$

and so the two sets are equal since each is contained in the other. Conversely suppose that $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. Then in particular by (1) $\mathbf{v} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}\} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, which is what we wanted to show.

6. Left to you. (Use (5)).

7. It follows from (2) and (3) that $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is a vector subspace of $\mathbb{R}^n$, in other words it is closed under the vector operations and contains $\mathbf{0}$. If W is a vector subspace of $\mathbb{R}^n$ containing $\mathbf{v}_1, \ldots, \mathbf{v}_d$, then the fact that W is closed under the vector operations easily implies that, for all $t_1, \ldots, t_d \in \mathbb{R}$, $t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d \in W$. Hence $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} \subseteq W$.

□

Note that (7) essentially says that $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is the *smallest* vector subspace of $\mathbb{R}^n$ containing all of the $\mathbf{v}_i$. As we shall later see, every vector subspace of $\mathbb{R}^n$ is of the form $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ for some sequence $\mathbf{v}_1, \ldots, \mathbf{v}_d$ of vectors in $\mathbb{R}^n$.

Definition 2.4. A sequence of vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$ such that

$$V = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$$

will be said to *span* V.

Thus to say that $\mathbf{v} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is to say that the vector equation

$$\mathbf{v} = t_1\mathbf{v}_1 + \cdots + t_d\mathbf{v}_d$$

always has a solution in real numbers t_i (not necessarily unique). We will also rewrite this in terms of systems of linear equations shortly.

A vector subspace spanned by a single nonzero vector $\mathbf{v}$ will be called a *line*. Likewise a vector subspace spanned by two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, such that $\mathbf{v}_1 \neq \mathbf{0}$ and $\mathbf{v}_2$ does not lie in the line spanned by $\mathbf{v}_1$, will be called a *plane*. For another example, let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be the standard unit vectors $\mathbf{e}_1 = (1, 0, \ldots, 0), \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \mathbf{e}_n = (0, \ldots, 0, 1)$. Thus all components of $\mathbf{e}_i$ are equal to zero except the i^{th} one, which is 1. (In case $n = 3$, one also uses the notation $\mathbf{e}_1 = \mathbf{i}, \mathbf{e}_2 = \mathbf{j}, \mathbf{e}_3 = \mathbf{k}$.) Then $\text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} = \mathbb{R}^n$, since any vector $\mathbf{v} = (x_1, \ldots, x_n) = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$. More generally, for $i \leq n$,

$$\text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_i\} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_{i+1} = \cdots = x_n = 0\}.$$

The next piece of terminology is there to deal with the fact that we might have chosen a highly redundant set of vectors to span V. For example, if we start out with a nonzero $\mathbf{v} = \mathbf{v}_1$, then $\text{span}\{\mathbf{v}_1, \mathbf{v}_2\} = \text{span}\{\mathbf{v}_1\}$ if $\mathbf{v}_2$ lies on the line spanned by $\mathbf{v}_1$.

Definition 2.5. A sequence $\mathbf{w}_1, \ldots, \mathbf{w}_r \in \mathbb{R}^n$ is *linearly independent* if the following holds: if there exist real numbers t_i such that

$$t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r = \mathbf{0},$$

then $t_i = 0$ for all i. The sequence $\mathbf{w}_1, \ldots, \mathbf{w}_r$ is *linearly dependent* if it is not linearly independent.

Note that the definition of linear independence does **not** depend **only** on the set $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$—if there are any repeated vectors $\mathbf{w}_i = \mathbf{w}_j$, then we can express $\mathbf{0}$ as the nontrivial linear combination $\mathbf{w}_i - \mathbf{w}_j$. Likewise if one of the $\mathbf{w}_i$ is zero then the set is linearly dependent.

By definition or by logic, the empty set is linearly independent. For a less obscure example, $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are linearly independent since if $t_1\mathbf{e}_1 + \cdots + t_n\mathbf{e}_n = \mathbf{0}$, then $(t_1, \ldots, t_n)$ is the zero vector and thus $t_i = 0$ for all i. More generally, for all $j \leq n$, the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_j$ are linearly independent.

Lemma 2.6. *The vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent if and only if, for every $\mathbf{x} \in \mathbb{R}^n$, the equation*

$$\mathbf{x} = t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r$$

has at most one solution in real numbers t_i. Put another way, $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent if and only if, given real numbers t_i and s_i, $1 \leq i \leq r$, such that

$$t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r = s_1\mathbf{w}_1 + \cdots + s_r\mathbf{w}_r,$$

then $t_i = s_i$ for all i.

Proof. If $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent and if $t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r = s_1\mathbf{w}_1 + \cdots + s_r\mathbf{w}_r$, then after subtracting and rearranging we have $(t_1 - s_1)\mathbf{w}_1 + \cdots + (t_r - s_r)\mathbf{w}_r\mathbf{0}$. Thus by the definition of linear independence $t_i - s_i = 0$ for every i, i.e. $s_i = t_i$. Conversely, if the last statement of the lemma holds and if $t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r = \mathbf{0}$, then from

$$t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r = \mathbf{0} = 0 \cdot \mathbf{w}_1 + \cdots + 0 \cdot \mathbf{w}_r,$$

we conclude that $t_i = 0$ for all i. Hence $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent. □

Thus to say that the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent is to say that the vector equation $\mathbf{x} = t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r$ has **at most** one solution in real numbers t_i. Thus to say $\mathbf{x} \in \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ is a statement about the *existence* of a solution to the vector equation $\mathbf{x} = t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r$, while to say that $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent is a statment about the *uniqueness* of a solution to the vector equation $\mathbf{x} = t_1\mathbf{w}_1 + \cdots + t_r\mathbf{w}_r$.

Clearly, if $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are linearly independent, then so is any reordering of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$, and likewise any smaller sequence (not allowing repeats), for example $\mathbf{w}_1, \ldots, \mathbf{w}_s$ with $s \leq r$. A related argument shows:

Lemma 2.7. *The vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$ are **not** linearly independent if and only if we can write at least one of the $\mathbf{w}_i$ as a linear combination of the $\mathbf{w}_j, j \neq i$.*

The proof is left as an exercise.

Definition 2.8. Let V be a vector subspace of $\mathbb{R}^n$. The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are a *basis* of V if they are linearly independent and $V = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. For example, the standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are a basis for $\mathbb{R}^n$.

Thus to say that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are a basis of V is to say that the vector equation $\mathbf{x} = t_1\mathbf{v}_1 + \cdots + t_r\mathbf{v}_r$ has a unique solution in real numbers t_i if and only if $\mathbf{x} \in V$. We can say the same thing in terms of linear equations: given the system

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1; \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{k1}x_1 & + & a_{k2}x_2 & + & \cdots & + & a_{kn}x_n & = & b_k, \end{array}$$

define vectors

$$\mathbf{v}_1 = (a_{11}, a_{21}, \ldots, a_{k1}), \mathbf{v}_2 = (a_{12}, a_{22}, \ldots, a_{k2}), \ldots, \mathbf{v}_n = (a_{1n}, a_{2n}, \ldots, a_{kn}),$$

and set $\mathbf{b} = (b_1, \ldots, b_k)$. Thus the vectors $\mathbf{v}_i$ and $\mathbf{b}$ are in $\mathbb{R}^k$. Then the system of equations has a solution in real numbers $x_1, \ldots, x_n$ if and only if $\mathbf{b} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Such a solution is unique for every $\mathbf{b}$ if and only if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent.

Lemma 2.9 (Main counting argument)**.** *Suppose that* $\mathbf{w}_1, \ldots, \mathbf{w}_b$ *are linearly independent vectors contained in* $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_a\}$*. Then* $b \leq a$*.*

Proof. We shall show that, possibly after relabeling the $\mathbf{v}_i$,

$$\begin{aligned}\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_a\} &= \text{span}\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\} = \text{span}\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_3, \ldots, \mathbf{v}_a\} \\ &= \cdots = \text{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_b, \mathbf{v}_{b+1}, \ldots, \mathbf{v}_a\}.\end{aligned}$$

From this we will be able to conclude that $b \leq a$.

To begin, we may suppose that $\{\mathbf{w}_1, \ldots, \mathbf{w}_b\} \neq \emptyset$. (If $\{\mathbf{w}_1, \ldots, \mathbf{w}_b\} = \emptyset$, then $b = 0$ and the conclusion $b \leq a$ is automatic.) Moreover none of the $\mathbf{w}_i$ is zero. Given $\mathbf{w}_1$, we can write it as a linear combination of the $\mathbf{v}_i$:

$$\mathbf{w}_1 = \sum_{i=1}^{a} t_i \mathbf{v}_i.$$

Since $\mathbf{w}_1 \neq \mathbf{0}$, at least one of the $t_i \neq 0$. After relabeling the $\mathbf{v}_i$, we can assume that $t_1 \neq 0$. Thus we can solve for $\mathbf{v}_1$ in terms of $\mathbf{w}_1$ and the $\mathbf{v}_i, i > 1$:

$$\mathbf{v}_1 = \frac{1}{t_1}\mathbf{w}_1 + \sum_{i=2}^{a} \left(-\frac{t_i}{t_1}\right) \mathbf{v}_i.$$

It follows that $\mathbf{v}_1 \in \text{span}\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\}$. Now using some of the properties of span listed above, we have

$$\text{span}\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\} = \text{span}\{\mathbf{v}_1, \mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\} = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\},$$

where the second equality holds since $\mathbf{w}_1 \in \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\}$.

Continuing in this way, write $\mathbf{w}_2$ as a vector in $\text{span}\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\}$:

$$\mathbf{w}_2 = t_1 \mathbf{w}_1 + \sum_{i=2}^{a} t_i \mathbf{v}_i.$$

For some $i \geq 2$, we must have $t_i \neq 0$, for otherwise we would have $\mathbf{w}_2 = t_1\mathbf{w}_1$ and thus there would exist a nontrivial linear combination $t_1\mathbf{w}_1 + (-1)\mathbf{w}_2 = \mathbf{0}$, contradicting the linear independence of the $\mathbf{w}_i$. After relabeling, we can assume that $t_2 \neq 0$; notice that in particular we must have $a \geq 2$. Arguing as before, we may write

$$\mathbf{v}_2 = -\frac{t_1}{t_2}\mathbf{w}_1 + \frac{1}{t_2}\mathbf{w}_2 + \sum_{i=3}^{a}\left(-\frac{t_i}{t_2}\right)\mathbf{v}_i,$$

and thus we can solve for $\mathbf{v}_2$ in terms of $\mathbf{w}_1, \mathbf{w}_2$, and $\mathbf{v}_i, i \geq 3$ and so

$$\operatorname{span}\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\} = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_3, \ldots, \mathbf{v}_a\}.$$

By induction, for a fixed $i < b$, suppose that we have showed that $i \leq a$ and that after some relabeling of the $\mathbf{v}_i$ we have

$$\begin{gathered}\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_a\} = \operatorname{span}\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_a\} = \\ = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_3, \ldots, \mathbf{v}_a\} = \cdots = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_a\}.\end{gathered}$$

We claim that the same is true for $i + 1$. Write

$$\mathbf{w}_{i+1} = t_1\mathbf{w}_1 + \cdots + t_i\mathbf{w}_i + t_{i+1}\mathbf{v}_{i+1} + \cdots + t_a\mathbf{v}_a,$$

which is possible as

$$\mathbf{w}_{i+1} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_a\} = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_a\}.$$

At least one of the numbers $t_{i+1}, \ldots, t_a$ is nonzero, for otherwise $\mathbf{w}_{i+1} = t_1\mathbf{w}_1 + \cdots + t_i\mathbf{w}_i$, which would say that the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{i+1}$ are not linearly independent. In particular this says that $i+1 \leq a$. After relabeling, we may assume that $t_{i+1} \neq 0$. Then as before we can solve for $\mathbf{v}_{i+1}$ in terms of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{i+1}, \mathbf{v}_{i+2}, \ldots, \mathbf{v}_a$. It follows that

$$\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_a\} = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_i, \mathbf{w}_{i+1}, \mathbf{v}_{i+2}, \ldots, \mathbf{v}_a\}$$

and we have completed the inductive step. So for all $i \leq b$, $i \leq a$, and in particular $b \leq a$. □

This rather complicated argument has the following consequences:

Corollary 2.10. *1. If $\mathbf{w}_1, \ldots, \mathbf{w}_\ell$ are linearly independent vectors in $\mathbb{R}^n$, then $\ell \leq n$.*

2. *Every two bases for a vector subspace V of $\mathbb{R}^n$ have the same number of elements—call this number the* dimension *of V which we write as* $\dim V$. *Thus* $\dim \mathbb{R}^n = n$.

3. *If* $V = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ *then some subset of* $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ *is a basis for V. Hence* $\dim V \leq d$, *and if* $\dim V = d$ *then* $\mathbf{v}_1, \ldots, \mathbf{v}_d$ *is a basis for V.*

4. *If V is a vector subspace of $\mathbb{R}^n$ and $\mathbf{w}_1, \ldots, \mathbf{w}_\ell$ are linearly independent vectors in V, then there exist vectors*

$$\mathbf{w}_{\ell+1}, \ldots, \mathbf{w}_r \in V$$

such that $\mathbf{w}_1, \ldots, \mathbf{w}_\ell, \mathbf{w}_{\ell+1}, \ldots, \mathbf{w}_r$ is a basis for V. Hence $\dim V \geq \ell$, *and if* $\dim V = \ell$ *then $\mathbf{w}_1, \ldots, \mathbf{w}_\ell$ is a basis for V.*

5. *If V_1 and V_2 are two vector subspaces of $\mathbb{R}^n$ and $V_1 \subseteq V_2$, then* $\dim V_1 \leq \dim V_2$. *Moreover* $\dim V_1 = \dim V_2$ *if and only if $V_1 = V_2$. In particular, for every vector subspace V of $\mathbb{R}^n$,* $\dim V \leq n$ *and* $\dim V = n$ *if and only if $V = \mathbb{R}^n$.*

Proof. (1) Apply the lemma to the subspace $V = \mathbb{R}^n$ of $\mathbb{R}^n$ itself, which is the span of $\mathbf{e}_1, \ldots, \mathbf{e}_n$, and to the linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_\ell \in \mathbb{R}^n$, to conclude that $\ell \leq n$.

(2) If $\{\mathbf{w}_1, \ldots, \mathbf{w}_b\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_a\}$ are two bases of V, then by definition $\mathbf{w}_1, \ldots, \mathbf{w}_b$ are linearly independent vectors, and, for all i, $\mathbf{w}_i \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_a\}$. Thus $b \leq a$. But also $\{\mathbf{v}_1, \ldots, \mathbf{v}_a\}$ is a sequence of linearly independent vectors contained in $\operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_b\}$, so $a \leq b$. Thus $a = b$.

(3) If $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is not linearly independent, then, by Lemma 2.7, one of the vectors $\mathbf{v}_i$ is expressed as a linear combination of the others. After relabeling we may assume that $\mathbf{v}_d$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{d-1}$. By general properties,

$$\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{d-1}\} = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}.$$

Continue in this way until it is no longer possible to do so. The remaining vectors will be linearly independent and their span will still be V, so we get a basis of V. By definition the number of left over vectors will be $\dim V$. So this number is at most d and equal to d exactly when we didn't throw any vectors out, in which case $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is a basis.

(4) If $\operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\} \neq V$, then there exists a vector, call it $\mathbf{w}_{\ell+1} \in V$ with $\mathbf{w}_{\ell+1} \notin \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$. It follows that $\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell, \mathbf{w}_{\ell+1}\}$ is still

linearly independent: if there exist real numbers t_i, not all 0, such that $\mathbf{0} = \sum_{i=1}^{\ell+1} t_i \mathbf{w}_i$, then we must have $t_{\ell+1} \neq 0$ since $\mathbf{w}_1, \ldots, \mathbf{w}_\ell$ are linearly independent. But that would say that $\mathbf{w}_{\ell+1} \in \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$, contradicting our choice. So $\mathbf{w}_1, \ldots, \mathbf{w}_\ell, \mathbf{w}_{\ell+1}$ are still linearly independent. We continue in this way. Since the number of elements in a linearly independent sequence of vectors in $\mathbb{R}^n$ is at most n, this procedure has to stop after at most $n - \ell$ stages. At this point we have found a linearly independent sequence which spans V and thus is a basis. To see the last statement, note that if $\operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\} \neq V$, then there exist $\ell + 1$ linearly independent vectors in V, and hence $\dim V \geq \ell + 1$.

(5) Choosing a basis of V_1 and applying (4) (since the elements of a basis are linearly independent) we see that it can be completed to a basis of V_2. Thus $\dim V_1 \leq \dim V_2$. Moreover $\dim V_1 = \dim V_2$ if and only if the basis we chose for V_1 was already a basis for V_2, i.e. $V_1 = V_2$. □

Proposition 2.11. *A subset V of $\mathbb{R}^n$ is a vector subspace of $\mathbb{R}^n$ if and only if it is of the form $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ for some sequence $\mathbf{v}_1, \ldots, \mathbf{v}_d$ of vectors in $\mathbb{R}^n$.*

Proof. We have already noted that a set of the form $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is a vector subspace of $\mathbb{R}^n$. Conversely let V be a vector subspace of $\mathbb{R}^n$. The proof of (3) of the above corollary shows how to find a basis of V. In particular V is of the form $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. □

Let V be an abstract vector space. If $\mathbf{v}_1, \ldots, \mathbf{v}_d \in V$, then we can still define the span of $\mathbf{v}_1, \ldots, \mathbf{v}_d$ or their linear independence, since these can be stated in terms of the vector operations alone. However, it is no longer true that every vector subspace of V is of the form $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ for some $\mathbf{v}_1, \ldots, \mathbf{v}_d \in V$. For example, V itself might not be of this form. We make the following definition:

Definition 2.12. An abstract vector space V is a *finite dimensional vector space* if there exist $\mathbf{v}_1, \ldots, \mathbf{v}_d \in V$ such that $V = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. The vector space V is *infinite dimensional* if it is not finite dimensional.

Thus for example $\mathbb{R}^n$, the space $\mathbb{M}_{k,n}$ of $k \times n$ matrices, and the space of all polynomials with real coefficients of degree at most d are examples of finite dimensional vector spaces. (What are natural choices of bases in the last two examples?) On the other hand, the space of all polynomials, or of all continuous functions on an interval, are not finite dimensional. **Essentially every statement we shall make about $\mathbb{R}^n$ would hold if we replace $\mathbb{R}^n$ by a finite dimensional vector space**. However,

for infinite dimensional vector spaces, many statements no longer hold, and some would have to be modified in order to continue to hold.

2.5 Sums and direct sums of subspaces

Let $V_1, \ldots, V_k$ be vector subspaces of $\mathbb{R}^n$ (or of an arbitrary vector space). The sum of the spaces V_i is analogous to the span of a sequence of vectors:

Definition 2.13. The *sum* $V_1 + \cdots + V_k$ of the subspaces $V_1, \ldots, V_k$ is the set

$$\{v_1 + \cdots + v_k : v_i \in V_i\}.$$

Thus $V_1 + \cdots + V_k$ consists of all possible sums of k vectors $v_1 + \cdots + v_k$ such that the i^{th} vector v_i is in V_i.

Example 2.14. (1) If $L_i = \text{span}\{\mathbf{v}_i\} = \{t\mathbf{v}_i : t \in \mathbb{R}\}$ is the line spanned by $\mathbf{v}_i$, then $L_1 + \cdots + L_k = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.

(2) Similarly, if $V_1 = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and $V_2 = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$, then

$$V_1 + V_2 = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{w}_1, \ldots, \mathbf{w}_\ell\}.$$

Proposition 2.15. *If $V_1, \ldots, V_k$ are vector subspaces of $\mathbb{R}^n$, then $V_1 + \cdots + V_k$ is also a vector subspace of $\mathbb{R}^n$. Moreover, if W is any vector subspace of $\mathbb{R}^n$ such that $V_i \subseteq W$ for all i, then $V_1 + \cdots + V_k \subseteq W$.*

Proof. We leave this as a straightforward exercise. □

The second conclusion of Proposition 2.15 says essentially that $V_1 + \cdots + V_k$ is the *smallest* vector subspace of $\mathbb{R}^n$ containing all of the V_i. In that sense, $V_1 + \cdots + V_k$ is the analogue of the union of the subspaces $V_1, \ldots, V_n$, which is almost never a subspace.

It is natural to ask if there is an analogue of linear independence for subspaces. We begin with the following observation, in the case of two subspaces:

Lemma 2.16. *Let V_1 and V_2 be two subspaces of $\mathbb{R}^n$. Then the following are equivalent:*

1. *Every element of $V_1 + V_2$ is of the form $\mathbf{v}_1 + \mathbf{v}_2$ for unique vectors $\mathbf{v}_1 \in V_1$ and $\mathbf{v}_2 \in V_2$. In other words, if $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{w}_1 + \mathbf{w}_2$, where $\mathbf{v}_1, \mathbf{w}_1 \in V_1$ and $\mathbf{v}_2, \mathbf{w}_2 \in V_2$, then $\mathbf{v}_1 = \mathbf{w}_1$ and $\mathbf{v}_2 = \mathbf{w}_2$.*
2. *If $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{0}$, where $\mathbf{v}_1 \in V_1$ and $\mathbf{v}_2 \in V_2$, then $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{0}$.*

3. $V_1 \cap V_2 = \{\mathbf{0}\}$.

Proof. (1) $\implies$ (2): (2) is the special case of (1) where $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{0}$.
(2) $\implies$ (3): Assume that (2) holds. If $\mathbf{v} \in V_1 \cap V_2$, then $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$, where on the left hand side we view $\mathbf{v}$ as an element of V_1 and $-\mathbf{v}$ as an element of V_2. By (2), we must have $\mathbf{v} = \mathbf{0}$, so that $V_1 \cap V_2 = \{\mathbf{0}\}$.
(3) $\implies$ (1): If (3) holds, i.e. $V_1 \cap V_2 = \{\mathbf{0}\}$, and if $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{w}_1 + \mathbf{w}_2$, where $\mathbf{v}_1, \mathbf{w}_1 \in V_1$ and $\mathbf{v}_2, \mathbf{w}_2 \in V_2$, then $\mathbf{v}_1 - \mathbf{w}_1 = \mathbf{w}_2 - \mathbf{v}_2$. But $\mathbf{v}_1 - \mathbf{w}_1 \in V_1$ and $\mathbf{w}_2 - \mathbf{v}_2 \in V_2$, so that $\mathbf{v}_1 - \mathbf{w}_1 = \mathbf{w}_2 - \mathbf{v}_2 \in V_1 \cap V_2 = \{\mathbf{0}\}$. It follows that $\mathbf{v}_1 - \mathbf{w}_1 = \mathbf{0}$, so that $\mathbf{v}_1 = \mathbf{w}_1$, and similarly $\mathbf{v}_2 = \mathbf{w}_2$. □

Definition 2.17. If V_1 and V_2 are two subspaces of $\mathbb{R}^n$ such that either of the equivalent conditions of Lemma 2.16 hold and $V = V_1 + V_2$, we say that V is the *direct sum* of V_1 and V_2 and write $V = V_1 \oplus V_2$.

The following is then a straightforward argument:

Proposition 2.18. *Suppose that $V_1 \cap V_2 = \{\mathbf{0}\}$. If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a basis for V_1 and $\mathbf{w}_1, \ldots, \mathbf{w}_\ell$ is a basis for V_2, then $\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{w}_1, \ldots, \mathbf{w}_\ell$ is a basis for $V_1 \oplus V_2$. Hence*

$$\dim(V_1 \oplus V_2) = \dim V_1 + \dim V_2.$$

Remark 2.19. For a sum of two subspaces, not necessarily direct, there is a more general formula:

$$\dim(V_1 + V_2) = \dim V_1 + \dim V_2 - \dim(V_1 \cap V_2).$$

While it is easy to prove this directly, we will give a proof in the exercises to the next chapter.

To deal with the concept of a direct sum of more than two subspaces, we have the following:

Proposition 2.20. *Let $V_1, \ldots, V_k$ be subspaces of $\mathbb{R}^n$. Then the following are equivalent:*

1. *Every element of $V_1 + \cdots + V_k$ is of the form $\mathbf{v}_1 + \cdots + \mathbf{v}_k$ for unique vectors $\mathbf{v}_1 \in V_1, \ldots, \mathbf{v}_k \in V_k$. In other words, if $\mathbf{v}_1 + \cdots + \mathbf{v}_k = \mathbf{w}_1 + \cdots + \mathbf{w}_k$, where $\mathbf{v}_i, \mathbf{w}_i \in V_i$ for all i, then $\mathbf{v}_i = \mathbf{w}_i$ for all i.*
2. *If $\mathbf{v}_1 + \cdots + \mathbf{v}_k = \mathbf{0}$, where $\mathbf{v}_i \in V_i$ for all i, then $\mathbf{v}_i = \mathbf{0}$ for all i.*
3. *For all i, if $W_i = V_1 + \cdots + V_{i-1} + V_{i+1} + \cdots + V_k$ is the sum of all of the subspaces V_j with $j \neq i$, then $V_i \cap W_i = \{\mathbf{0}\}$.*

If any of the equivalent conditions above holds, we say that $V_1 + \cdots + V_k = V$ is the *direct sum* of the V_i and write $V = V_1 \oplus \cdots \oplus V_k$. For example, if $L_i = \text{span}\{\mathbf{v}_i\} = \{t\mathbf{v}_i : t \in \mathbb{R}\}$ is the line spanned by $\mathbf{v}_i$, then the sum of the L_i is a direct sum $\iff$ $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent.

2.6 Vector spaces over general fields

The discussion about $\mathbb{R}^n$ and its vector subspaces needed the following basic facts about real numbers, which we use without comment: there are the two basic algebraic operations of addition and multiplication on $\mathbb{R}$, which are technically functions from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$, and which satisfy: both addition and multiplication are commutative and associative, multiplication distributes over addition, there exists an additive identity (0) and additive inverses (the additive inverse of the real number t is $-t$), there exists a multiplicative identity (1), and every nonzero real number has a multiplicative identity. Of course, there are lots of other sets of numbers with these properties. For example, the rational numbers $\mathbb{Q}$ have this property, as do the complex numbers $\mathbb{C}$. Another, less familiar example is the following, defined by analogy with the complex numbers $\mathbb{C}$:

$$\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}.$$

(Here the hard part is checking that multiplicative inverses exist, which follows by the device of "rationalizing the denominator.")

In general, we define a *field* F to be a set with two algebraic operations (traditionally denoted by $+$ and $\cdot$), which satisfy: both $+$ and $\cdot$ are commutative and associative, $\cdot$ distributes over $+$, there exists an additive identity (traditionally denoted by 0) and additive inverses (the additive inverse of an element a of F is usually denoted by $-a$), there exists a multiplicative identity (traditionally denoted by 1), and every element $a \in F$ such that $a \neq 0$ has a multiplicative identity (usually denoted by a^{-1}). Finally, to avoid the trivial case of the field with one element we assume that the additive identity is not equal to the multiplicative identity, i.e. $0 \neq 1$. (It is easy to see by the usual arguments that if the other properties are satisfied then. for all $a \in F$, $0 \cdot a = 0$, so if $0 = 1$ then $F = \{0\}$.) Equivalently, a field always must contain *at least* two elements. A word about closure, which is sometimes thrown in as an axiom about algebraic operations: for us, an algebraic operation such as addition is really a function $F \times F \to F$, whose value on (a, b) is denoted $a + b$. To say that F is closed under addition

is just saying that the range of the addition function is F, which has been built into the definition.

There exist many different kinds of fields. For example, by the remarks above, every field has to have at least two elements. In fact, there exists a field F with exactly two elements. By the above, we must have $F = \{0, 1\}$, and the only question is how to add and multiply. By the definitions and properties of 0 and 1, we must have $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$, and $1 \cdot 1 = 1$. So the only question is how to define $1 + 1$. But if 1 is to have an additive inverse, it can only be 1 and so we must have $1+1 = 0$. Another way to see this is as follows: the only two possibilities for $1 + 1$ are either 0 or 1. But if $1 + 1 = 1$, then by adding an additive inverse for 1 we would be able to cancel it, getting $1 = 0$, which contradicts our assumption. In any case, with these rules for addition and multiplication, one can check that F is a field. There are other examples of finite fields, and it turns out that the number of elements in a finite field is always of the form p^n, where p is a prime number. It is easy to see from the discussion above that every field with just two elements must essentially look like the field $\{0, 1\}$ we have just described. For this reason, we often call $\{0, 1\}$, with the operations described above, *the* field with two elements and write it as $\{0, 1\} = \mathbb{F}_2$.

Remark 2.21. It may look a little strange to write an equation like "$1+1 = 0$." It may help to think that we are just using the familiar symbols 0 and 1 to denote elements in some abstract set F, which have some of the familiar properties of the usual 0 and 1, and we also use the familiar symbol $+$ to denote some abstract way of combining 0 and 1. Some people prefer, at least at the beginning, to use different symbols for 0 and 1, say e and f, and different symbols for the operations, writing for example $e * e$ for addition. Another approach is to decide that 0 and 1 have the usual meaning but that addition and multiplication have been redefined to suit our purposes. In the above example of a field with two elements, the given multiplication is usual multiplication, but addition is different and could be denoted by some new symbol such as $\boxplus$. In fact, in this special example we could also define $\boxplus$ by an explicit closed formula such as $a \boxplus b = |a - b|$, where $|\cdot|$ is the usual absolute value (since $|0 - 0| = |1 - 1| = 0$ and $|0 - 1| = |1 - 0| = 0$).

With this said, let F be any field. As with $\mathbb{R}^n$, we write elements in the n-fold Cartesian product F^n, i.e. ordered n-tuples $(a_1, \ldots, a_n)$ with $a_i \in F$ for all i, as vectors $\mathbf{v}$. Given two elements $\mathbf{v} = (a_1, \ldots, a_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$ in F^n, we add them in the usual way:

$$\mathbf{v} + \mathbf{w} = (a_1, \ldots, a_n) + (w_1, \ldots, w_n) = (a_1 + b_1, \ldots, a_n + b_n).$$

Likewise there is "scalar multiplication" by an element of F: Given $\mathbf{v} = (a_1, \ldots, a_n) \in F^n$ and $a \in F$, then we set $a\mathbf{v} = (aa_1, \ldots, aa_n) \in F^n$. Here the "scalars" are the field F. These two operations of vector addition and scalar multiplication satisfy the same eight properties that we wrote down for $\mathbb{R}^n$. More generally, we can define a *vector space over* F or an F*-vector space* in the same way as for $F = \mathbb{R}$. Then subspaces, span, linear independence, basis, and dimension are all defined as for $\mathbb{R}$.

In this course we shall primarily be concerned with real vector spaces (the case $F = \mathbb{R}$). However, most of the algebraic side of the story works for a general field F. Two examples that are particularly important are 1) the case $F = \mathbb{C}$, where we allow the entries of a vector and the scalars to be complex numbers, and 2) the case $F = \mathbb{F}_2 = \{0, 1\}$ is the field with two elements. In this case an element of $\mathbb{F}_2$ is a string of n 0's and 1's, and a *code* is defined to be a vector subspace of $\mathbb{F}_2$. A few of our results in linear algebra will only make sense for the field of real numbers, for example properties of length and measurement of angles in Chapter 5, which use positivity of a sum of squares and the fact that every positive real number has a (unique) positive square root. On the other hand, a few results that work for the complex numbers $\mathbb{C}$ (such as existence of eigenvalues) do not work for the real numbers. We will try to mention this where appropriate.

Let us say a few more words about complex vector spaces. Recall that a complex number can be written as $a + bi$, where $a, b \in \mathbb{R}$. The formulas for addition and multiplication are as follows:

$$(a_1 + b_1 i) + (a_2 + b_2 i) = (a_1 + a_2) + (b_1 + b_2)i;$$
$$(a_1 + b_1 i)(a_2 + b_2 i) = (a_1 a_2 - b_1 b_2) + (a_1 b_2 + a_2 b_1)i.$$

These are the formulas consistent with $i^2 = -1$. Then one can check that, if $a + bi \neq 0$, then

$$(a + bi)^{-1} = \frac{a}{a^2 + b^2} - \frac{b}{a^2 + b^2} i = \frac{1}{a^2 + b^2}(a - bi).$$

The quantity $a - bi$ which appears in the above formula is important enough to have a special name: if $\alpha = a + bi \in \mathbb{C}$, then its *complex conjugate* $\bar{\alpha}$ is $a - bi$. Straightforward computation then verifies (if $\alpha = a + bi$):

$$\overline{\alpha_1 + \alpha_2} = \bar{\alpha}_1 + \bar{\alpha}_2; \qquad \overline{\alpha_1 \alpha_2} = \bar{\alpha}_1 \bar{\alpha}_2; \qquad \alpha\bar{\alpha} = a^2 + b^2.$$

Moreover, if $\alpha = a + bi$, then $a = \frac{1}{2}(\alpha + \bar{\alpha})$ and $b = \frac{1}{2i}(\alpha - \bar{\alpha})$. We write $a = \operatorname{Re}\alpha$, the *real part* of α, and $b = \operatorname{Im}\alpha$, the *imaginary part* of α. (Note

however that the imaginary part of α is defined to be a real number.) Hence, α is real $\iff \bar{\alpha} = \alpha$.

Applying this to the complex vector space $\mathbb{C}^n$, we see that $\mathbb{R}^n$ is a subset of $\mathbb{C}^n$, closed under vector addition and multiplication by real scalars, but not closed under multiplication by complex scalars (and is thus not a *complex vector subspace* of $\mathbb{C}^n$). If $\mathbf{v} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$, then we can decompose $\mathbf{v}$ into real and imaginary parts by taking the real and imaginary parts a_i, b_i of α_i: setting $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, then $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and $\mathbf{v} = \mathbf{a} + i\mathbf{b}$. Also, if we define $\bar{\mathbf{v}} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$, then

$$\mathbf{a} = \frac{1}{2}(\mathbf{v} + \bar{\mathbf{v}}); \qquad \mathbf{b} = \frac{1}{2i}(\mathbf{v} - \bar{\mathbf{v}}),$$

and $\mathbf{v} \in \mathbb{R}^n \iff \mathbf{v} = \bar{\mathbf{v}}$.

Chapter 3

Linear maps

3.1 Definition of a linear map

Let us begin by recalling some properties of a map (function) $f\colon X \to Y$ from a set X to a set Y. (More details are given in the appendix.) The function f is *one-to-one* or *injective* if, for all elements x_1 and x_2 of X, $f(x_1) = f(x_2)$ if and only if $x_1 = x_2$. The function f is *onto* or *surjective* if for all $y \in Y$, there exists an $x \in X$ such that $f(x) = y$. Thus being one-to-one asserts that, given $y \in Y$, there is at most one solution to the equation

$$f(x) = y,$$

i.e. solutions may not exist but are unique if they do, whereas being onto asserts that the equation always has at least one solution (which may not be unique). A function is a *one-to-one correspondence* (or a *bijection*) if it is both one-to-one and onto: this says that for every $y \in Y$ there is a unique $x \in X$ such that $f(x) = y$. A function $f\colon X \to Y$ is one-to-one if and only if it has a left inverse, i.e. a function $g\colon Y \to X$ such that $g \circ f = \mathrm{Id}_X$. A function $f\colon X \to Y$ is onto if and only if it has a right inverse, i.e. a function $h\colon Y \to X$ such that $f \circ h = \mathrm{Id}_Y$. A function $f\colon X \to Y$ is a one-to-one correspondence if and only if it has both a left and a right inverse, which are then necessarily equal.

In general we can always make a function $f\colon X \to Y$ onto by restricting its range to the set of values it actually takes: define $\mathrm{Im}\, f \subseteq Y$ by

$$\mathrm{Im}\, f = \{y \in Y : \text{ there exists an } x \in X \text{ with } f(x) = y\}.$$

Thus f is onto Y if and only if $\mathrm{Im}\, f = Y$, and a function is always onto its image. Here we are being mathematically imprecise, since technically we

must distinguish between the function $f\colon X \to Y$ with values in Y and the function $f_0\colon X \to \operatorname{Im} f$ with values in $\operatorname{Im} f$, but which is otherwise defined "by the same formula," i.e. $f_0(x) = f(x)$ for all $x \in X$.

With this said, we can now define linear maps. The idea is, given an $n \times k$ matrix A, to think of the matrix multiplication $A \cdot \mathbf{x}$ as defining a function (or in other words a map) from $\mathbb{R}^k$ to $\mathbb{R}^n$. However, we shall set things up a little differently: Fix vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$–these are the columns of A. Define the function $F\colon \mathbb{R}^k \to \mathbb{R}^n$ by the formula:

$$F(t_1, \ldots, t_k) = \sum_{i=1}^{k} t_i \mathbf{v}_i.$$

Thus $F(t_1, \ldots, t_k)$ is the linear combination of the $\mathbf{v}_i$ with coefficients given by the t_i.

Example 3.1. Here are some basic examples of maps of the type we are considering:

1. The map $F\colon \mathbb{R}^k \to \mathbb{R}^n$ defined by $F(t_1, \ldots, t_k) = \mathbf{0}$ for all $(t_1, \ldots, t_k) \in \mathbb{R}^k$. Here $\mathbf{v}_i = \mathbf{0}$ for all i.

2. The map $F\colon \mathbb{R}^k \to \mathbb{R}^k$ defined by $F(t_1, \ldots, t_k) = (t_1, \ldots, t_k)$ for all $(t_1, \ldots, t_k) \in \mathbb{R}^k$. Here $F = \mathrm{Id} = \mathrm{Id}_{\mathbb{R}^k}$ and $\mathbf{v}_i = \mathbf{e}_i$ for all i, $1 \leq i \leq k$.

3. If $n \leq k$, the map $F\colon \mathbb{R}^k \to \mathbb{R}^n$ defined by $F(t_1, \ldots, t_k) = (t_1, \ldots, t_n)$ for all $(t_1, \ldots, t_k) \in \mathbb{R}^k$. Here $\mathbf{v}_i = \mathbf{e}_i$ for all i, $1 \leq i \leq n$ (where the $\mathbf{e}_i \in \mathbb{R}^n$), and $\mathbf{v}_i = \mathbf{0}$ for $i > n$. The map F is called *projection from $\mathbb{R}^k$ to $\mathbb{R}^n$*. Of course, we could have singled out any subset I of $\{1, \ldots, k\}$ with n elements, not necessarily the first n elements, and defined projection from $\mathbb{R}^k$ to $\mathbb{R}^n$ by setting $F(t_1, \ldots, t_k)$ to be the vector whose components are the t_i, $i \in I$ (in some fixed order, for example increasing order).

4. If $n \geq k$, the map $F\colon \mathbb{R}^k \to \mathbb{R}^n$ defined by

$$F(t_1, \ldots, t_k) = (t_1, \ldots, t_k, 0, \ldots, 0)$$

for all $(t_1, \ldots, t_k) \in \mathbb{R}^k$. Here $\mathbf{v}_i = \mathbf{e}_i$ for all i, $1 \leq i \leq k$ (where the $\mathbf{e}_i \in \mathbb{R}^n$). The map F is called the *inclusion of $\mathbb{R}^k$ in $\mathbb{R}^n$*. Of course, there are variations on this construction as in the case of projection.

Let $F\colon \mathbb{R}^k \to \mathbb{R}^n$ be defined via $F(t_1, \ldots, t_k) = \sum_{i=1}^{k} t_i \mathbf{v}_i$. We observe the following easy properties of F:

1. $F(\mathbf{0}) = \mathbf{0}$.
2. $F(\mathbf{e}_i) = \mathbf{v}_i$ for every standard basis vector $\mathbf{e}_i$.
3. Let $\mathbf{w} = (t_1, \ldots, t_k) \in \mathbb{R}^k$ and let $\mathbf{u} = (s_1, \ldots, s_k) \in \mathbb{R}^k$. Then
$$F(\mathbf{w} + \mathbf{u}) = F(\mathbf{w}) + F(\mathbf{u}),$$
where the addition $\mathbf{w} + \mathbf{u}$ takes place in $\mathbb{R}^k$ and the addition $F(\mathbf{w}) + F(\mathbf{u})$ takes place in $\mathbb{R}^n$.
4. Let $\mathbf{w} = (t_1, \ldots, t_k) \in \mathbb{R}^k$ and let $t \in \mathbb{R}$. Then
$$F(t\mathbf{w}) = tF(\mathbf{w}).$$

Note that (2) says that we can recover the vectors $\mathbf{v}_i$ used to define F from the knowledge of F as a function.

For general vector spaces, we will define functions that behave like the functions described above:

Definition 3.2. Let $F\colon \mathbb{R}^k \to \mathbb{R}^n$ be a function. Then F is *linear* if it satisfies the following two properties:

(i) For all $t \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^k$, $F(t\mathbf{v}) = tF(\mathbf{v})$;

(ii) For all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^k$, $F(\mathbf{v} + \mathbf{w}) = F(\mathbf{v}) + F(\mathbf{w})$.

Proposition 3.3. *A function $F\colon \mathbb{R}^k \to \mathbb{R}^n$ is linear if and only if it is of the form $F(t_1, \ldots, t_k) = \sum_{i=1}^k t_i \mathbf{v}_i$ for some vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$. In this case the $\mathbf{v}_i$ are given by the formula $F(\mathbf{e}_i) = \mathbf{v}_i$.*

Proof. We have seen that if F is of the form $F(t_1, \ldots, t_k) = \sum_{i=1}^k t_i \mathbf{v}_i$ for some vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, then it is linear and the $\mathbf{v}_i$ are given by the formula $F(\mathbf{e}_i) = \mathbf{v}_i$. Conversely, suppose that F is linear. Let $\mathbf{v}_i = F(\mathbf{e}_i)$. Then every vector $(t_1, \ldots, t_k) \in \mathbb{R}^k$ can be written in terms of the standard basis: $(t_1, \ldots, t_k) = \sum_{i=1}^n t_i \mathbf{e}_i$. Repeated use of the properties (i) and (ii) of linear functions enables us to expand out:

$$\begin{aligned} F(t_1, \ldots, t_k) &= F(\sum_{i=1}^n t_i \mathbf{e}_i) \\ = \sum_{i=1}^n F(t_i \mathbf{e}_i) &= \sum_{i=1}^n t_i F(\mathbf{e}_i) = \sum_{i=1}^n t_i \mathbf{v}_i. \end{aligned}$$

Thus F is as described. □

For general vector spaces, there are functions which satisfy the definition of being a linear map but which are not given by taking a linear combination of a finite number of vectors. For example, the derivative or a definite integral are linear functions on appropriate vector spaces of functions but they are not specified by a finite number of vectors (functions). Thus Definition 3.2 is the correct definition for an abstract vector space.

Roughly speaking, a function $F\colon \mathbb{R}^k \to \mathbb{R}^n$ is linear if it involves the variables t_i to the first power only (with no constant terms). For example, in our definition the only linear functions from $\mathbb{R}$ to $\mathbb{R}$ are of the form $f(x) = ax$ for some $a \in \mathbb{R}$; the function $f(x) = ax + b$ with $b \neq 0$ is no longer called a linear function. A function involving both the variables t_i to the first power and constants is sometimes called an *affine linear function*.

Proposition 3.4. *Let $F(t_1, \ldots, t_k) = \sum_{i=1}^{k} t_i \mathbf{v}_i$ be a linear map.*

1. *F is onto $\mathbb{R}^n$ if and only if the $\mathbf{v}_i$ span $\mathbb{R}^n$. More generally, the image $\operatorname{Im} F$ is always equal to the span of the $\mathbf{v}_i$. In particular, $\operatorname{Im} F$ is always a vector subspace of $\mathbb{R}^n$.*

2. *F is one-to-one if and only if the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent.*

3. *F is one-to-one and onto if and only if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are a basis for $\mathbb{R}^n$. In this case we necessarily have $k = n$.*

4. *Suppose that F is a linear map from $\mathbb{R}^n$ to itself (i.e. $k = n$ in the notation above). Then F is a 1-1 correspondence $\iff$ F is one to one $\iff$ F is onto $\iff$ F has an inverse map F^{-1}.*

Proof. (1) is a matter of the definition of $\operatorname{Im} F$ and the span of the $\mathbf{v}_i$. (2) says that if the $\mathbf{v}_i$ are linearly independent, then a vector in $\mathbb{R}^n$ is a linear combination of the $\mathbf{v}_i$ in at most one way, and conversely. (3) follows by putting (1) and (2) together.

To see (4), first note that F has an inverse function if and only if F is a one-to-one correspondence (this has nothing to do with F being linear). Also F is one-to-one and onto if and only if the $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are a basis for $\mathbb{R}^n$. Moreover if F is one-to-one, then the $\mathbf{v}_i$ are linearly independent. Since there are n of them, and they are vectors in $\mathbb{R}^n$, they must automatically span $\mathbb{R}^n$ and thus be a basis. Hence F is onto. A similar argument shows that F is onto implies that it is one-to-one. □

Property (4) above only holds in the context of $\mathbb{R}^n$, or of finite-dimensional vector spaces, in the sense that if V is an infinite-dimensional vector space,

then there exist linear maps $F\colon V \to V$ which are injective but not surjective, and linear maps $F\colon V \to V$ which are surjective but not injective.

A linear function $F\colon \mathbb{R}^n \to \mathbb{R}^n$ has a right inverse $\iff$ F has a left inverse, and the two must agree. This is a very special property of *linear* functions from $\mathbb{R}^n$ to itself; a general function would not have this property. In case a linear function has an inverse, the inverse function is automatically linear:

Proposition 3.5. *Let $F\colon \mathbb{R}^n \to \mathbb{R}^n$ be a linear function, and suppose that F^{-1} is an inverse function for F. Then F^{-1} is again a linear map.*

Proof. Using the previous proposition, we see that we must show that, for all $\mathbf{w}$ and $\mathbf{u}$ in $\mathbb{R}^n$, $F^{-1}(\mathbf{w}+\mathbf{u}) = F^{-1}(\mathbf{w}) + F^{-1}(\mathbf{u})$, and similarly $F^{-1}(t\mathbf{w}) = tF^{-1}(\mathbf{w})$ for all $t \in \mathbb{R}$. To see the first equality, use the fact that F is linear to obtain

$$F\big(F^{-1}(\mathbf{w}) + F^{-1}(\mathbf{u})\big) = F\big(F^{-1}(\mathbf{w})\big) + F\big(F^{-1}(\mathbf{u})\big).$$

By definition of an inverse function, $F\big(F^{-1}(\mathbf{w})\big) = \mathbf{w}$ and $F\big(F^{-1}(\mathbf{u})\big) = \mathbf{u}$. Thus

$$F\big(F^{-1}(\mathbf{w}) + F^{-1}(\mathbf{u})\big) = \mathbf{w} + \mathbf{u}.$$

Again by the definition of an inverse function, this says that

$$F^{-1}(\mathbf{w}) + F^{-1}(\mathbf{u}) = F^{-1}(\mathbf{w}+\mathbf{u}).$$

The argument for scalar multiplication is similar. □

In general, if the $\mathbf{v}_i$ are linearly independent, let V be the vector subspace that they span. Then $F\colon \mathbb{R}^k \to V$ is a 1-1 correspondence from $\mathbb{R}^k$ to V. In this way, a vector subspace of dimension k looks just like $\mathbb{R}^k$, and the linear property says that this 1-1 correspondence preserves the basic vector operations (although not in general length or angle). We can think of the function F as parametrizing the subspace V, and the real numbers t_i as coordinates which describe the points of V. In general, given two vector spaces V and W, a bijective linear map $F\colon V \to W$ is called an *isomorphism* or *linear isomorphism* from V to W and two vector spaces V and W are *isomorphic* if there exists an isomorphism $F\colon V \to W$. For example, $\mathbb{R}^n$ and $\mathbb{R}^m$ are isomorphic $\iff$ $n = m$. (Why?) For another example, a vector subspace of $\mathbb{R}^n$ of dimension d is isomorphic to $\mathbb{R}^d$. The vector space of all polynomials of degree at most d is isomorphic to $\mathbb{R}^{d+1}$.

3.2 Linear functions and matrices

Given a linear function $F\colon \mathbb{R}^n \to \mathbb{R}^k$ with $F(x_1, \ldots, x_n) = \sum_i x_i \mathbf{v}_i$, recall that we can associate an $k \times n$ matrix to F as follows: write the vectors $\mathbf{v}_i = (a_{1i}, \ldots, a_{ki})$. Then to F we associate the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kn} \end{pmatrix}.$$

Here the columns of A are the vectors $\mathbf{v}_i$, written vertically, and the linear map $F(x_1, \ldots, x_n)$ corresponds to the matrix product $A \cdot \mathbf{x}$, where $A \cdot \mathbf{x}$ is the $n \times 1$ matrix (column vector) whose j^{th} entry is $\sum_{i=1}^n a_{ji} x_i$. In particular $A \cdot \mathbf{e}_i = \mathbf{v}_i$, written as a column vector; its j^{th} entry is a_{ji} and it is equal to $\sum_{j=1}^k a_{ji} \mathbf{e}_j$, where in the equality

$$A \cdot \mathbf{e}_i = \sum_{j=1}^{k} a_{ji} \mathbf{e}_j$$

the $\mathbf{e}_i$ on the left is a basis vector in $\mathbb{R}^n$ and the $\mathbf{e}_j$ on the right is a basis vector in $\mathbb{R}^k$. Note the reversal of the indices! The case $F\colon \mathbb{R}^n \to \mathbb{R}^n$ corresponds to square $(n \times n)$ matrices. More generally we recall that we can multiply two matrices A and B as long as A is an $k \times n$ matrix and B is a $m \times k$ matrix, in the order $B \cdot A$. Here $B \cdot A$ will be an $m \times n$ matrix. Suppose that the $(i, \ell)^{\text{th}}$ entry of B is $b_{i\ell}$ and that the $(\ell, j)^{\text{th}}$ entry of A is $a_{\ell,j}$. Then the $(i, j)^{\text{th}}$ entry of $B \cdot A$ is $\sum_{\ell=1}^k b_{i\ell} a_{\ell j}$.

The importance of matrix multiplication comes from its connection with function composition. If $F\colon \mathbb{R}^n \to \mathbb{R}^k$ is a linear map and $G\colon \mathbb{R}^k \to \mathbb{R}^m$ is a linear map, then we have the composition $G \circ F\colon \mathbb{R}^n \to \mathbb{R}^m$.

Proposition 3.6. *If $F\colon \mathbb{R}^n \to \mathbb{R}^k$ and $G\colon \mathbb{R}^k \to \mathbb{R}^m$ are linear maps, and A and B are the matrices corresponding to F and G respectively, then $G \circ F$ is again linear and the matrix corresponding to $G \circ F$ is the matrix product $B \cdot A$.*

Proof. Let us first verify that $G \circ F$ is linear. For example, we have

$$\begin{aligned} G \circ F(\mathbf{w} + \mathbf{u}) &= G(F(\mathbf{w} + \mathbf{u})) \\ = G(F(\mathbf{w}) + F(\mathbf{u})) &= G(F(\mathbf{w})) + G(F(\mathbf{u})) = G \circ F(\mathbf{w}) + G \circ F(\mathbf{u}). \end{aligned}$$

The check for scalar multiplication is similar.

To verify that $B \cdot A$ corresponds to $G \circ F$, note the following: we have $F(\mathbf{e}_i) = \sum_{\ell=1}^{k} a_{\ell i}\mathbf{e}_\ell$ and $G(\mathbf{e}_\ell) = \sum_{j=1}^{m} b_{j\ell}\mathbf{e}_j$. Thus using linearity and expanding out,

$$\begin{aligned} G \circ F(\mathbf{e}_i) &= G\left(\sum_{\ell=1}^{k} a_{\ell i}\mathbf{e}_\ell\right) = \sum_{\ell=1}^{k} a_{\ell i} G(\mathbf{e}_\ell) \\ &= \sum_{\ell=1}^{k} a_{\ell i}\left(\sum_{j=1}^{m} b_{j\ell}\mathbf{e}_j\right) = \sum_{j=1}^{m}\left(\sum_{\ell=1}^{k} b_{j\ell}a_{\ell i}\right)\mathbf{e}_j, \end{aligned}$$

after a little thought about the double sums. This says (using the fact that the indices get reversed) that $G \circ F$ corresponds to the matrix with $(i,j)^{\text{th}}$ entry equal to $b_{i\ell}a_{\ell j}$, in other words to $B \cdot A$. □

As a consequence, we have another (and cleaner) proof that matrix multiplication is associative, and the other properties of matrix multiplication can also be checked by showing the corresponding properties for linear maps.

3.3 Image and kernel

Definition 3.7. Let $F\colon \mathbb{R}^n \to \mathbb{R}^k$ be a linear map. The *kernel* of F (written $\operatorname{Ker} F$) is the set

$$\{\mathbf{w} \in \mathbb{R}^n : F(\mathbf{w}) = \mathbf{0}\}.$$

The image of F is of course the set of $\mathbf{v} \in \mathbb{R}^k$ such that there is a $\mathbf{w} \in \mathbb{R}^n$ with $F(\mathbf{w}) = \mathbf{v}$. Thus if $F(t_1, \ldots, t_n) = \sum_{i=1}^{n} t_i\mathbf{v}_i$, the image of F is the span of the $\mathbf{v}_i$ and so is a vector subspace of $\mathbb{R}^k$. By contrast, we can think of the kernel of F as the set of those scalars $(t_1, \ldots, t_n)$ for which the linear combination $\sum_i t_i\mathbf{v}_i = \mathbf{0}$. Thus the kernel describes all the ways we can write $\mathbf{0}$ as a linear combination of the $\mathbf{v}_i$. It is also the set of solutions to the matrix equation $A \cdot \mathbf{x} = \mathbf{0}$. We have the following:

Proposition 3.8. *The kernel of F is a vector subspace of $\mathbb{R}^n$. Moreover, F is one-to-one if and only if $\operatorname{Ker} F = \{\mathbf{0}\}$.*

Proof. Clearly $F(\mathbf{0}) = \mathbf{0}$. If $F(\mathbf{w}) = \mathbf{0}$ and $F(\mathbf{u}) = \mathbf{0}$, then

$$F(\mathbf{w} + \mathbf{u}) = F(\mathbf{w}) + F(\mathbf{u}) = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

Thus the kernel of F is closed under vector addition. Likewise it is closed under scalar multiplication. So it is a vector subspace. The last statement is a homework problem. □

Thus to every linear map we have two important associated subspaces, the image and the kernel. (Of course, just knowing the image and the kernel does not tell us what the linear map is.) As we will see, every vector subspace of $\mathbb{R}^n$ is the kernel of a linear map. In this way we have defined vector subspaces by the vector equation $\{\mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) = \mathbf{0}\}$, which is the same thing as k linear equations, i.e. implicitly instead of parametrically. This is of course another (and very familiar) way to think about lines and planes in $\mathbb{R}^2$ or $\mathbb{R}^3$: a line through the origin in $\mathbb{R}^2$ is given by an equation $Ax + By = 0$, a plane through the origin in $\mathbb{R}^3$ is given by an equation $Ax + By + Cz = 0$, and a line through the origin in $\mathbb{R}^3$ is given by *two* equations $A_1x + B_1y + C_1z = 0$ and $A_2x + B_2y + C_2z = 0$, where the vectors (A_1, B_1, C_1) and (A_2, B_2, C_2) are linearly independent.

One reason for the importance of the kernel is the following. Suppose we want to write down all solutions to the vector equation $F(\mathbf{x}) = \mathbf{b}$. To decide if there are any solutions is to decide if $\mathbf{b} \in \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ (if $F(t_1, \dots, t_n) = \sum_{i=1}^n t_i\mathbf{v}_i$). If there is one particular solution $\mathbf{x}_0$, then the set of all solutions is exactly the set

$$\{\mathbf{x}_0 + \mathbf{w} : \mathbf{w} \in \operatorname{Ker} F\}.$$

Indeed, if $\mathbf{x}$ also satisfies $F(\mathbf{x}) = \mathbf{b}$, then $F(\mathbf{x}) = F(\mathbf{x}_0)$ and thus $F(\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$. So $\mathbf{x} - \mathbf{x}_0 \in \operatorname{Ker} F$, say $\mathbf{x} - \mathbf{x}_0 = \mathbf{w}$. Thus $\mathbf{x} = \mathbf{x}_0 + \mathbf{w}$ for some $\mathbf{w} \in \operatorname{Ker} F$. Conversely, if $\mathbf{x} = \mathbf{x}_0 + \mathbf{w}$ for some $\mathbf{w} \in \operatorname{Ker} F$, then $F(\mathbf{x}) = F(\mathbf{x}_0 + \mathbf{w}) = F(\mathbf{x}_0) + f(\mathbf{w}) = \mathbf{b} + \mathbf{0} = \mathbf{b}$.

Summarizing, then, the image of F describes for which vectors $\mathbf{b}$ the equation $F(\mathbf{x}) = \mathbf{b}$ has a solution, whereas the kernel describes the set of all possible solutions to $F(\mathbf{x}) = \mathbf{b}$, assuming that one exists. Note that a subset of $\mathbb{R}^n$ of the form $\mathbf{x}_0 + \operatorname{Ker} F$ is not a vector subspace of $\mathbb{R}^n$, but rather a vector subspace plus a fixed vector, which we can think of as a geometric object "parallel" to the vector subspace but which does not necessarily pass through the origin. Such an object, namely a subset of $\mathbb{R}^n$ of the form $\mathbf{x}_0 + W$, where W is a vector subspace of $\mathbb{R}^n$, is called an *affine subspace* of $\mathbb{R}^n$. It can be described by saying that it contains the line through any two points in it. The solution set to $F(\mathbf{x}) = \mathbf{b}$ is then an affine subspace of $\mathbb{R}^n$.

Finally we have the following formula:

Proposition 3.9. *If $F\colon \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, then*

$$\dim \operatorname{Ker} F + \dim \operatorname{Im} F = n.$$

Proof. Let $\mathbf{w}_1, \dots, \mathbf{w}_a$ be a basis for $\operatorname{Ker} F$ and let $\mathbf{v}'_1, \dots, \mathbf{v}'_b$ be a basis for $\operatorname{Im} F$. By definition of $\operatorname{Im} F$, for every i, there exist $\mathbf{v}_i \in \mathbb{R}^n$ such that

$F(\mathbf{v}_i) = \mathbf{v}'_i$. We claim that $\mathbf{w}_1, \ldots, \mathbf{w}_a, \mathbf{v}_1, \ldots, \mathbf{v}_b$ is a basis for $\mathbb{R}^n$. By the fact that every basis has the same number of elements, it will then follow that $a + b = n$, so by definition $\dim \operatorname{Ker} F + \dim \operatorname{Im} F = n$.

To see the claim, we shall show that $\mathbf{w}_1, \ldots, \mathbf{w}_a, \mathbf{v}_1, \ldots, \mathbf{v}_b$ span $\mathbb{R}^n$ and are linearly independent. To see that they span $\mathbb{R}^n$, let $\mathbf{x} \in \mathbb{R}^n$. Then $F(\mathbf{x}) \in \operatorname{Im} F$, so that there exist $t_1, \ldots, t_b \in \mathbb{R}$ such that $F(\mathbf{x}) = \sum_{i=1}^{b} t_i \mathbf{v}'_i$. (By choice of the $\mathbf{v}'_i$, they are a basis for $\operatorname{Im} F$.) Now consider $\mathbf{x} - \sum_i t_i \mathbf{v}_i$. By construction $F(\mathbf{x}) = F(\sum_i t_i \mathbf{v}_i)$, and so $\mathbf{x} - \sum_i t_i \mathbf{v}_i \in \operatorname{Ker} F$. Thus there are $s_i \in \mathbb{R}$ such that $\mathbf{x} - \sum_i t_i \mathbf{v}_i = \sum_{i=1}^{a} s_i \mathbf{w}_i$. It follows that $\mathbf{x} = \sum_i t_i \mathbf{v}_i + \sum_{i=1}^{a} s_i \mathbf{w}_i$, and thus that $\mathbf{x}$ is in the span of the $\mathbf{w}_i$ and $\mathbf{v}_j$.

Finally we show that $\mathbf{w}_1, \ldots, \mathbf{w}_a, \mathbf{v}_1, \ldots, \mathbf{v}_b$ are linearly independent. Suppose that $\sum_i t_i \mathbf{v}_i + \sum_{i=1}^{a} s_i \mathbf{w}_i = \mathbf{0}$ for some real numbers t_i, s_i. We must show that they are all zero. Applying F, we see that $F(\sum_i t_i \mathbf{v}_i + \sum_{i=1}^{a} s_i \mathbf{w}_i) = \mathbf{0}$. Since $F(\mathbf{w}_i) = \mathbf{0}$ and $F(\mathbf{v}_i) = \mathbf{v}'_i$, we see that $\sum_i t_i \mathbf{v}'_i = \mathbf{0}$. As the $\mathbf{v}'_i$ are a basis for $\operatorname{Im} F$, they are linearly independent and so $t_i = 0$ for all i. Thus since $\sum_i t_i \mathbf{v}_i + \sum_{i=1}^{a} s_i \mathbf{w}_i = \mathbf{0}$, $\sum_{i=1}^{a} s_i \mathbf{w} = \mathbf{0}$. But as the $\mathbf{w}_i$ are a basis for $\operatorname{Ker} F$, they are also linearly independent and so $s_i = 0$ for all i as well. So $\mathbf{w}_1, \ldots, \mathbf{w}_a, \mathbf{v}_1, \ldots, \mathbf{v}_b$ are linearly independent. □

Remark 3.10. The proof above shows the following: In the notation of the proof, let $V = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_b\}$. Then (i) $\mathbb{R}^n = \operatorname{Ker} F \oplus V$ and (ii) the restriction of F to V is an isomorphism of vector spaces from V to $\operatorname{Im} F$.

3.4 Change of basis

As we have seen above, we are able to define a linear map $F\colon \mathbb{R}^n \to \mathbb{R}^m$ by arbitrarily prescribing its values on the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Since all bases of $\mathbb{R}^n$ look the same in some sense, we should be able to specify a linear map by prescribing its values on an arbitrary basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$. This is indeed the case:

Lemma 3.11. *Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be a basis of $\mathbb{R}^n$, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be n vectors in $\mathbb{R}^m$. Then there exists a unique linear map $F\colon \mathbb{R}^n \to \mathbb{R}^m$ such that $F(\mathbf{w}_i) = \mathbf{v}_i$ for every i.*

Proof. Let us first show that, if F exists, it is unique. Since $\mathbf{w}_1, \ldots, \mathbf{w}_n$ is a basis of $\mathbb{R}^n$, for each $\mathbf{x} \in \mathbb{R}^n$, there exist unique real numbers $t_i, 1 \le i \le n$ such that $\mathbf{x} = \sum_{i=1}^{n} t_i \mathbf{w}_i$. By linearity of F, if it exists, we must have $F(\mathbf{x}) = \sum_{i=1}^{n} t_i F(\mathbf{w}_i) = \sum_{i=1}^{n} t_i \mathbf{v}_i$. Thus shows that F is unique. Conversely let us define F by the rule: $F(\sum_{i=1}^{n} t_i \mathbf{w}_i) = \sum_{i=1}^{n} t_i \mathbf{v}_i$. This is well-defined and easily seen to be linear. □

We remark that, in the above proof, if we let $H\colon \mathbb{R}^n \to \mathbb{R}^n$ be the "change of basis" map defined by $H(t_1, \ldots, t_n) = \sum_i t_i \mathbf{w}_i$, then H is one-to-one and onto and so has an inverse H^{-1}, which is again linear. Let $G\colon \mathbb{R}^n \to \mathbb{R}^m$ be the map $G(t_1, \ldots, t_n) = \sum_i t_i \mathbf{v}_i$. Then the map F is equal to $G \circ H^{-1}$, giving another (and perhaps clearer) proof of the existence.

A related question is the following: Suppose that $\mathbf{w}_1, \ldots, \mathbf{w}_n$ is a basis of $\mathbb{R}^n$ and that $\mathbf{w}'_1, \ldots, \mathbf{w}'_m$ is a basis of $\mathbb{R}^m$. As we have just seen, given n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in $\mathbb{R}^m$, we can specify a unique linear map $F\colon \mathbb{R}^n \to \mathbb{R}^m$ via the rule $F(\mathbf{w}_i) = \mathbf{v}_i$. Now the vectors $\mathbf{v}_i$ can be written in terms of the basis $\mathbf{w}'_1, \ldots, \mathbf{w}'_m$. Writing $\mathbf{v}_i = \sum_{j=1}^m b_{ji} \mathbf{w}'_j$ defines an $n \times m$ matrix $B = (b_{ij})$. Note that we are reversing the indices so as to be consistent with our notation for matrices which we read off from the standard bases. On the other hand, F is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ and so there is an associated $n \times m$ matrix $A = (a_{ij})$ which describes F in terms of the standard bases for $\mathbb{R}^n$ and $\mathbb{R}^m$: $F(\mathbf{e}_i) = \sum_{j=1}^k a_{ji} \mathbf{e}_j$. What is the relationship between the matrices A and B?

This question can be answered most simply in terms of linear maps. Let $G\colon \mathbb{R}^n \to \mathbb{R}^m$ be the linear map corresponding to the matrix B in the usual way: $G(\mathbf{e}_i) = \sum_{j=1}^k b_{ji} \mathbf{e}_j$. As above, let $H_1\colon \mathbb{R}^n \to \mathbb{R}^n$ be the "change of basis" map in $\mathbb{R}^n$ defined by $H_1(\mathbf{e}_i) = \mathbf{w}_i$, and let $H_2\colon \mathbb{R}^m \to \mathbb{R}^m$ be the corresponding change of basis map in $\mathbb{R}^m$, defined by $H_2(\mathbf{e}_j) = \mathbf{w}'_j$. Then clearly

$$F = H_2 \circ G \circ H_1^{-1}.$$

Thus, if H_1 corresponds to the $n \times n$ (square) matrix C_1 and H_2 corresponds to the $m \times m$ matrix C_2, then F, which corresponds to the matrix A, also corresponds to the matrix $C_2 \cdot B \cdot C_1^{-1}$, and thus we have the following equality of $m \times n$ matrices:

$$A = C_2 \cdot B \cdot C_1^{-1}.$$

A case that often arises is when $n = m$, so that A and B are square matrices, and the two bases $\mathbf{w}_1, \ldots, \mathbf{w}_n$ and $\mathbf{w}'_1, \ldots, \mathbf{w}'_n$ are equal. In this case $C_1 = C_2 = C$, say, and the above formula reads

$$A = C \cdot B \cdot C^{-1}.$$

We say that A is obtained by *conjugating* B *by* C.

Chapter 4

Row reduction

4.1 Outline of the problem

There are many explicit computational techniques for dealing with various related computational problems in linear algebra, and they go by many names: row reduction, Gaussian elimination, Gauss-Jordan elimination, ... Among the problems we would like to analyze are the following:

1. Given $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, are the $\mathbf{v}_i$ are linearly independent and what is their span? If A is the matrix with columns $\mathbf{v}_i$, we are asking equivalently to decide when the system $A\mathbf{x} = \mathbf{b}$ has a solution (what is $\operatorname{Im} A = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$?), and to explicitly exhibit one or all solutions if possible.

2. With $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and A as above, find all $t_1, \ldots, t_k \in \mathbb{R}$ such that $t_1\mathbf{v}_1 + \cdots + t_k\mathbf{v}_k = \mathbf{0}$. Equivalently, are there are any solutions to $A\mathbf{x} = \mathbf{0}$ aside from the trivial solution $\mathbf{x} = \mathbf{0}$ (is $\operatorname{Ker} A = \{\mathbf{0}\}$), in other words are the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ linearly independent?). More generally, we would like to find a basis for $\operatorname{Ker} A$.

Here and later, we will write a matrix equation $A\mathbf{x} = \mathbf{b}$ where the vectors $\mathbf{x}$ and $\mathbf{b}$ written as **row** vectors, even though they must be **column** vectors for this matrix equation to make sense, and likewise speak of $\operatorname{Ker} A$ and $\operatorname{Im} A$, written as sets of row vectors, when we really mean the kernel and image of the linear map corresponding to A.

Such questions and many others can be solved fairly efficiently with the method we now describe. This procedure is based upon the following two easy observations: First, in general it is hard to look at a collection of vectors and decide if another vector is in their span. However, if the vectors are in

a very special form it can be quite easy. For example, it is easy to describe $\text{span}\{(1,0,-3,1),(0,1,2,-2)\}$: if

$$\mathbf{x} = (x_1, x_2, x_3, x_4) = t_1(1,0,-3,1) + t_2(0,1,2,-2),$$

then $t_1 = x_1$, $t_2 = x_2$, and then $x_3 = -3x_1 + 2x_2$ and $x_4 = x_1 - 2x_2$. It is easy to see that these last two equations characterize the span. We might have to consider slightly more general situations. For example,

$$V = \text{span}\{(0,0,1,2,3,0,7,2),(0,0,0,0,0,1,-5,6),(0,0,0,0,0,0,0,0)\}$$

can likewise be described via

$$V = (x_1,\ldots,x_8) : x_1 = x_2 = 0, x_4 = 2x_1, x_5 = 3x_1, x_7 = 7x_1 - 5x_2, x_8 = 2x_1 + 6x_2\}.$$

So we would like to find a sequence of operations that puts $\mathbf{v}_1,\ldots,\mathbf{v}_k$ in a standard form (sometimes called *row echelon form*) without changing their span.

The second observation is the following: Consider the following operations applied to the sequence of vectors $\mathbf{v}_1,\ldots,\mathbf{v}_k$, yielding a new sequence $\mathbf{v}'_1,\ldots,\mathbf{v}'_k$:

1. Switch $\mathbf{v}_i$ and $\mathbf{v}_j$, leaving the other vectors alone;
2. Replace $\mathbf{v}_i$ by $t\mathbf{v}_i$, where $t \in \mathbb{R}$ and $t \neq 0$, leaving the other vectors alone;
3. Replace $\mathbf{v}_i$ by $\mathbf{v}'_i = \mathbf{v}_i - t\mathbf{v}_j$, for some $j \neq i$, leaving all other vectors (including $\mathbf{v}_j$) alone.

Then these operations do not affect $\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$, and moreover $\mathbf{v}_1,\ldots,\mathbf{v}_k$ are linearly independent if and only if the same is true after performing one of the operations.

To see this, note that all three operations applied to a sequence $\mathbf{v}_1,\ldots,\mathbf{v}_k$ have inverse operations (applioed to the new sequence $\mathbf{v}'_1,\ldots,\mathbf{v}'_k$) of the same type. For example (1) is its own inverse: If we switch $\mathbf{v}_i$ and $\mathbf{v}_j$, and then switch them back, the sequence is the one we started with. The inverse operation to (2) is: Replace $\mathbf{v}'_i$ by $t^{-1}\mathbf{v}'_i = \mathbf{v}_i$. For (3), replace $\mathbf{v}'_i = \mathbf{v}_i - t\mathbf{v}_j$ by $\mathbf{v}'_i - t\mathbf{v}'_j = \mathbf{v}_i$, which is again an operation of type (3) applied to the sequence $\mathbf{v}'_1,\ldots,\mathbf{v}'_k$.

Now note that, by construction, in all of the cases (1)—(3), we clearly have $\text{span}\{\mathbf{v}'_1,\ldots,\mathbf{v}'_k\} \subseteq \text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$. By the symmetry of the construction we also have $\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_k\} \subseteq \text{span}\{\mathbf{v}'_1,\ldots,\mathbf{v}'_k\}$ and so

$$\text{span}\{\mathbf{v}'_1,\ldots,\mathbf{v}'_k\} = \text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}.$$

Thus the spans are the same. We must also check that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent $\iff$ $\mathbf{v}'_1, \ldots, \mathbf{v}'_k$ are linearly independent. We will just do this for an operation of type (3) (the others are easier). Suppose for example that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent and that $\sum_{\ell=1}^{k} s_\ell \mathbf{v}'_\ell = \mathbf{0}$. Then, since $\mathbf{v}'_\ell = \mathbf{v}_\ell$ for $\ell \neq i$ and $\mathbf{v}'_i = \mathbf{v}_i - t\mathbf{v}_j$, $\sum_{\ell=1}^{k} t_\ell \mathbf{v}_\ell = \mathbf{0}$, where $t_\ell = s_\ell$ for $\ell \neq j$ and $t_j = s_j - ts_i$. Since $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent, $t_\ell = 0$ for all ℓ. Thus $s_\ell = 0$ for $\ell \neq j$, and $s_j = t_j + ts_i = t_j$ as well, since $s_i = 0$. Conversely, if the $\mathbf{v}'_1, \ldots, \mathbf{v}'_k$ are linearly independent, then since $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are obtained from $\mathbf{v}'_1, \ldots, \mathbf{v}'_k$ by an operation of type (3), they are linearly independent as well.

4.2 How to find the span of a sequence of vectors

We now illustrate how these operations can be used to solve the problems outlined at the beginning of the chapter. The first question is to decide if a sequence of vectors is linearly independent and to describe its span.

Consider the example $\mathbf{v}_1 = (2, 3, -6, 4)$, $\mathbf{v}_2 = (1, 1, 5, 2)$. We write the two vectors as $\begin{matrix} 2 & 3 & -6 & 4 \\ 1 & 1 & 5 & 2 \end{matrix}$, without commas or parentheses, to save time. First multiply the top row by $\frac{1}{2}$ to make the first entry 1. Then subtract off the first row from the second to make the second row start with a zero. Then multiply the second row by -2 to make the second entry of the second row 1, then subtract off $\frac{3}{2}$ times the second row from the first to make the second entry of the first row 0. The sequence of operations is denoted as:

$$\begin{matrix} 2 & 3 & -6 & 4 \\ 1 & 1 & 5 & 2 \end{matrix} \rightarrow \begin{matrix} \underline{1} & \frac{3}{2} & -3 & 2 \\ 1 & 1 & 5 & 2 \end{matrix} \rightarrow \begin{matrix} 1 & \frac{3}{2} & -3 & 2 \\ \underline{0} & -\frac{1}{2} & 8 & 0 \end{matrix}$$

$$\rightarrow \begin{matrix} 1 & \frac{3}{2} & -3 & 2 \\ 0 & \underline{1} & -16 & 0 \end{matrix} \rightarrow \begin{matrix} 1 & \underline{0} & 21 & 2 \\ 0 & 1 & -16 & 0 \end{matrix},$$

where we have underlined at each stage the term we were trying to achieve. This tells us that $\mathbf{v}_1 = (2, 3, -6, 4)$, $\mathbf{v}_2 = (1, 1, 5, 2)$ are linearly independent and that their span is the same as the span of $(1, 0, 21, 2)$ and $(0, 1, -16, 0)$, so that $\mathbf{x} = (x_1, x_2, x_3, x_4) \in \text{span}\{\mathbf{v}_1, \mathbf{v}_2\} \iff x_3 = 21x_1 - 16x_2$ and $x_4 = 2x_1$. Note that the original vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ do in fact satisfy these

equations. To rephrase our results, the system of equations

$$\begin{aligned} 2x_1 + x_2 &= b_1 \\ 3x_1 + x_2 &= b_2 \\ -6x_1 + 5x_2 &= b_3 \\ 4x_1 + 2x_2 &= b_4 \end{aligned}$$

has a solution $\iff$ $b_3 = 21b_1 - 16b_2$ and $b_4 = 2b_1$. Equivalently, if A is the matrix

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 1 \\ -6 & 5 \\ 4 & 2 \end{pmatrix}$$

then $(x_1, x_2, x_3, x_4) \in \operatorname{Im} A \iff x_3 = 21x_1 - 16x_2$ and $x_4 = 2x_1$. Note that in both cases we switched columns into rows.

Here is another example where we have to switch the rows and are unable to do anything with the first column: let $\mathbf{v}_1 = (0, 0, 2, 1, 3)$, $\mathbf{v}_2 = (0, -2, 4, 0, -8)$. Then the procedure would look as follows:

$$\begin{matrix} 0 & 0 & 2 & 1 & 3 \\ 0 & -2 & 4 & 0 & -8 \end{matrix} \rightarrow \begin{matrix} 0 & \underline{-2} & 4 & 0 & -8 \\ 0 & 0 & 2 & 1 & 3 \end{matrix} \rightarrow \begin{matrix} 0 & \underline{1} & -2 & 0 & 4 \\ 0 & 0 & 2 & 1 & 3 \end{matrix}$$

$$\rightarrow \begin{matrix} 0 & 1 & -2 & 0 & 4 \\ 0 & 0 & \underline{1} & \frac{1}{2} & \frac{3}{2} \end{matrix} \rightarrow \begin{matrix} 0 & 1 & \underline{0} & 1 & 7 \\ 0 & 0 & 1 & \frac{1}{2} & \frac{3}{2} \end{matrix} .$$

The final description of the span of $\mathbf{v}_1 = (0, 0, 2, 1, 3)$, $\mathbf{v}_2 = (0, -2, 4, 0, -8)$ is then:

$$\{(x_1, x_2, x_3, x_4, x_5) : x_1 = 0, x_4 = x_2 + \frac{1}{2}x_3, x_5 = 7x_2 + \frac{3}{2}x_3\}.$$

We now work through a more complicated example in detail Suppose that we wish to solve the system

$$\begin{aligned} 2x_1 + x_2 + 0x_3 - 2x_4 &= b_1; \\ 2x_1 + 3x_2 + 9x_3 + 4x_4 &= b_2; \\ 4x_1 + 2x_2 + 3x_3 + 2x_4 &= b_3; \\ 6x_1 + 7x_2 + 12x_3 - 6x_4 &= b_4. \end{aligned}$$

In this case, the vectors in question are $\mathbf{v}_1 = (2, 2, 4, 6), \mathbf{v}_2 = (1, 3, 2, 7), \mathbf{v}_3 = (0, 9, 3, 12), \mathbf{v}_4 = (-2, 4, 2, -6)$.

The procedure is then as follows:

(A) Search for a $\mathbf{v}_i$ with nonzero first component; in our case $i = 1$ works. In general switch $\mathbf{v}_i$ and $\mathbf{v}_1$ if $\mathbf{v}_1$ doesn't have nonzero first component. It might be the case that all of the $\mathbf{v}_i$ have zero first component. In this case look for $\mathbf{v}_i$ which has a nonzero second component, and so on. In the general case we find the first $\mathbf{v}_i$ whose a^{th} component is nonzero, where a is the smallest natural number such that some $\mathbf{v}_j$ has a nonzero a^{th} component.

(B) After multiplying $\mathbf{v}_i$ by the appropriate nonzero scalar, its first component becomes 1. For example, we replace $\mathbf{v}_1$ by $(1, 1, 2, 3)$ above. We write:

$$\begin{matrix} 2 & 2 & 4 & 6 \\ 1 & 3 & 2 & 7 \\ 0 & 9 & 3 & 12 \\ -2 & 4 & 2 & -6 \end{matrix} \quad \rightarrow \quad \begin{matrix} 1 & 1 & 2 & 3 \\ 1 & 3 & 2 & 7 \\ 0 & 9 & 3 & 12 \\ -2 & 4 & 2 & -6 \end{matrix}$$

(C) Next, after subtracting off an appropriate multiple of $\mathbf{v}_1$ from $\mathbf{v}_i, i > 1$, we can assume that $\mathbf{v}_i$ has first component zero for $i > 1$. For example, subtracting off $(1, 1, 2, 3)$ from $\mathbf{v}_2 = (1, 3, 2, 7)$ gives $(0, 2, 0, 4)$. Note that we can divide the coefficients of this by 2. This observation isn't part of the algorithm, nor is it ever necessary, but it usually helps to keep the numbers from getting too big. Likewise, we can divide the coefficients of $\mathbf{v}_3$ by 3; in this case, we don't need to subtract off $\mathbf{v}_1$ since the first coordinate is already zero. For $\mathbf{v}_4$ we have to subtract off $-2\mathbf{v}_1$, i.e. add $2\mathbf{v}_1$. Symbolically:

$$\begin{matrix} 1 & 1 & 2 & 3 \\ 1 & 3 & 2 & 7 \\ 0 & 9 & 3 & 12 \\ -2 & 4 & 2 & -6 \end{matrix} \rightarrow \begin{matrix} 1 & 1 & 2 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & 9 & 3 & 12 \\ -2 & 4 & 2 & -6 \end{matrix} \rightarrow \begin{matrix} 1 & 1 & 2 & 3 \\ 0 & 1 & 0 & 4 \\ 0 & 3 & 1 & 12 \\ -2 & 4 & 2 & -6 \end{matrix} \rightarrow \begin{matrix} 1 & 1 & 2 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 3 & 1 & 4 \\ 0 & 6 & 6 & 0 \end{matrix}$$

Notice that we could have replaced $(-2, 4, 2, -6)$ by $(-1, 2, 1, -3)$ but chose not to; however, we will divide the final result $(0, 6, 6, 0)$ by 6 to get $(0, 1, 1, 0)$. Thus we have (so far) the vectors

$$(1, 1, 2, 3), (0, 1, 0, 2), (0, 3, 1, 4), (0, 1, 1, 0).$$

(D) Now we have a vector $\mathbf{v}_1$ with first component 1 and the others all have first component zero (or a vector $\mathbf{v}_1$ whose first i components are zero, whose $(i+1)^{\text{st}}$ component is 1 and for all the other vectors $\mathbf{v}_j$ the first $i + 1$ components of $\mathbf{v}_j$ are zero. Look for a vector $\mathbf{v}_2$ whose second

component is nonzero and switch it with $\mathbf{v}_2$. In our example, the next vector $\mathbf{v}_2 = (0, 1, 0, 2)$ already works. After multiplying by a nonzero scalar we can assume that its first component is 1. For our example we don't have to do this, since the second component is already one. But if we had $(0, 3, 1, 4)$ down as our second vector instead, we would replace it by $(0, 1, \frac{1}{3}, \frac{4}{3})$. Of course, in some examples all the second (or $(i+2)^{\text{nd}}$) components of the remaining vectors will be zero, so again we look for the first nonzero one and switch it up to $\mathbf{v}_2$.

(E) Using this new $\mathbf{v}_2$, make the second component of $\mathbf{v}_1$ and all the $\mathbf{v}_j$ for $j \geq 3$ equal to zero. For example:

$$\begin{matrix} 1 & 1 & 2 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 3 & 1 & 4 \\ 0 & 1 & 1 & 0 \end{matrix} \quad \rightarrow \quad \begin{matrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & -2 \end{matrix}$$

Note at this stage it is already clear that the vectors are not linearly independent—and so the original sequence of vectors was not.

(F) Continue this procedure until there is nothing left to do. For example, in our sequence of vectors we take the vector $\mathbf{v}_3$, which already has third component nonzero and indeed equal to one. Using it, we eliminate the third component of all of the vectors $\mathbf{v}_i$:

$$\begin{matrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & -2 \end{matrix} \quad \rightarrow \quad \begin{matrix} 1 & 0 & 0 & 5 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 \end{matrix}$$

If instead we had found a nonzero vector $\mathbf{v}_4$, it would necessarily have been of the form $(0, 0, 0, t)$ for some $t \in \mathbb{R}$, and we would have been able to use it to produce the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_4$. That would have told us that the original four vectors were linearly independent and that their span was $\mathbb{R}^4$. In this case, we see that we could find three independent vectors in the span, and that the span is the same as the span of $(1, 0, 0, 5), (0, 1, 0, 2), (0, 0, 1, -2)$. A linear combination $x_1(1, 0, 0, 5) + x_2(0, 1, 0, 2) + x_3(0, 0, 1, -2)$ of the above three vectors must satisfy

$$x_4 = 5x_1 + 2x_2 - 2x_3,$$

and conversely all vectors satisfying this equation are a linear combination of the above three vectors.

As a check, we notice that all three of the original vectors

$$(2, 2, 4, 6), (1, 3, 2, 7), (0, 9, 3, 12), (-2, 4, 2, -6)$$

do in fact satisfy this equation. This tells us that we didn't make any arithmetic mistakes (or they were lucky enough to cancel each other out). We also can say when we can solve the original system for a vector (b_1, b_2, b_3, b_4): a solution exists if and only if $b_4 = 5b_1 + 2b_2 - 2b_3$. Finally we should add that in general you should expect a lot of fractions to appear in the calculation, even though they did not in our example, and there are an incredible number of potential arithmetic mistakes waiting to be made.

4.3 Row reduction with bookkeeping

We return to the example of $\mathbf{v}_1 = (2, 3, -6, 4)$, $\mathbf{v}_2 = (1, 1, 5, 2)$. Given the vector $\mathbf{b} = (4, 5, 4, 8)$, we ask not only if it is in the span of $\mathbf{v}_1$ and $\mathbf{v}_2$, but, if so, how we can write it as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. (Note that the original discussion of the span of $\mathbf{v}_1$, $\mathbf{v}_2$ already shows that $\mathbf{b}$ is in the span.) To do so, we write down all three vectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{b}$ and add a 3×3 identity matrix to the right of the matrix whose rows are $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{b}$. Here the 3 is the number of the $\mathbf{v}_i$ (in this case 2) plus one more for the vector $\mathbf{b}$. To help you keep track of what we are doing in this example, we will also write down the vectors appearing in the left-hand 3×4 matrix in terms of $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{b}$. (In general we would omit this step.) Thus we will apply row reduction to the matrix

$$\begin{array}{cccc|ccc} 2 & 3 & -6 & 4 & 1 & 0 & 0 \\ 1 & 1 & 5 & 2 & 0 & 1 & 0 \\ 4 & 5 & 4 & 8 & 0 & 0 & 1 \end{array} .$$

Running through the algorithm, and noting the vectors that appear in the left-hand 3×4 matrix in terms of $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{b}$, we find:

$$\begin{array}{rcccccccccc}
 & \mathbf{v}_1 & 2 & 3 & -6 & 4 & 1 & 0 & 0 & & \\
 & \mathbf{v}_2 & 1 & 1 & 5 & 2 & 0 & 1 & 0 & \to & \\
 & \mathbf{b} & 4 & 5 & 4 & 8 & 0 & 0 & 1 & & \\
 & \frac{1}{2}\mathbf{v}_1 & 1 & \frac{3}{2} & -3 & 2 & \frac{1}{2} & 0 & 0 & & \\
\to & \mathbf{v}_2 & 1 & 1 & 5 & 2 & 0 & 1 & 0 & \to & \\
 & \mathbf{b} & 4 & 5 & 4 & 8 & 0 & 0 & 1 & & \\
 & \frac{1}{2}\mathbf{v}_1 & 1 & \frac{3}{2} & -3 & 2 & \frac{1}{2} & 0 & 0 & & \\
\to & -\frac{1}{2}\mathbf{v}_1 + \mathbf{v}_2 & 0 & -\frac{1}{2} & 8 & 0 & -\frac{1}{2} & 1 & 0 & \to & \\
 & -2\mathbf{v}_1 + \mathbf{b} & 0 & -1 & 16 & 0 & -2 & 0 & 1 & & \\
 & \frac{1}{2}\mathbf{v}_1 & 1 & \frac{3}{2} & -3 & 2 & \frac{1}{2} & 0 & 0 & & \\
\to & \mathbf{v}_1 - 2\mathbf{v}_2 & 0 & 1 & -16 & 0 & 1 & -2 & 0 & \to & \\
 & -2\mathbf{v}_1 + \mathbf{b} & 0 & -1 & 16 & 0 & -2 & 0 & 1 & & \\
 & -\mathbf{v}_1 + 3\mathbf{v}_2 & 1 & 0 & 21 & 2 & -1 & 3 & 0 & & \\
\to & \mathbf{v}_1 - 2\mathbf{v}_2 & 0 & 1 & -16 & 0 & 1 & -2 & 0 & . & \\
 & -\mathbf{v}_1 - 2\mathbf{v}_2 + \mathbf{b} & 0 & 0 & 0 & 0 & -1 & -2 & 1 & &
\end{array}$$

For example, note that at the last step (Step 5), the term $-\mathbf{v}_1 + 3\mathbf{v}_2 = \frac{1}{2}\mathbf{v}_1 - \frac{3}{2}(\mathbf{v}_1 - 2\mathbf{v}_2)$ in the first row is obtained by subtracting $\frac{3}{2} \times (\mathbf{v}_1 - 2\mathbf{v}_2)$ in the second row in Step 4 from the term $\frac{1}{2}\mathbf{v}_1$ in the first row of Step 4, and similarly the term in the third row $-\mathbf{v}_1 - 2\mathbf{v}_2 + \mathbf{b}$ in Step 5 is obtained by adding the second row term $\mathbf{v}_1 - 2\mathbf{v}_2$ in Step 4 to the term $-2\mathbf{v}_1 + \mathbf{b}$ in Step 4. Comparing the 3×3 matrices on the right with the linear combinations on the left, we see that the entries of the 3×3 matrix keep track of the coefficients of $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{b}$ that show up in computing the vector in the middle. In particular, this says that $(1, 0, 21, 2) = -\mathbf{v}_1 + 3\mathbf{v}_2$ and $(0, 1, -16, 0) = \mathbf{v}_1 - 2\mathbf{v}_2$. But we are really interested only in the last equality, which says that $\mathbf{0} = -\mathbf{v}_1 - 2\mathbf{v}_2 + \mathbf{b}$, in other words that $\mathbf{b} = \mathbf{v}_1 + 2\mathbf{v}_2$. In this way we have explicitly written $\mathbf{b}$ as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$.

The next example illustrates the case where there is a nontrivial kernel. Consider the system of equations

$$\begin{aligned} x_1 + 5x_2 + 7x_3 - 2x_4 &= 1; \\ 3x_1 - x_2 + x_3 + 2x_4 &= -1, \end{aligned}$$

corresponding to the matrix $A = \begin{pmatrix} 1 & 5 & 7 & -2 \\ 3 & -1 & 1 & 2 \end{pmatrix}$, or to the span of the vectors $\mathbf{v}_1 = (1, 3)$, $\mathbf{v}_2 = (5, -1)$, $\mathbf{v}_3 = (7, 1)$, $\mathbf{v}_4 = (-2, 2)$, and $\mathbf{b} = (1, -1)$.

Running through the algorithm, turning the columns of A into **row** vectors, gives:

$$\begin{array}{cccccccccc}
1 & 3 & 1 & 0 & 0 & 0 & 0 & & 1 & 3 & 1 & 0 & 0 & 0 & 0 & \\
5 & -1 & 0 & 1 & 0 & 0 & 0 & & 0 & -16 & -5 & 1 & 0 & 0 & 0 & \\
7 & 1 & 0 & 0 & 1 & 0 & 0 & \to & 0 & -20 & -7 & 0 & 1 & 0 & 0 & \to \\
-2 & 2 & 0 & 0 & 0 & 1 & 0 & & 0 & 8 & 2 & 0 & 0 & 1 & 0 & \\
1 & -1 & 0 & 0 & 0 & 0 & 1 & & 0 & -4 & -1 & 0 & 0 & 0 & 1 &
\end{array}$$

$$\begin{array}{cccccccccccccc}
 & 1 & 3 & 1 & 0 & 0 & 0 & 0 & & 1 & 0 & \frac{1}{16} & -\frac{3}{16} & 0 & 0 & 0 \\
 & 0 & 1 & \frac{5}{16} & -\frac{1}{16} & 0 & 0 & 0 & & 0 & 1 & \frac{5}{16} & -\frac{1}{16} & 0 & 0 & 0 \\
\to & 0 & -20 & -7 & 0 & 1 & 0 & 0 & \to & 0 & 0 & -\frac{3}{4} & -\frac{5}{4} & 1 & 0 & 0 \\
 & 0 & 8 & 2 & 0 & 0 & 1 & 0 & & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 0 & 1 & 0 \\
 & 0 & -4 & -1 & 0 & 0 & 0 & 1 & & 0 & 0 & \frac{1}{4} & -\frac{1}{4} & 0 & 0 & 1
\end{array},$$

Reading the fifth row we see that $(1,-1) = -\frac{1}{4}(1,3) + \frac{1}{4}(5,-1)$, giving a particular solution $(-\frac{1}{4}, \frac{1}{4}, 0, 0)$ to $A \cdot \mathbf{x} = (1,-1)$, which is written as a linear combination of the first two vectors, $(1,3)$ and $(5,-1)$. This is something of an artifact of the procedure, since we can readily identify a simpler particular solution $(0,0,0,-\frac{1}{2})$. The third and fourth rows of the final output tell us that

$$-\frac{3}{4}(1,3) - \frac{5}{4}(5,-1) + (7,1) = (0,0);$$
$$-\frac{1}{2}(1,3) + \frac{1}{2}(5,-1) + (-2,2) = (0,0).$$

In terms of the system which began this example, a particular solution is $(-\frac{1}{4}, \frac{1}{4}, 0, 0)$, and a basis for the kernel of A (written as row vectors) is $(-\frac{3}{4}, -\frac{5}{4}, 1, 0), (-\frac{1}{2}, \frac{1}{2}, 0, 1)$. Thus, the set of all solutions of the system is given by

$$\{(-\frac{1}{4} - \frac{3}{4}t_1 - \frac{1}{2}t_2, \frac{1}{4} - \frac{5}{4}t_1 + \frac{1}{2}t_2, t_1, t_2) : t_1, t_2 \in \mathbb{R}\}.$$

For another example, consider the four vectors in $\mathbb{R}^3$:

$$\mathbf{v}_1 = (1,1,2), \mathbf{v}_2 = (2,3,0), \mathbf{v}_3 = (4,7,-4), \mathbf{v}_4 = (-9,-14,2).$$

Let $\mathbf{b} = (5,-2,38)$, and let A be the matrix whose columns are given by the $\mathbf{v}_i$, so that we are trying to solve the equation $A\mathbf{x} = \mathbf{b}$. To determine the solutions, write down the matrix whose **rows** are the $\mathbf{v}_i$, along with one final row which is $\mathbf{b}$. To keep track of the steps in row reduction, we add in

a matrix which is the 5×5 identity matrix (5 is the number of the $\mathbf{v}_i$'s plus one more for $\mathbf{b}$).

$$\text{Now apply row reduction to } \begin{matrix} 1 & 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 7 & -4 & 0 & 0 & 1 & 0 & 0 \\ -9 & -14 & -2 & 0 & 0 & 0 & 1 & 0 \\ 5 & -2 & 38 & 0 & 0 & 0 & 0 & 1 \end{matrix} :$$

$$\begin{matrix} 1 & 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 7 & -4 & 0 & 0 & 1 & 0 & 0 \\ -9 & -14 & -2 & 0 & 0 & 0 & 1 & 0 \\ 5 & -2 & 38 & 0 & 0 & 0 & 0 & 1 \end{matrix} \rightarrow \begin{matrix} 1 & 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & -2 & 1 & 0 & 0 & 0 \\ 0 & 3 & -12 & -4 & 0 & 1 & 0 & 0 \\ 0 & -5 & -20 & 9 & 0 & 0 & 1 & 0 \\ 0 & -7 & 28 & -5 & 0 & 0 & 0 & 1 \end{matrix}$$

$$\rightarrow \begin{matrix} 1 & 0 & 6 & 3 & -1 & 0 & 0 & 0 \\ 0 & 1 & -4 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -3 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 5 & 0 & 1 & 0 \\ 0 & 0 & 0 & -19 & 7 & 0 & 0 & 1 \end{matrix}.$$

Note that, at the second step, the vectors are now $\mathbf{v}_1, -2\mathbf{v}_1 + \mathbf{v}_2, -4\mathbf{v}_1 + \mathbf{v}_3, 9\mathbf{v}_1 + \mathbf{v}_4$, and $\mathbf{b}$ is replaced by $-5\mathbf{v}_1 + \mathbf{b}$. At the last step, $\mathbf{v}_1$ has been replaced by $\mathbf{v}_1 + (-1)(-2\mathbf{v}_1 + \mathbf{v}_2) = 3\mathbf{v}_1 - \mathbf{v}_2$. Thus the entries of the first row on the right keep track of how to write $(1, 0, 6)$ as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. Likewise, the second row tells us how to write $(0, 1, -4)$ as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. This is not the most interesting information we want to extract from the final matrix—instead, we will use the information in the last three rows. The third row tells us that $\mathbf{0} = 2\mathbf{v}_1 - 3\mathbf{v}_2 + \mathbf{v}_3$. On the one hand, this says that $\mathbf{v}_3 = -2\mathbf{v}_1 + 3\mathbf{v}_2$, in other words, how to write $\mathbf{v}_3$ as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. Also, this tells us that $(2, -3, 1, 0)$ is in the kernel of A. Likewise, the fourth row tells us that $\mathbf{0} = -\mathbf{v}_1 + 5\mathbf{v}_2 + \mathbf{v}_4$, so that $\mathbf{v}_4 = \mathbf{v}_1 - 5\mathbf{v}_2$ and that that $(-1, 5, 0, 1)$ is in the kernel of A. We read off that the dimension of the span of the $\mathbf{v}_i$ is 2, and hence that $\dim \operatorname{Ker} A = 4 - 2 = 2$. Since $(2, -3, 1, 0)$ and $(-1, 5, 0, 1)$ are linearly independent, they are a basis for $\operatorname{Ker} A$. In fact, this works generally for any matrix A: the elements of the kernel corresponding to the entries in the final row reduced form which are all 0's give a basis for $\operatorname{Ker} A$. You can see this by noting that the final entries (reading right to left and bottom

to top) for the elements of the kernel look like the standard basis vectors, so that they are linearly independent, and that there are $k - \dim \text{span}\{\mathbf{v}_i\}$ of them, which is the dimension of $\text{Ker}\, A$. Finally, the last row of the row reduced matrix tells us how to write $\mathbf{b}$ as a linear combination of the $\mathbf{v}_i$: $-19\mathbf{v}_1 + 7\mathbf{v}_2 + \mathbf{b} = \mathbf{0}$, so that $\mathbf{b} = 19\mathbf{v}_1 - 7\mathbf{v}_2$. Note that $\mathbf{b}$ is just written in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$. This makes sense, since if $\mathbf{b} \in \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$, then since $\mathbf{v}_3$ and $\mathbf{v}_4$ are linear combinations of $\mathbf{v}_1$ and $\mathbf{v}_2$, we will be able to write $\mathbf{b}$ as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. Finally, a particular solution to $A\mathbf{x} = \mathbf{b}$ is $(-19, 7, 0, 0)$, and so the general solution to the equation can be uniquely written as

$$\mathbf{x} = (-19, 7, 0, 0) + t_1(2, -3, 1, 0) + t_2(-1, 5, 0, 1),$$

and every such vector $\mathbf{x}$ is a solution.

4.4 Finding the inverse of a matrix

A similar method works to find the inverse of a given matrix A. To explain this procedure, first note that, as we have seen, if A is a $k \times n$ matrix with columns $\mathbf{c}_1, \ldots, \mathbf{c}_n$, and $\mathbf{x}$ is an $n \times 1$ column vector with entries $x_1, \ldots, x_n$, then $A \cdot \mathbf{x}$ is the $k \times 1$ column vector $x_1\mathbf{c}_1 + \cdots + x_n\mathbf{c}_n$. More generally, if B is an $n \times m$ matrix with columns $\mathbf{b}_i$, $i = 1, \ldots, m$, then $A \cdot B$ is the $k \times n$ matrix whose columns are $\sum_{i=1}^n b_{i1}\mathbf{c}_i, \ldots, \sum_{i=1}^n b_{im}\mathbf{c}_i$. In particular, if we want B to be an $n \times n$ matrix such that $A \cdot B = I_n$, then the j^{th} column of I_n, namely $\mathbf{e}_j$, is the linear combinations of the $\mathbf{c}_i$ given by $\sum_{i=1}^n b_{ij}\mathbf{c}_i$. Now, assuming that A is invertible, using row reduction to put A in row echelon form outputs the identity matrix I_n, and bookkeeping exhibits the **rows** of the identity matrix as linear combinations of the **rows** of A. To avoid having to switch between rows and columns in this case, we note that, if A is a $k \times n$ matrix with **rows** $\mathbf{r}_1, \ldots, \mathbf{r}_k$, and $\mathbf{x}$ is the $1 \times n$ vector $(x_1, \ldots, x_n)$, then a short calculation shows that $\mathbf{x} \cdot A = x_1\mathbf{r}_1 + \cdots + x_k\mathbf{r}_k$. More generally, if B is an $m \times k$ matrix with rows $\mathbf{b}_i$, $i = 1, \ldots, m$, then $B \cdot A$ is the $m \times n$ matrix whose rows are $\sum_{i=1}^k b_{1i}\mathbf{r}_i, \ldots, \sum_{i=1}^k b_{mi}\mathbf{r}_i$. Thus, if A is an invertible $n \times n$ matrix and we want to find an $n \times n$ matrix B such that $B \cdot A = I_n$, then we want to find b_{ij} such that $\sum_{i=1}^k b_{ji}\mathbf{r}_i = \mathbf{e}_j$, $j = 1, \ldots n$, and in this case B is both a left and a right inverse for A and thus $B = A^{-1}$. But this is what row reduction with bookkeeping does for us: add a! n $n \times n$ identity matrix I_n to the right of A and run through row reduction. The output will be a matrix which is I_n on the left and A^{-1} on the right.

For example, to find the inverse of $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ -1 & -1 & -2 \end{pmatrix}$, we apply the method as follows:

$$\begin{array}{cccccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ -1 & -1 & -2 & 0 & 0 & 1 \end{array} \rightarrow \begin{array}{cccccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{array}$$

$$\rightarrow \begin{array}{cccccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 2 & \frac{4}{3} & -\frac{1}{3} & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{array} \rightarrow \begin{array}{cccccc} 1 & 0 & -1 & -\frac{5}{3} & \frac{2}{3} & 0 \\ 0 & 1 & 2 & \frac{4}{3} & -\frac{1}{3} & 0 \\ 0 & 0 & -1 & -\frac{1}{3} & \frac{1}{3} & 1 \end{array}$$

$$\rightarrow \begin{array}{cccccc} 1 & 0 & -1 & -\frac{5}{3} & \frac{2}{3} & 0 \\ 0 & 1 & 2 & \frac{4}{3} & -\frac{1}{3} & 0 \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{3} & -1 \end{array} \rightarrow \begin{array}{cccccc} 1 & 0 & 0 & -\frac{4}{3} & \frac{1}{3} & -1 \\ 0 & 1 & 0 & \frac{2}{3} & \frac{1}{3} & 2 \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{3} & -1 \end{array}.$$

You can then check that $\begin{pmatrix} -\frac{4}{3} & \frac{1}{3} & -1 \\ \frac{2}{3} & \frac{1}{3} & 2 \\ \frac{1}{3} & -\frac{1}{3} & -1 \end{pmatrix}$ is the inverse of $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ -1 & -1 & -2 \end{pmatrix}$, and we did not have to replace the columns of the matrix by its rows.

There are many variations of these methods. We have not for example tried to give the most computationally efficient methods, but this is an important question if one is doing large scale machine computations. A modified form of the methods in this chapter can be used to give efficient methods for finding determinants. Also, in case we are just interested in finding a basis of Ker A, one approach is to apply row reduction to the **rows** of A. We shall outline how this works later.

Chapter 5

Inner products and orthogonality

5.1 Inner product and length

Let $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$. We define the *inner product* (or *dot product* or *scalar product*) of $\mathbf{v}$ and $\mathbf{w}$ by the following formula:

$$\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + \cdots + v_n w_n.$$

Define the *length* or *norm* of $\mathbf{v}$ by the formula

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{v_1^2 + \cdots + v_n^2}.$$

Note that we can define $\langle \mathbf{v}, \mathbf{w} \rangle$ for the vector space F^n, where F is any field, but $\|\mathbf{v}\|$ only makes sense for $F = \mathbb{R}$.

Proposition 5.1. *We have the following properties for the inner product:*

1. *For all* $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^n$, $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$.
2. *For all* $\mathbf{v}$, $\mathbf{u}$, $w \in \mathbb{R}^n$, $\langle \mathbf{v} + \mathbf{u}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$.
3. *For all* $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^n$ *and* $t \in \mathbb{R}$, $\langle t\mathbf{v}, \mathbf{w} \rangle = t\langle \mathbf{v}, \mathbf{w} \rangle$.
4. *For all* $\mathbf{v}$ *and all* i, $\langle \mathbf{v}, \mathbf{e}_i \rangle = v_i$. *Thus* $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ *for all* $\mathbf{w} \in \mathbb{R}^n$ *if and only if* $\mathbf{v} = \mathbf{0}$.

Proof. These are routine calculations. □

Of course, there are properties similar to (2) and (3) for the second variable w, by the symmetry property (1). These properties say that $\langle \mathbf{v}, \mathbf{w} \rangle$ is a linear function in $\mathbf{v}$ when $\mathbf{w}$ is held fixed, and likewise for $\mathbf{w}$ when $\mathbf{v}$ is held fixed. Such a function of two (vector) variables $\mathbf{v}$ and $\mathbf{w}$ is called *bilinear*. The property (1) is called symmetry, and the inner product is a symmetric bilinear function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$.

It is clear from the definition of a linear map that every linear function $F\colon \mathbb{R}^n$ to $\mathbb{R}$ is of the form

$$F(\mathbf{x}) = a_1 x_1 + \cdots + a_n x_n$$

for unique real numbers a_i. Thus $F(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle$ for some fixed vector $\mathbf{c} \in \mathbb{R}^n$, and in fact $\mathbf{a}$ is uniquely determined by F. More generally, if $F\colon \mathbb{R}^n$ to $\mathbb{R}$ is a linear map, then $F(\mathbf{x}) = (\langle \mathbf{r}_1, \mathbf{x} \rangle, \ldots, \langle \mathbf{r}_k, \mathbf{x} \rangle)$, where the vectors $\mathbf{r}_i$ are the rows of the matrix corresponding to F.

Proposition 5.2. *We have the following properties of the length $\| \cdot \|$:*

1. *For all $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\| \geq 0$, and $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$. As a corollary, we see again that $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in \mathbb{R}^n$ if and only if $\mathbf{v} = \mathbf{0}$.*

2. *For all $\mathbf{v} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, $\|t\mathbf{v}\| = |t|\|\mathbf{v}\|$.*

3. ***(Cauchy-Schwarz inequality)** For all $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^n$, we have*

$$|\langle \mathbf{v}, \mathbf{w} \rangle| \leq \|\mathbf{v}\|\|\mathbf{w}\|,$$

   with equality holding if and only if $\mathbf{v}$ and $\mathbf{w}$ are linearly dependent.

4. ***(Triangle inequality)** For all $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^n$, we have*

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|.$$

   (Equality holds if and only if $\mathbf{v}$ is a nonnegative multiple of $\mathbf{w}$ or vice-versa.)

Proof. (1) and (2) are clear. To see (3), we may assume that $\mathbf{v} \neq \mathbf{0}$, otherwise the statement is trivial. Consider the real valued function

$$f(t) = \langle t\mathbf{v} + \mathbf{w}, t\mathbf{v} + \mathbf{w} \rangle = t^2\|\mathbf{v}\|^2 + 2t\langle \mathbf{v}, \mathbf{w} \rangle + \|\mathbf{w}\|^2.$$

Here we find the second expression for $f(t)$ by expanding out and using the basic properties of the inner product. Thus we see that $f(t)$ is a quadratic

function of t. Moreover $f(t) \geq 0$ for all t, and in particular plugging in the minimum value $= -\dfrac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\|^2}$ (which we could find by setting the derivative equal to 0, or more simply by completing the square) gives

$$\frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{v}\|^4} \|\mathbf{v}\|^2 - 2\frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{v}\|^2} + \|\mathbf{w}\|^2 \geq 0.$$

Simplifying and collecting terms gives

$$\|\mathbf{w}\|^2 \geq \frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{v}\|^2},$$

so that when we cross multiply we get

$$\|\mathbf{v}\|^2 \|\mathbf{w}\|^2 \geq \langle \mathbf{v}, \mathbf{w} \rangle^2.$$

Taking square roots gives (3). Note that equality holds if and only if $\langle t\mathbf{v} + \mathbf{w}, t\mathbf{v} + \mathbf{w} \rangle = 0$ for the minimum value, which says that $t\mathbf{v} + \mathbf{w} = \mathbf{0}$ and that $\mathbf{w}$ is a scalar multiple of $\mathbf{v}$.

Finally, to see (4), we have

$$\begin{aligned} \|\mathbf{v} + \mathbf{w}\|^2 &= \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle \\ &= \langle \mathbf{v}, \mathbf{v} \rangle + 2\langle \mathbf{v}, \mathbf{w} \rangle + \mathbf{w}, \mathbf{w} \rangle \\ &= \|\mathbf{v}\|^2 + 2\langle \mathbf{v}, \mathbf{w} \rangle + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2|\langle \mathbf{v}, \mathbf{w} \rangle| + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2\|\mathbf{v}\|\|\mathbf{w}\| + \|\mathbf{w}\|^2 \\ &= (\|\mathbf{v}\| + \|\mathbf{w}\|)^2. \end{aligned}$$

We leave it as an exercise to see when equality is attained. □

The geometric meaning of the triangle inequality is as follows: there is a triangle in $\mathbb{R}^n$ with side lengths $\|\mathbf{v}\|$, $\|\mathbf{w}\|$ and $\|\mathbf{v} + \mathbf{w}\|$ and the inequality asserts that the length of one side of the triangle is smaller than the sum of the other two lengths.

A related argument shows:

Lemma 5.3. *For all* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$,

$$|\|\mathbf{v}\| - \|\mathbf{w}\|| \leq \|\mathbf{v} - \mathbf{w}\|.$$

Proof. Apply the triangle inequality to $\mathbf{u} = \mathbf{v} - \mathbf{w}$ and $\mathbf{w}$: we get

$$\|\mathbf{v}\| = \|\mathbf{u} + \mathbf{w}\| \leq \|\mathbf{u}\| + \|\mathbf{w}\| = \|\mathbf{v} - \mathbf{w}\| + \|\mathbf{w}\|.$$

Thus

$$\|\mathbf{v}\| - \|\mathbf{w}\| \leq \|\mathbf{v} - \mathbf{w}\|.$$

Exchanging the roles of $\mathbf{v}$ and $\mathbf{w}$ gives

$$\|\mathbf{w}\| - \|\mathbf{v}\| \leq \|\mathbf{w} - \mathbf{v}\| = \|\mathbf{v} - \mathbf{w}\|.$$

So $\pm(\|\mathbf{w}\| - \|\mathbf{v}\|) \leq \|\mathbf{v} - \mathbf{w}\|$, which says that $\Big|\|\mathbf{v}\| - \|\mathbf{w}\|\Big| \leq \|\mathbf{v} - \mathbf{w}\|$. □

We view $\|\mathbf{v}\|$ as the distance from $\mathbf{v}$ to the origin and $\|\mathbf{v} - \mathbf{w}\|$ as the distance from $\mathbf{v}$ to $\mathbf{w}$. We also have the following interpretation for the inner product in $\mathbb{R}^2$ and $\mathbb{R}^3$ (and by extension in general): If θ is the angle made by $\mathbf{v}$ and $\mathbf{w}$, then

$$\cos\theta = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\|\|\mathbf{w}\|}.$$

Note that both sides are meaningless if one of $\mathbf{v}$ or $\mathbf{w}$ is zero. Also, the Cauchy-Schwarz inequality assures us that the absolute value of the right hand side is always at most one (if it is defined), and so it is of the form $\cos\theta$ for some real number θ. In particular $\mathbf{v}$ and $\mathbf{w}$ are perpendicular if and only if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$.

For an arbitrary field F, it is possible for $\langle \mathbf{v}, \mathbf{v} \rangle$ to equal 0. For example, this happens for $F = \mathbb{C}$ and $\mathbf{v} = (1, i)$.

5.2 Orthogonality

In this section, we shall discuss results about inner products which are purely algebraic and which work over any field.

Definition 5.4. Let $\mathbf{v}$ and $\mathbf{w} \in \mathbb{R}^n$. Then $\mathbf{v}$ and $\mathbf{w}$ are *orthogonal* if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$.

Given $\mathbf{v} \in \mathbb{R}^n$, define $\mathbf{v}^\perp = \{\mathbf{w} \in \mathbb{R}^n : \langle \mathbf{v}, \mathbf{w} \rangle = 0\}$. More generally, for $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, define

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp = \{\mathbf{w} \in \mathbb{R}^n : \langle \mathbf{v}_i, \mathbf{w} \rangle = 0 \text{ for all } i\}.$$

Clearly

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp = \mathbf{v}_1^\perp \cap \cdots \cap \mathbf{v}_k^\perp.$$

If $F(\mathbf{x}) = (\langle \mathbf{r}_1, \mathbf{x}\rangle, \ldots, \langle \mathbf{r}_k, \mathbf{x}\rangle)$ is the linear map corresponding to the matrix A, where the $\mathbf{r}_i$ are the rows of A, then clearly

$$\operatorname{Ker} F = \{\mathbf{r}_1, \ldots, \mathbf{r}_k\}^\perp.$$

We can define $X^\perp$ similarly for every subset X of $\mathbb{R}^n$:

$$X^\perp = \{\mathbf{v} \in \mathbb{R}^n : \langle \mathbf{v}, \mathbf{x}\rangle = 0 \text{ for all } \mathbf{x} \in X\}.$$

Note that $(\mathbb{R}^n)^\perp = \{\mathbf{0}\}$ and that $\{\mathbf{0}\}^\perp = \mathbb{R}^n$. Likewise, by logic $\emptyset^\perp = \mathbb{R}^n$.

Lemma 5.5. *If $X \subseteq \mathbb{R}^n$, then $X^\perp$ is a vector subspace of $\mathbb{R}^n$. If $X_1 \subseteq X_2$, then $X_2^\perp \subseteq X_1^\perp$. Finally, $X \subseteq X^{\perp\perp}$.*

Proof. For arbitrary X, we always have $\mathbf{0} \in X^\perp$. If $\mathbf{v}, \mathbf{w} \in X^\perp$, then $\mathbf{v} + \mathbf{w}$ and $t\mathbf{v} \in X^\perp$. Thus $X^\perp$ is a vector subspace of $\mathbb{R}^n$. The second statement is clear. To see the last statement, note that, by definition,

$$X^{\perp\perp} = \{\mathbf{y} \in \mathbb{R}^n : \text{ if } \langle \mathbf{v}, \mathbf{x}\rangle = 0 \text{ for all } \mathbf{x} \in X, \text{ then } \langle \mathbf{v}, \mathbf{y}\rangle = 0\}.$$

In other words, $X^{\perp\perp}$ is the set of all vectors which are orthogonal to all vectors which are orthogonal to every vector in X. Thus $X \subseteq X^{\perp\perp}$. □

In particular we can define $V^\perp$ if V is a vector subspace of $\mathbb{R}^n$. However, if $V = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, then $V^\perp = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$:

Lemma 5.6. $(\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})^\perp = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$.

Proof. Using the fact that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subseteq \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and the last statement in the above lemma, we see that $(\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})^\perp \subseteq \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$. On the other hand, if $\mathbf{v} \in \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$, then $\langle \mathbf{v}, \mathbf{v}_i\rangle = 0$ for all i. Thus $\mathbf{v}$ is orthogonal to every linear combination of the $\mathbf{v}_i$, by expanding out, and so $\mathbf{v} \in (\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})^\perp$. It follows that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp \subseteq (\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})^\perp$ and so $(\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})^\perp = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$. □

Our main result about orthogonal spaces is the following:

Theorem 5.7. *If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent vectors in $\mathbb{R}^n$, then*

$$\dim\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp = n - k.$$

Equivalently, if V is a vector subspace of $\mathbb{R}^n$, then

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

To prove the theorem, we first show:

Proposition 5.8. *Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ are linearly independent. Then for every $i, 1 \leq i \leq k$, there exists a $\mathbf{c}_i \in \mathbb{R}^n$ such that $\langle \mathbf{c}_i, \mathbf{v}_i \rangle = 1$ and $\langle \mathbf{c}_i, \mathbf{v}_j \rangle = 0$ if $i \neq j$.*

Proof. Complete the linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ to a basis of $\mathbb{R}^k$, say $\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$. For every i such that $1 \leq i \leq k$, define a linear map $F_i \colon \mathbb{R}^n \to \mathbb{R}$ as follows: $F_i(\mathbf{v}_i) = 1$ and $F_i(\mathbf{v}_j) = 0$ if $j \neq i$. That such a linear map exists follows from Lemma 3.11, which says more generally that we can specify a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ by specifying its values on a basis of $\mathbb{R}^n$. But we have seen that every map $F \colon \mathbb{R}^n \to \mathbb{R}$ is of the form $F(\mathbf{x}) = \langle \mathbf{x}, \mathbf{c} \rangle$ for a unique $\mathbf{c} \in \mathbb{R}^n$. For our particular choice of F_i, this then says that $F_i(\mathbf{x}) = \langle \mathbf{x}, \mathbf{c}_i \rangle$. Hence, by the choice of F_i, we must have $\langle \mathbf{v}_j, \mathbf{c}_i \rangle = 0$ for all $j \neq i$ and $\langle \mathbf{v}_i, \mathbf{c}_i \rangle = 1$. □

Let us now give the proof of the theorem above:

Proof of the theorem. Given $\mathbf{v}_1, \ldots, \mathbf{v}_k$ linearly independent in $\mathbb{R}^n$, define a map $F \colon \mathbb{R}^n \to \mathbb{R}^k$ by $F(\mathbf{x}) = (\langle \mathbf{v}_1, \mathbf{x} \rangle, \ldots, \langle \mathbf{v}_k, \mathbf{x} \rangle)$. Then F is linear, and its kernel is just $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$. Moreover we claim that F is onto $\mathbb{R}^k$. Indeed, choosing $\mathbf{c}_i$ as in the above proposition, we see that the image of F contains $\mathbf{e}_i$ for every i and is thus equal to $\mathbb{R}^k$. Thus $n = \dim \mathbb{R}^n = \dim \operatorname{Im} F + \dim \operatorname{Ker} F = k + \dim\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp$, so that $\dim\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^\perp = n - k$. □

Let us see the connection with linear equations and linear maps. Let A be a $k \times n$ matrix corresponding to a linear map from $\mathbb{R}^n$ to $\mathbb{R}^k$. We can think of A as built up out of n column vectors $\mathbf{c}_1, \ldots \mathbf{c}_n \in \mathbb{R}^k$ or out of k row vectors $\mathbf{r}_1, \ldots, \mathbf{r}_k \in \mathbb{R}^n$. Thus A is a linear map either by the formula $A \cdot \mathbf{x} = \sum_{i=1}^n x_i \mathbf{c}_i$ or $A \cdot \mathbf{x} = (\langle \mathbf{x}, \mathbf{r}_1 \rangle, \ldots, \langle \mathbf{x}, \mathbf{r}_k \rangle)$. The image of A is the span of $\mathbf{c}_1, \ldots, \mathbf{c}_n$ and clearly the kernel of A is $\{\mathbf{r}_1, \ldots, \mathbf{r}_k\}^\perp$. Thus if we want to describe the kernel of A as a subspace of $\mathbb{R}^n$, we will need an efficient way to describe $\{\mathbf{r}_1, \ldots, \mathbf{r}_k\}^\perp$. Instead of trying to write down a general recipe, we will just give an example: suppose that we have applied row reduction to the **rows** $\mathbf{r}_1, \ldots, \mathbf{r}_n$ of A. Be sure not to confuse this with the problem of finding the **image** of A, where we apply row reduction to the **columns** of A. Since row reduction doesn't change the span of $\mathbf{r}_1, \ldots, \mathbf{r}_k$, it is enough to find what is perpendicular to the end result, and this is easy to do in practice. For example, first suppose that $\mathbf{r}_1 = (1, 0, a_1, a_2, 0)$ and $\mathbf{r}_2 = (0, 1, b_1, b_2, 0)$. Then $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ is perpendicular to both $\mathbf{r}_1$

and $\mathbf{r}_2$ if and only if:

$$x_1 + a_1x_3 + a_2x_4 = 0;$$
$$x_2 + b_1x_3 + b_2x_4 = 0.$$

This says that x_3, x_4, x_5 can be arbitrary and that $x_1 = -a_1x_3 - a_2x_4$, $x_2 = -b_1x_3 - b_2x_4$. Taking $x_3 = 1, x_4 = x_5 = 0$ gives the vector $(-a_1, -b_1, 1, 0, 0)$. Taking $x_4 = 1, x_3 = x_5 = 0$ gives the vector $(-a_2, -b_2, 0, 1, 0)$. Taking $x_5 = 1, x_3 = x_4 = 0$ gives the vector $(0, 0, 0, 0, 1)$. It is then easy to check that

$$\{(1, 0, a_1, a_2, 0), (0, 1, b_1, b_2, 0)\}^\perp =$$
$$\text{span}\{(-a_1, -b_1, 1, 0, 0), (-a_2, -b_2, 0, 1, 0), (0, 0, 0, 0, 1)\}.$$

Once we have found three linearly independent vectors in $V = \{\mathbf{r}_1, \mathbf{r}_2\}^\perp$, we have in fact found a basis of V since $\dim V = 5 - 2 = 3$. It is easy to check that the three vectors we found above are linearly independent (because of the arrangement of 1's and 0's). Thus we have indeed found a basis for V. For a somewhat more general example, let us find a basis for $\text{Ker}\, A$, where $A = \begin{pmatrix} 1 & 5 & 7 & -2 \\ 3 & -1 & 1 & 2 \end{pmatrix}$. We thus want to find a basis for $\{\mathbf{r}_1, \mathbf{r}_2\}^\perp$, where $\mathbf{r}_1 = (1, 5, 7, -2)$ and $\mathbf{r}_2 = (3, -1, 1, 2)$. Applying row reduction to the row vectors $\mathbf{r}_1, \mathbf{r}_2$ gives:

$$\begin{matrix} 1 & 5 & 7 & -2 \\ 3 & -1 & 1 & 2 \end{matrix} \quad \to \quad \begin{matrix} 1 & 5 & 7 & -2 \\ 0 & -16 & -20 & 8 \end{matrix}$$
$$\to \quad \begin{matrix} 1 & 5 & 7 & -2 \\ 0 & 1 & \frac{5}{4} & -\frac{1}{2} \end{matrix} \quad \to \quad \begin{matrix} 1 & 0 & \frac{3}{4} & \frac{1}{2} \\ 0 & 1 & \frac{5}{4} & -\frac{1}{2} \end{matrix} .$$

Thus $\text{Ker}\, A = \{(x_1, x_2, x_3, x_4) : x_3 = \frac{3}{4}x_1 + \frac{5}{4}x_2, x_4 = \frac{1}{2}x_2 - \frac{1}{2}x_2\}$ and a basis is given by $(-\frac{3}{4}, -\frac{5}{4}, 1, 0)$, $(-\frac{1}{2}, \frac{1}{2}, 0, 1)$.

We return now to the study of a general matrix A.

Definition 5.9. Let A be a $k \times n$ matrix. The *rank* of A is the dimension of $\text{Im}\, A$, i.e. the dimension of the span of the n column vectors $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathbb{R}^k$.

We then have the (non-obvious) result:

Proposition 5.10. *The rank of A is equal to the dimension of the span of the rows of A. In other words, the dimension of the span of the n column vectors $\mathbf{c}_1, \ldots, \mathbf{c}_n$ in $\mathbb{R}^k$ is equal to the dimension of the span of the k row vectors $\mathbf{r}_1, \ldots, \mathbf{r}_k$ in $\mathbb{R}^n$.*

Proof. First we know that $\dim \operatorname{Im} A + \dim \operatorname{Ker} A = n$. Now let W be the span of the rows $\mathbf{r}_1, \ldots, \mathbf{r}_k$. Then $\operatorname{Ker} A = W^\perp$. But we also know that $\dim W + \dim W^\perp = n$. This says that

$$\begin{aligned} n &= \dim \operatorname{Im} A + \dim \operatorname{Ker} A \\ &= \dim \operatorname{Im} A + \dim W^\perp \\ &= \dim W + \dim W^\perp. \end{aligned}$$

Thus $\dim \operatorname{Im} A = \dim W$, which is the statement of the proposition. □

Proposition 5.11. *For $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}^{\perp\perp} = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Equivalently, if V is a vector subspace of $\mathbb{R}^n$, then $V^{\perp\perp} = V$.*

Proof. We shall prove the second version of the proposition. Suppose that V is a vector subspace of $\mathbb{R}^n$ with $\dim V = d$. Then $\dim V^\perp = n - d$, by Theorem 5.7, and so $\dim V^{\perp\perp} = d$. Clearly $V \subseteq V^{\perp\perp}$, and since they have the same dimension they must be equal. □

The above tells us how to write a vector subspace V as the set of zeroes of a set of linear equations. Suppose that $\dim V = k$. Choose a basis $\mathbf{w}_1, \ldots, \mathbf{w}_{n-k}$ for $V^\perp$. Then $V = \{\mathbf{w}_1, \ldots, \mathbf{w}_{n-k}\}^\perp$. In other words, V is defined by

$$V = \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{w}_1 \rangle = \cdots = \langle \mathbf{x}, \mathbf{w}_{n-k} \rangle = 0\}.$$

These are $n - k$ linear equations defining V (and $n - k$ is the minimum number of equations needed to define V—why?) To sum up, then: a vector subspace V of $\mathbb{R}^n$ may be described either parametrically, as the span of a sequence of vectors, or by linear equations:

Theorem 5.12. *Let V be a vector subspace of $\mathbb{R}^n$ of dimension k.*

1. *There exist $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$ such that $V = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Moreover, every set of vectors in V which spans V has at least k elements.*

2. *There exist $\mathbf{w}_1, \ldots, \mathbf{w}_{n-k} \in \mathbb{R}^n$ such that $V = \{\mathbf{w}_1, \ldots, \mathbf{w}_{n-k}\}^\perp$, and every set of vectors in $\mathbb{R}^n$ with this property has at least $n-k$ elements.*

□

5.3 Orthonormal bases

In this section, we describe results which are special to $\mathbb{R}^n$.

Lemma 5.13. *Let V be a vector subspace of $\mathbb{R}^n$. Then $\mathbb{R}^n = V \oplus V^\perp$. In other words, every vector $\mathbf{x}$ in $\mathbb{R}^n$ can be uniquely written as $\mathbf{x} = \mathbf{v} + \mathbf{w}$, where $\mathbf{v} \in V$ and $\mathbf{v} \in V^\perp$.*

Proof. First note that, if $\mathbf{v} \in V \cap V^\perp$, then by definition $\langle \mathbf{v}, \mathbf{v} \rangle = 0$, and hence $\mathbf{v} = 0$. It follows that $V + V^\perp = V \oplus V^\perp$. Now $\dim(V \oplus V^\perp) = \dim V + \dim V^\perp$. If $\dim V = d$, then $\dim V^\perp = n - d$ and hence $\dim V + \dim V^\perp = n$. It follows that the subspace $V \oplus V^\perp$ of $\mathbb{R}^n$ has dimension n, and hence $V \oplus V^\perp = \mathbb{R}^n$. □

Given a vector subspace V and a vector $\mathbf{x}$ in $\mathbb{R}^n$. we want an explicit recipe for writing $\mathbf{x} = \mathbf{v} + \mathbf{w}$, where $\mathbf{v} \in V$ and $\mathbf{v} \in V^\perp$. This is the analogue of "dropping a perpendicular from a point to a line," familiar from plane geometry. We begin as follows: Given a line through the origin spanned by $\mathbf{v}$, we can always make $\mathbf{v}$ of unit length by dividing by its length. We want a similar procedure for dealing with vector subspaces of larger dimension.

Definition 5.14. A basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ for $\mathbb{R}^n$ is a *orthonormal* basis if $\|\mathbf{u}_i\| = 1$ for all i and $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for all i and j if $i \neq j$. An orthonormal basis for a vector subspace V of $\mathbb{R}^n$ is similarly defined.

Given an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$, for every vector $\mathbf{v} \in \mathbb{R}^n$, we call $\langle \mathbf{v}, \mathbf{u}_i \rangle$ the *component* of V along $\mathbf{u}_i$. It is easy to check the following:

Proposition 5.15. *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis of $\mathbb{R}^n$. Then, for every $\mathbf{v} \in \mathbb{R}^n$,*

1. $\mathbf{v} = \sum_{i=1}^n \langle \mathbf{v}, \mathbf{u}_i \rangle \mathbf{u}_i$;
2. $\|\mathbf{v}\|^2 = \sum_{i=1}^n (\langle \mathbf{v}, \mathbf{u}_i \rangle)^2$.

Proof. To see (1), consider $\mathbf{v} - \sum_{i=1}^n \langle \mathbf{v}, \mathbf{u}_i \rangle \mathbf{u}_i = \mathbf{w}$. Then

$$\langle \mathbf{w}, \mathbf{u}_j \rangle = \langle \mathbf{v}, \mathbf{u}_j \rangle - \sum_i \langle \mathbf{v}, \mathbf{u}_i \rangle \langle \mathbf{u}_i, \mathbf{u}_j \rangle.$$

Since $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ if $i \neq j$ and $= 1$ if $i = j$, the sum reduces to $\langle \mathbf{v}, \mathbf{u}_i \rangle$. Thus the above expression is zero. This says that $\mathbf{w}$ is orthogonal to $\mathbf{u}_j$ for every j, and so is orthogonal to the span of the $\mathbf{u}_j$, i.e. to $\mathbb{R}^n$. Hence $\mathbf{w} = \mathbf{0}$, so that $\mathbf{v} = \sum_{i=1}^n \langle \mathbf{v}, \mathbf{u}_i \rangle \mathbf{u}_i$.

To see (2), consider

$$\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = \langle \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{u}_i \rangle \mathbf{u}_i, \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{u}_i \rangle \mathbf{u}_i \rangle$$
$$= \sum_{i,j} \langle \mathbf{v}, \mathbf{u}_i \rangle \langle \mathbf{v}, \mathbf{u}_j \rangle \langle \mathbf{u}_i, \mathbf{u}_j \rangle.$$

Again using $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ if $i \neq j$ and $= 1$ if $i = j$, the sum reduces to $\sum_i \langle \mathbf{v}, \mathbf{u}_i \rangle$. □

The meaning of the above proposition is as follows: in general, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathbb{R}^n$, we know that we can write every vector $\mathbf{v}$ as a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$: $\mathbf{v} = \sum_{i=1}^n t_i \mathbf{v}_i$. But it is hard to find the numbers t_i explicitly. In fact, if A is the matrix whose columns are given by the $\mathbf{v}_i$, then the equation $\mathbf{v} = \sum_{i=1}^n t_i \mathbf{v}_i$ says that $\mathbf{v} = A\mathbf{t}$, and equivalently that $\mathbf{t} = A^{-1} \cdot \mathbf{v}$. Thus, in practice, to find the t_i we have to find A^{-1}. But for an orthonormal basis, the recipe is much simpler. Likewise, (2) says that it is easy to find the length of a vector $\mathbf{v} = \sum_{i=1}^n t_i \mathbf{u}_i$: it is just $\sqrt{t_1^2 + \cdots + t_n^2}$. (In fancier terms, this says that the linear map F defined by the $\mathbf{u}_i$ is an *isometry*.) Again, for an arbitrary basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$, there is no simple formula which gives the length of $\sum_{i=1}^n t_i \mathbf{v}_i$ in terms of the t_i.

We have a procedure (the Gram-Schmidt procedure) for starting with an arbitrary basis and making it orthonormal. The procedure is as follows: suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a set of linearly independent vectors in $\mathbb{R}^n$ (not necessarily a basis). Let $\mathbf{u}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|$. Then $\|\mathbf{u}_1\| = \|\mathbf{v}_1\| / \|\mathbf{v}_1\| = 1$. Next take $\mathbf{v}_2$, and replace it by $\mathbf{v}_2' = \mathbf{v}_2 - p_{\mathbf{u}_1}(\mathbf{v}_2) = \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1$. Since $\mathbf{v}_2'$ is the difference of $\mathbf{v}_2$ and a multiple of $\mathbf{u}_1$,

$$\operatorname{span}\{\mathbf{u}_1, \mathbf{v}_2'\} = \operatorname{span}\{\mathbf{u}_1, \mathbf{v}_2\} = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}.$$

In particular $\dim \operatorname{span}\{\mathbf{u}_1, \mathbf{v}_2'\} = 2$, so that $\mathbf{u}_1$ and $\mathbf{v}_2'$ are linearly independent (and hence $\mathbf{v}_2' \neq \mathbf{0}$). Now $\langle \mathbf{v}_2', \mathbf{u}_1 \rangle = \langle \mathbf{v}_2, \mathbf{u}_1 \rangle - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle = 0$. Thus $\mathbf{v}_2'$ and $\mathbf{u}_1$ are two linearly independent, orthogonal vectors. Of course, $\mathbf{v}_2'$ need not have unit length, so replace it by $\mathbf{v}_2' / \|\mathbf{v}_2'\| = \mathbf{u}_2$.

By induction, suppose that we have found $\mathbf{u}_1, \ldots, \mathbf{u}_i$ such that the $\mathbf{u}_j$, $j \leq i$ are orthonormal (i.e. are an orthonormal basis of their span) and $\operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_i\} = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_i\}$. If $i < k$, define $\mathbf{v}_{i+1}' = \mathbf{v}_{i+1} - \sum_{j=1}^i \langle \mathbf{v}_{i+1}, \mathbf{u}_j \rangle \mathbf{u}_j$. Clearly

$$\operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_i, \mathbf{v}_{i+1}'\} = \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_i, \mathbf{v}_{i+1}\} = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{i+1}\}.$$

Thus $\mathbf{u}_1, \ldots, \mathbf{u}_i, \mathbf{v}'_{i+1}$ are linearly independent (and in particular $\mathbf{v}'_{i+1} \neq \mathbf{0}$). Moreover, for $j \leq i$, since $\langle \mathbf{u}_j, \mathbf{u}_\ell \rangle = 0$ if $j \neq \ell$ and $= 1$ if $j = \ell$, we see that

$$\langle \mathbf{v}'_{i+1}, \mathbf{u}_\ell \rangle = \langle \mathbf{v}_{i+1}, \mathbf{u}_\ell \rangle - \langle \mathbf{v}_{i+1}, \mathbf{u}_\ell \rangle = 0.$$

Thus $\mathbf{v}_{i+1}$ is orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_i$. Finally, replace $\mathbf{v}'_{i+1}$ by $\mathbf{u}_{i+1} = \mathbf{v}'_{i+1}/\|\mathbf{v}'_{i+1}\|$ to get the next element in the orthonormal set.

What this procedure shows in particular is the following:

Proposition 5.16. *1. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis of $\mathbb{R}^n$. Then there exists an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathbb{R}^n$ such that, for every i with $1 \leq i \leq n$,*

$$\operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_i\} = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_i\}.$$

2. Every vector subspace V of $\mathbb{R}^n$ has an orthonormal basis.

3. Every set $\mathbf{u}_1, \ldots \mathbf{u}_k$ of vectors such that $\|\mathbf{u}_i\| = 1$ for all i and $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for all i and j if $i \neq j$ can be completed to an orthonormal basis for $\mathbb{R}^n$.

Here is how this works in practice: let us apply Gram-Schmidt to the vectors $\mathbf{v}_1 = (1, 2, 2)$, $\mathbf{v}_2 = (2, 0, 2)$, $\mathbf{v}_3 = (1, 1, 7)$. We replace $\mathbf{v}_1$ by $\mathbf{u}_1 = (\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$. Now take

$$\begin{aligned}\mathbf{v}'_2 &= \mathbf{v}_2 - p_{\mathbf{u}_1}(\mathbf{v}_2) = \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1 \\ &= (2, 0, 2) - (\frac{2}{3} + \frac{4}{3})(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}) = (2, 0, 2) - (\frac{2}{3}, \frac{4}{3}, \frac{4}{3}) = (\frac{4}{3}, -\frac{4}{3}, \frac{2}{3}).\end{aligned}$$

Note that $\mathbf{v}'_2$ is indeed orthogonal to $\mathbf{v}_1$. Replace $\mathbf{v}'_2$ by the unit vector $\mathbf{u}_2$ pointing in the same direction. Since $\mathbf{v}'_2 = \frac{2}{3}(2, -2, 1)$, clearly $\mathbf{u}_2 = (\frac{2}{3}, -\frac{2}{3}, \frac{1}{3})$. Next we take

$$\begin{aligned}\mathbf{v}'_3 &= \mathbf{v}_3 - p_{\mathbf{u}_1}(\mathbf{v}_3) - p_{\mathbf{u}_1}(\mathbf{v}_3) \\ &= (1, 1, 7) - \frac{1}{9}(17)(1, 2, 2) - \frac{1}{9}(7)(2, -2, 1) = (1, 1, 7) - \frac{1}{9}(31, 20, 41) \\ &= \frac{1}{9}[(9, 9, 63) - (31, 20, 41)] = \frac{11}{9}(-2, -1, 2).\end{aligned}$$

Up to a positive scalar, this gives $\mathbf{v}'_3 = (-2, -1, 2)$, which is orthogonal to $(1, 2, 2)$ and $(2, 0, 2)$, and making it of unit length gives $\mathbf{u}_3 = (-\frac{2}{3}, -\frac{1}{3}, \frac{2}{3})$. Of course, in most examples there will be many square roots to contend with.

Next let us define the orthogonal projection onto the line spanned by a nonzero vector $\mathbf{v}$. Fix $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v} \neq \mathbf{0}$. Define

$$p_{\mathbf{v}}(\mathbf{w}) = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\|^2} \mathbf{v}.$$

Note that if $\|\mathbf{v}\| = 1$, so that $\mathbf{v}$ has unit length, then this has the simpler form $p_{\mathbf{v}}(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle \mathbf{v}$. We call $p_{\mathbf{v}}$ the *projection of* $\mathbf{v}$ *along* $\mathbf{w}$ (Figure 5.1). It is a linear map from $\mathbb{R}^n$ to itself whose kernel is $\mathbf{v}^\perp$ and whose image is $L_{\mathbf{v}}$, the line spanned by $\mathbf{v}$. Moreover $p_{\mathbf{v}}(\mathbf{w}) = \mathbf{w}$ if and only if $\mathbf{w} \in L_{\mathbf{v}}$. We also have the easy calculation: for all $\mathbf{w} \in \mathbb{R}^n$,

$$\langle \mathbf{w} - p_{\mathbf{v}}(\mathbf{w}), \mathbf{v} \rangle = 0.$$

Hence, when writing $\mathbf{w} = (\mathbf{w} - p_{\mathbf{v}}(\mathbf{w})) + p_{\mathbf{v}}(\mathbf{w})$, we see that we have written $\mathbf{w}$ as the sum of a vector in $L_{\mathbf{v}}^\perp$ plus the sum of a vector in $L_{\mathbf{v}}$. By (2) of the lemma above, there is a unique way of doing this. It is also easy to check that $\|\mathbf{w} - p_{\mathbf{v}}(\mathbf{w})\|$ is the distance from $\mathbf{w}$ to $L_{\mathbf{v}}$, in the sense that

$$\|\mathbf{w} - p_{\mathbf{v}}(\mathbf{w})\| = \min_{t \in \mathbb{R}} \|\mathbf{w} - t\mathbf{v}\|.$$

Indeed, to minimize $\|\mathbf{w} - t\mathbf{v}\|$, it suffices to minimize $\|\mathbf{w} - t\mathbf{v}\|^2 = \|\mathbf{w}\|^2 - 2t\langle \mathbf{w}, \mathbf{v} \rangle + t^2 \|\mathbf{v}\|^2$. Differentiating with respect to t and setting the result equal to zero gives

$$t = \frac{\langle \mathbf{w}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}.$$

This value of t gives $\mathbf{w} - t\mathbf{v} = \mathbf{w} - p_{\mathbf{v}}(\mathbf{w})$. (Another proof which does not use any calculus is given later in the chapter, in Proposition 5.18.)

Using orthonormal bases, we can generalize the definition of projection of a vector onto a line to projection of a vector onto a subspace as follows. Let V be a vector subspace of $\mathbb{R}^n$. Then we can choose an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_k$ for V. We define the projection p_V of $\mathbb{R}^n$ onto V as follows:

$$p_V(\mathbf{w}) = \sum_{i=1}^{k} \langle \mathbf{w}, \mathbf{u}_i \rangle \mathbf{u}_i.$$

We have the following properties for p_V:

Proposition 5.17. *1.* $p_V(\mathbf{w}) = \mathbf{w}$ *if and only if* $\mathbf{w} \in V$*;*

2. $\operatorname{Im} p_V = V$*:*

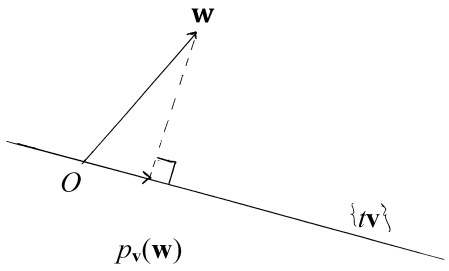

Figure 5.1: Dropping a perpendicular from a point to a line

3. $\operatorname{Ker} p_V = V^\perp$;

4. $\mathbf{w} - p_V(\mathbf{w}) \in V^\perp$ *for all* $\mathbf{w} \in \mathbb{R}^n$;

5. *For all* $\mathbf{w} \in \mathbb{R}^n$,
$$\mathbf{w} = (\mathbf{w} - p_V(\mathbf{w})) + p_V(\mathbf{w}),$$
where the first term is in $V^\perp$ *and the second is in* V.

6. *The linear map* p_V *is independent of the choice of orthonormal basis of* V.

Proof. (1) Since $\mathbf{u}_i \in V$, $p_V(\mathbf{w}) \in V$ for all $\mathbf{w} \in \mathbb{R}^n$. Now $\langle \mathbf{w} - p_V(\mathbf{w}), \mathbf{u}_i \rangle = \langle \mathbf{w}, \mathbf{u}_i \rangle - \langle \mathbf{w}, \mathbf{u}_i \rangle = 0$. Thus $\mathbf{w} - p_V(\mathbf{w})$ is orthogonal to $\mathbf{u}_i$ for every i, and thus $\mathbf{w} - p_V(\mathbf{w}) \in \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}^\perp = V^\perp$, proving (4). If in addition $\mathbf{w} \in V$, then $\mathbf{w} - p_V(\mathbf{w}) \in V$ and so $\mathbf{w} - p_V(\mathbf{w}) \in V \cap V^\perp = \{\mathbf{0}\}$. Thus $\mathbf{w} = p_V(\mathbf{w})$, proving (1). To prove (2), we have seen that $p_V(\mathbf{w}) \in V$ for all $\mathbf{w} \in \mathbb{R}^n$, so that $\operatorname{Im} p_V \subseteq V$. Since $\mathbf{w} = p_V(\mathbf{w})$ if $\mathbf{w} \in V$, we see that $V \subseteq \operatorname{Im} p_V$, and thus $\operatorname{Im} p_V = V$. If $\mathbf{w} \in \operatorname{Ker} p_V$, then (as $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is a basis for V) $\langle \mathbf{w}, \mathbf{u}_i \rangle = 0$ for all i, so that $\mathbf{w} \in \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}^\perp = V^\perp$. Thus $\operatorname{Ker} p_V \subseteq V^\perp$. Conversely, if $\mathbf{w} \in V^\perp$, then clearly $p_V(\mathbf{w}) = \mathbf{0}$. Thus

$\operatorname{Ker} p_V = V^\perp$, proving (3). (5) follows from (4) and (2). Finally, if $\{\mathbf{u}'_i\}$ is another choice of orthonormal basis for V and p'_V is defined via the basis $\{\mathbf{u}'_i\}$, then we have

$$\mathbf{w} = (\mathbf{w} - p_V(\mathbf{w})) + p_V(\mathbf{w}) = (\mathbf{w} - p'_V(\mathbf{w})) + p'_V(\mathbf{w}).$$

Applying Lemma 5.13, we conclude that $p_V(\mathbf{w}) = p'_V(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{R}^n$. Thus $p_V = p'_V$. □

As a simple consequence of (2) and (3) above, we can find a different proof of the fact that, for every vector subspace V of $\mathbb{R}^n$, $\dim V + \dim V^\perp = n$. Indeed, consider the map $p_V : \mathbb{R}^n \to \mathbb{R}^n$. Then $n = \dim \operatorname{Ker} p_V + \dim \operatorname{Im} p_V = \dim V^\perp + \dim V$.

We also note that projections give the shortest distance from a vector $\mathbf{w}$ to a subspace V:

Proposition 5.18. *If V is a vector subspace of $\mathbb{R}^n$, then, for all $\mathbf{w} \in \mathbb{R}^n$,*

$$\|\mathbf{w} - p_V(\mathbf{w})\| = \min_{\mathbf{v} \in V} \|\mathbf{w} - \mathbf{v}\|.$$

Proof. Write $\mathbf{w} = \mathbf{w} - p_V(\mathbf{w}) + p_V(\mathbf{w})$, where $p_V(\mathbf{w}) \in V$ and $\mathbf{w} - p_V(\mathbf{w}) \in V^\perp$. Thus, for $\mathbf{v} \in V$, $\mathbf{w} - \mathbf{v} = \mathbf{w} - p_V(\mathbf{w}) + p_V(\mathbf{w}) - \mathbf{v}$. If we set $\mathbf{u}_1 = \mathbf{w} - p_V(\mathbf{w})$ and $\mathbf{u}_2 = p_V(\mathbf{w}) - \mathbf{v}$, then this says that $\mathbf{w} - \mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2$ with $\mathbf{u}_1 \in V^\perp$ and $\mathbf{u}_2 \in V$. By the Pythagorean theorem (exercise) $\|\mathbf{w} - \mathbf{v}\|^2 = \|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2$. Now $\|\mathbf{u}_1\|^2 = \|\mathbf{w} - p_V(\mathbf{w})\|^2$ and $\|\mathbf{u}_2\|^2 = \|p_V(\mathbf{w}) - \mathbf{v}\|^2$. Thus $\|\mathbf{u}_2\|^2 \geq 0$ and $\|\mathbf{u}_2\|^2 = 0$ if and only if $\mathbf{u}_2 = \mathbf{0}$, or $\mathbf{v} = p_V(\mathbf{w})$. Thus

$$\|\mathbf{w} - \mathbf{v}\|^2 \geq \|\mathbf{w} - p_V(\mathbf{w})\|^2$$

for all $\mathbf{v} \in V$, with equality if and only if $\mathbf{v} = p_V(\mathbf{w})$. □

5.4 Bilinear and quadratic forms

Bilinear forms (and especially symmetric positive definite bilinear forms) are a generalization of inner products. More generally, symmetric bilinear forms are a way of describing all degree two quadratic polynomials in several variables with no linear or constant terms, much as linear functions from $\mathbb{R}^n$ to $\mathbb{R}$ are the same thing as polynomials of degree one with no constant term. (In what follows, "form" is just one of the many synonyms for function, which it is traditional to use in this context.)

Definition 5.19. A *bilinear form* (or *bilinear function*) on $\mathbb{R}^n$ is a function $B\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfying:

(i) For all $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w} \in \mathbb{R}^n$,

$$B(\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}) = B(\mathbf{v}_1, \mathbf{w}) + B(\mathbf{v}_2, \mathbf{w}).$$

(ii) For all $\mathbf{v}, \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$,

$$B(\mathbf{v}, \mathbf{w}_1 + \mathbf{w}_2) = B(\mathbf{v}, \mathbf{w}_1) + B(\mathbf{v}, \mathbf{w}_2).$$

(iii) For all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and $t \in \mathbb{R}$,

$$B(t\mathbf{v}, \mathbf{w}) = B(\mathbf{v}, t\mathbf{w}) = tB(\mathbf{v}, \mathbf{w}).$$

Thus a bilinear form is a function $B\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that, for each fixed $\mathbf{w} \in \mathbb{R}^n$, the function $L_{\mathbf{w}}(\mathbf{x}) = B(\mathbf{x}, \mathbf{w})\colon \mathbb{R}^n \to \mathbb{R}$ is a linear function, and similarly for the function $M_{\mathbf{v}}(\mathbf{y}) = B(\mathbf{v}, \mathbf{y})\colon \mathbb{R}^n \to \mathbb{R}$.

A bilinear form $B(\mathbf{v}, \mathbf{w})$ is *symmetric* if, for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $B(\mathbf{v}, \mathbf{w}) = B(\mathbf{w}, \mathbf{v})$. A symmetric bilinear form B is *positive definite* if, for all $\mathbf{v} \in \mathbb{R}^n$, $B(\mathbf{v}, \mathbf{v}) \geq 0$, and $B(\mathbf{v}, \mathbf{v}) = 0 \iff \mathbf{v} = \mathbf{0}$. The form B is *positive semi-definite* if, for all $\mathbf{v} \in \mathbb{R}^n$, $B(\mathbf{v}, \mathbf{v}) \geq 0$, but we allow for the possibility that $B(\mathbf{v}, \mathbf{v}) = 0$ for some nonzero $\mathbf{v}$. Negative definite and negative semi-definite are defined in the same way: B is *negative definite* if, for all $\mathbf{v} \in \mathbb{R}^n$, $B(\mathbf{v}, \mathbf{v}) \leq 0$, and $B(\mathbf{v}, \mathbf{v}) = 0 \iff \mathbf{v} = \mathbf{0}$, and it is *negative semi-definite* if, for all $\mathbf{v} \in \mathbb{R}^n$, $B(\mathbf{v}, \mathbf{v}) \leq 0$. Clearly B is negative definite $\iff -B$ is positive definite, and similarly for negative semi-definite. Finally, B is *indefinite* if $B(\mathbf{v}, \mathbf{v})$ is positive for some $\mathbf{v}$ and negative for other $\mathbf{v}$.

The bilinear form B is *non-degenerate* if, for all $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v} \neq \mathbf{0}$, there exists a $\mathbf{w} \in \mathbb{R}^n$ such that $B(\mathbf{v}, \mathbf{w}) \neq 0$, and similarly in the second variable. For example, a positive definite symmetric bilinear form is non-degenerate. A non-degenerate bilinear form on an abstract vector space is defined similarly.

Note that we can define bilinear forms, symmetric bilinear forms, positive definite forms, etc. on an abstract vector space in exactly the same way.

We will call any positive definite bilinear form an *inner product*. The inner product $B(\mathbf{v}, \mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle$ will be called the *standard inner product* or *usual inner product*.

Let B be a bilinear form. Clearly, if we know the values $B(\mathbf{e}_i, \mathbf{e}_j) = a_{ij}$ for the standard basis vectors $\mathbf{e}_i$, then we know all values $B(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = \sum_i x_i \mathbf{e}_i$ and $\mathbf{y} = \sum_i y_i \mathbf{e}_i$. In fact,

$$B(\mathbf{x}, \mathbf{y}) = \sum_{i,j} a_{ij} x_i y_j.$$

Thus B determines and is determined by the $n \times n$ matrix $A = (a_{ij})$. In fact, we can write the above formula as $B(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, A \cdot \mathbf{y} \rangle$, which is also sometimes written as

$$B(\mathbf{x}, \mathbf{y}) = {}^t\mathbf{x} \cdot A \cdot \mathbf{y},$$

where ${}^t\mathbf{x}$ means that $\mathbf{x}$ is written as a row vector (and $\mathbf{y}$ is viewed as a column vector). For example, for the standard inner product $B(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$, the matrix A is the identity matrix Id. Note that B is symmetric $\iff a_{ij} = a_{ji}$ for all i, j. In this case, we say that the matrix $A = (a_{ij})$ is a *symmetric* matrix. We sometimes call the matrix $(B(\mathbf{e}_i, \mathbf{e}_j))$ the *intersection matrix* of B. More generally, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is any basis of $\mathbb{R}^n$, we call the matrix $(B(\mathbf{v}_i, \mathbf{v}_j))$ the *intersection matrix of B with respect to the basis* $\{\mathbf{v}_i\}$. Of course, we can make a similar definition for a bilinear form B on any finite dimensional vector space V once we have chosen a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for V. We will discuss shortly how the intersection matrix changes when we change the basis.

Note that, if A is the matrix associated to B, then B is symmetric $\iff$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle \mathbf{x}, A \cdot \mathbf{y} \rangle = \langle \mathbf{y}, A \cdot \mathbf{x} \rangle = \langle A \cdot \mathbf{x}, \mathbf{y} \rangle.$$

On the other hand, it is left as an exercise that, for every $n \times n$ matrix A and all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle \mathbf{x}, A \cdot \mathbf{y} \rangle = \langle {}^tA \cdot \mathbf{x}, \mathbf{y} \rangle,$$

where ${}^tA$ is the *transpose* of A; its $(i, j)^{\text{th}}$ entry is a_{ji}. Comparing the two formulas, we see that B is symmetric $\iff$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle A \cdot \mathbf{x}, \mathbf{y} \rangle = \langle {}^tA \cdot \mathbf{x}, \mathbf{y} \rangle.$$

or equivalently using bilinearity $\iff$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle A \cdot \mathbf{x} - {}^tA \cdot \mathbf{x}, \mathbf{y} \rangle = \mathbf{0}.$$

But $A \cdot \mathbf{x} - {}^tA \cdot \mathbf{x}$ is perpendicular to every vector $\mathbf{y} \iff A \cdot \mathbf{x} - {}^tA \cdot \mathbf{x}$ is always zero, or equivalently $\iff A \cdot \mathbf{x} = {}^tA \cdot \mathbf{x}$ for every vector $\mathbf{x}$. Thus

we recover (by a slightly longer route) the statement that B is symmetric $\iff$ A is symmetric.

The above ideas also give us a way to describe how the matrix A changes when we use a different basis of $\mathbb{R}^n$ to compute B. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis of $\mathbb{R}^n$ and define, for $\mathbf{s} = (s_1, \ldots, s_n), \mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$

$$B'(\mathbf{s}, \mathbf{t}) = B(\sum_{i=1}^{n} s_i \mathbf{v}_i, \sum_{i=1}^{n} t_i \mathbf{v}_i).$$

Then B' is a bilinear form in $\mathbf{s}$ and $\mathbf{t}$ and so can be described via a matrix. To see what this matrix is, let C be the change of basis matrix defined by $C\mathbf{e}_i = \mathbf{v}_i$. Then $\sum_{i=1}^{n} s_i \mathbf{v}_i = C\mathbf{s}$ and similarly $\sum_{i=1}^{n} t_i \mathbf{v}_i = C\mathbf{t}$. Thus

$$B'(\mathbf{s}, \mathbf{t}) = B(C\mathbf{s}, C\mathbf{t}) = \langle C\mathbf{s}, AC\mathbf{t} \rangle = \langle \mathbf{s}, {}^tCAC\mathbf{t} \rangle.$$

Thus we see that:

Proposition 5.20. *If $B(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, A \cdot \mathbf{y} \rangle$ is a bilinear form, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathbb{R}^n$ and $B'(\mathbf{s}, \mathbf{t})$ is the bilinear form defined by*

$$B'(\mathbf{s}, \mathbf{t}) = B(\sum_{i=1}^{n} s_i \mathbf{v}_i, \sum_{i=1}^{n} t_i \mathbf{v}_i),$$

then $B'(\mathbf{s}, \mathbf{t}) = \langle \mathbf{s}, {}^tCAC\mathbf{t} \rangle$, i.e. B' corresponds to the matrix ${}^tCAC$. □

Given a bilinear form B, we define the *associated quadratic form*

$$Q(\mathbf{x}) = B(\mathbf{x}, \mathbf{x}).$$

If B corresponds to the matrix $A = (a_{ij})$, then $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$; explicitly

$$Q(\mathbf{x}) = \sum_{i,j} a_{ij} x_i x_j = \sum_i a_{ii} x_i^2 + \sum_{i<j} (a_{ij} + a_{ji}) x_i x_j.$$

Clearly, $Q(\mathbf{x})$ is a sum of monomials x_i^2 and $x_i x_j$, $i < j$, of degree two, hence the name quadratic form. Note that, if B is symmetric, then we can rewrite this as

$$Q(\mathbf{x}) = \sum_i a_{ii} x_i^2 + 2 \sum_{i<j} a_{ij} x_i x_j.$$

Conversely, if $P(\mathbf{x})$ is any degree two polynomial in the variables $x_1, \ldots, x_n$ with no constant or linear term, then by definition

$$P(\mathbf{x}) = \sum_i c_i x_i^2 + \sum_{i<j} d_{ij} x_i x_j,$$

and so uniquely determines the symmetric matrix $A = (a_{ij})$, where $a_{ii} = c_i$ and, for $i \neq j$, $a_{ij} = a_{ji} = \frac{1}{2}d_{ij}$. **Note the factor of $\frac{1}{2}$ in the off-diagonal terms!** Somewhat more intrinsically, a quadratic function is said to determine a bilinear form by *polarization*: if $Q(\mathbf{x}) = B(\mathbf{x}, \mathbf{x})$, where B is a *symmetric* bilinear form then it is easy to check that

$$B(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\Big(B(\mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}) - B(\mathbf{x}, \mathbf{x}) - B(\mathbf{y}, \mathbf{y})\Big) = \frac{1}{2}\Big(Q(\mathbf{x}+\mathbf{y}) - Q(\mathbf{x}) - Q(\mathbf{y})\Big).$$

From now on we shall just consider symmetric bilinear forms. We shall call the associated quadratic form Q positive definite, negative definite, indefinite, non-degenerate, etc. if the bilinear form B has the corresponding property. Likewise, we shall call a symmetric matrix A positive definite, negative definite, indefinite, non-degenerate, etc. if the bilinear form $B(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, A\mathbf{y} \rangle$ has the corresponding property.

Example 5.21. Here are some of the standard examples of symmetric bilinear forms and their associated quadratic forms:

1. The usual inner product

   $$B(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i$$

   with associated quadratic form

   $$Q(\mathbf{x}) = \|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2.$$

2. The inner product

   $$B(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{k} x_i y_i - \sum_{i=k+1}^{n} x_i y_i$$

   with associated quadratic form

   $$Q(\mathbf{x}) = x_1^2 + \cdots + x_k^2 - x_{k+1}^2 - \cdots - x_n^2.$$

   In case $k = 1$, this quadratic form is called *Lorentzian* or *hyperbolic*, and in case $k = 1$, $n = 4$, the $\mathbb{R}^4$ with this inner product called *Minkowski space*; it is important in special relativity. Note that, if $0 < k < n$, so that there exist vectors $\mathbf{x}$ such that $Q(\mathbf{x}) > 0$ as well as vectors $\mathbf{x}$ such that $Q(\mathbf{x}) < 0$, and $i \leq k$ and $j \geq k + 1$, then $Q(\mathbf{e}_i + \mathbf{e}_j) = 0$.

3. The *hyperbolic inner product* on $\mathbb{R}^2$:

$$B((x_1, x_2), (y_1, y_2)) = x_1y_2 + x_2y_1; \qquad Q(x_1, x_2) = 2x_1x_2.$$

Note that $Q(\mathbf{e}_1) = Q(\mathbf{e}_2) = 0$, and in fact the only vectors $\mathbf{x}$ such that $Q(\mathbf{x}) = 0$ are the vectors $t\mathbf{e}_1$ or $t\mathbf{e}_2$ for some $t \in \mathbb{R}$.

4. Let V be a vector subspace of $\mathbb{R}^n$ of dimension k. Then the standard inner product restricts to a positive definite symmetric bilinear form on V. Choosing an arbitrary basis of V gives an isomorphism from V to $\mathbb{R}^k$ and hence a positive definite symmetric bilinear form on $\mathbb{R}^k$. Of course, in this example we could have chosen an orthonormal basis of V, in which case the corresponding bilinear form on $\mathbb{R}^k$ would have been the usual one.

Note that, the inner products $Q(\mathbf{x}) = x_1^2 + \cdots + x_k^2 - x_{k+1}^2 - \cdots - x_n^2$, $1 \leq k < n$, and $Q(x_1, x_2) = 2x_1x_2$ are non-degenerate, despite the existence of vectors $\mathbf{x}$ such that $Q(\mathbf{x}) = 0$. Clearly both of these forms are indefinite.

Definition 5.22. If B is a symmetric bilinear form on $\mathbb{R}^n$, then the *null space* of B is the set

$$N = \{\mathbf{v} \in \mathbb{R}^n : B(\mathbf{v}, \mathbf{w}) = 0 \text{ for all } \mathbf{w} \in \mathbb{R}^n\}.$$

Note that B is non-degenerate $\iff$ the null space of B is $\{\mathbf{0}\}$.

Lemma 5.23. *If B is a symmetric bilinear form on $\mathbb{R}^n$ corresponding to the matrix A and N is null space N of B, then $N = \operatorname{Ker} A$. Hence N is a vector subspace of $\mathbb{R}^n$. If W is any subspace of $\mathbb{R}^n$ such that $N \oplus W = \mathbb{R}^n$, then the restriction of B to W is non-degenerate.*

Proof. To see the first statement, note that $\mathbf{v} \in N \iff B(\mathbf{v}, \mathbf{w}) = 0$ for all $\mathbf{w} \in \mathbb{R}^n \iff B(\mathbf{w}, \mathbf{v}) = 0$ for all $\mathbf{w} \in \mathbb{R}^n \iff \langle \mathbf{w}, A\mathbf{v} \rangle = 0$ for all $\mathbf{w} \in \mathbb{R}^n \iff A\mathbf{v} = \mathbf{0} \iff \mathbf{v} \in \operatorname{Ker} A$. Thus $N = \operatorname{Ker} A$ and hence it is a vector subspace of $\mathbb{R}^n$. Of course, that N is a vector subspace can also be seen directly by a straightforward argument. To see the second, suppose that $N \oplus W = \mathbb{R}^n$ and let $\mathbf{w} \in W$. If $\mathbf{w} \neq \mathbf{0}$, then by definition there exists a $\mathbf{u} \in \mathbb{R}^n$ such that $B(\mathbf{w}, \mathbf{u}) \neq 0$. Then we can write $\mathbf{u} = \mathbf{v}_1 + \mathbf{w}_1$ with $\mathbf{v}_1 \in N$ and $\mathbf{w}_1 \in W$. By definition, $B(\mathbf{v}_1, \mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^n$. Hence $B(\mathbf{v}_1, \mathbf{w}) = 0$ and so $B(\mathbf{w}, \mathbf{u}) = B(\mathbf{w}, \mathbf{w}_1) \neq 0$. It follows that the restriction of B to W is non-degenerate. □

We can now state the main theorem concerning symmetric bilinear forms. To give some idea of what it means, first consider the case where B is positive definite. Then a minor modification of Gram-Schmidt procedure can be used to find an orthonormal basis for B, in other words a basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathbb{R}^n$ such that $B(\mathbf{u}_i, \mathbf{u}_j) = 0$ if $i \neq j$ and $B(\mathbf{u}_i, \mathbf{u}_i) = 1$. Equivalently, $Q(\sum_{i=1}^n x_i \mathbf{u}_i) = \sum_{i=1}^m x_i^2$, or the matrix for B in the basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ (i.e. the matrix whose $(i,j)^{\text{th}}$ entry is $B(\mathbf{u}_i, \mathbf{u}_j)$) is the identity matrix. Now for a general symmetric bilinear form B, we can't expect to find such a basis, because B may assume some negative values or be degenerate. So we show that there is a basis with the properties that can reasonably be expected under these circumstances: $B(\mathbf{u}_i, \mathbf{u}_j) = 0$ if $i \neq j$ and $B(\mathbf{u}_i, \mathbf{u}_i) = \pm 1$ or 0. More generally, a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ such that $B(\mathbf{v}_i, \mathbf{v}_j) = 0$ if $i \neq j$ is called a *diagonal basis* for the form B. The theorem will state that, for every symmetric bilinear form, there exists a diagonal basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$, such that the the diagonal entries of $(B(\mathbf{v}_i, \mathbf{v}_j)$ are $+1$, -1, or 0. For example, for the hyperbolic form B (or Q) on $\mathbb{R}^2$ given by $Q(x_1, x_2) = 2x_1x_2$, the basis

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}}(\mathbf{e}_1 + \mathbf{e}_2); \qquad \mathbf{v}_1 = \frac{1}{\sqrt{2}}(\mathbf{e}_1 - \mathbf{e}_2)$$

is a diagonal basis: $Q(\mathbf{v}_1) = 1$, $Q(\mathbf{v}_2) = -1$, $B(\mathbf{v}_1, \mathbf{v}_2) = 0$. Note that, in this case, if C is the change of basis matrix specified by $C\mathbf{e}_i = \mathbf{v}_i$, $i = 1, 2$, then $C = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, and a straightforward computation shows that

$${}^tC \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot C = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

as is predicted by our general formula for how the intersection matrix changes under a change of basis.

Theorem 5.24 (Sylvester)**.** *Let B be a symmetric bilinear form on $\mathbb{R}^n$, or more generally on a finite-dimensional vector space V, with associated quadratic form Q. Then there exists a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{R}^n$, or of V, and nonnegative integers r and s with $r+s \leq n$, such that, for all $x_1, \ldots, x_n \in \mathbb{R}$,*

$$Q(\sum_{i=1}^n x_i \mathbf{v}_i) = \sum_{i=1}^r x_i^2 - \sum_{i=r+1}^{r+s} x_i^2.$$

Of course, there is an equivalent statement in terms of B:

$$B(\sum_{i=1}^n x_i \mathbf{v}_i, \sum_{i=1}^n y_i \mathbf{v}_i) = \sum_{i=1}^r x_i y_i - \sum_{i=r+1}^{r+s} x_i y_i.$$

Here the integers r, s, and $t = n - r - s$ only depend on B and not on the choice of basis above.

Proof. Before we begin the proof, suppose that $\mathbf{v} \in \mathbb{R}^n$ (or V). We define $\{\mathbf{v}\}^{\perp_B}$, the orthogonal space with respect to the inner product B, to be the set

$$\{\mathbf{x} \in \mathbb{R}^n : B(\mathbf{x}, \mathbf{v}) = 0.$$

The orthogonal space of an arbitrary subset X of of $\mathbb{R}^n$ is defined similarly, and is denoted $X^{\perp_B}$. By the same arguments that we gave for the usual inner product, $\{\mathbf{v}\}^{\perp_B}$ and $X^{\perp_B}$ are subspaces of $\mathbb{R}^n$. However, if $B(\mathbf{v}, \mathbf{v}) = Q(\mathbf{v}) = 0$, then $\mathbf{v} \in \{\mathbf{v}\}^{\perp_B}$, which of course can happen even if B is non-degenerate. On the other hand, $Q(\mathbf{v}) \neq 0 \iff \mathbf{v} \notin \{\mathbf{v}\}^{\perp_B}$. Note that, if N is the null space of B, then $\mathbf{v} \in N \iff \{\mathbf{v}\}^{\perp_B} = \mathbb{R}^n$.

With this said, let us prove the existence part of Sylvester's theorem by induction on n. If $n = 1$, then $Q(x) = ax^2$, where $a = Q(1)$. If $a = 0$, then $N = \mathbb{R}$ and we are in the degenerate case of Sylvester's theorem where $r = s = 0$ and $t = 1$. So we may assume that $a \neq 0$. Choosing the new basis $\{1/\sqrt{|a|}\}$ of $\mathbb{R}$. Then $Q(x \cdot 1/\sqrt{|a|}) = (a/|a|) \cdot x^2 = \pm x$, where the sign depends on whether $a > 0$ or $a < 0$, i.e. on whether $Q(x) > 0$ for all nonzero x or $Q(x) < 0$ for all nonzero x. In the first case $r = 1$, $s = t = 0$, and in the second case $r = t = 0$, $s = 1$.

For the inductive step, suppose that the theorem has been established for all symmetric bilinear forms on $\mathbb{R}^{n-1}$.

Claim 5.25. *Let B be a symmetric bilinear form on $\mathbb{R}^n$, with associated quadratic form Q. If B is not identically zero, then there exists a $\mathbf{v} \in \mathbb{R}^n$ such that $Q(\mathbf{v}) \neq 0$.*

Proof. Since B is not identically zero, there exist $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ such that $B(\mathbf{v}, \mathbf{w}) \neq 0$. If $Q(\mathbf{v}) \neq 0$ we are done. If $Q(\mathbf{w}) \neq 0$ we can use $\mathbf{w}$ instead of $\mathbf{v}$. Otherwise, by bilinearity,

$$Q(\mathbf{v} + \mathbf{w}) = Q(\mathbf{v}) + 2B(\mathbf{v}, \mathbf{w}) + Q(\mathbf{w}) = 2B(\mathbf{v}, \mathbf{w}) \neq 0,$$

and we can simply use $\mathbf{v} + \mathbf{w}$ instead of $\mathbf{v}$. □

Returning to the proof of Sylvester's theorem, if B is identically zero then we are done: we can let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be any basis, and then the matrix for B in this basis is the zero matrix, which is of the desired form with $r = s = 0, t = n$. Otherwise, let $\mathbf{v}$ be some vector in $\mathbb{R}^n$ such that $Q(\mathbf{v}) \neq 0$. If $Q(\mathbf{v}) = a \neq 0$, then after replacing $\mathbf{v}$ by the scalar multiple $(1/\sqrt{|a|})\mathbf{v}$,

we have either $Q(\mathbf{v}) = 1$ or $Q(\mathbf{v}) = -1$. Now consider the vector subspace $\{\mathbf{v}\}^{\perp_B}$ of $\mathbb{R}^n$.

Claim 5.26. $\dim\{\mathbf{v}\}^{\perp_B} = n-1$, *and* $\operatorname{span}\{\mathbf{v}\} \oplus \{\mathbf{v}\}^{\perp_B} = \mathbb{R}^n$.

Proof. The linear function $L(\mathbf{x}) = B(\mathbf{x}, \mathbf{v})\colon \mathbb{R}^n \to \mathbb{R}$ is surjective, since B is non-degenerate, and its kernel by definition is $\{\mathbf{v}\}^{\perp_B}$. Hence $\dim\{\mathbf{v}\}^{\perp_B} = n-1$ (this argument only uses the fact that B is non-degenerate and that $\mathbf{v} \neq \mathbf{0}$). Since $Q(\mathbf{v}) \neq 0$, $\mathbf{v} \notin \{\mathbf{v}\}^{\perp_B}$ and hence $\operatorname{span}\{\mathbf{v}\} \cap \{\mathbf{v}\}^{\perp_B} = \{\mathbf{0}\}$. It follows that $\operatorname{span}\{\mathbf{v}\} + \{\mathbf{v}\}^{\perp_B} = \operatorname{span}\{\mathbf{v}\} \oplus \{\mathbf{v}\}^{\perp_B}$ is a vector subspace of $\mathbb{R}^n$ of dimension $1 + (n-1) = n$, and hence $\operatorname{span}\{\mathbf{v}\} \oplus \{\mathbf{v}\}^{\perp_B} = \mathbb{R}^n$. □

Let us now prove the existence part of Sylvester's theorem. By induction applied to the vector space $\{\mathbf{v}\}^{\perp_B}$, which can be viewed as $\mathbb{R}^{n-1}$ after choosing a basis, and to the bilinear form on $\{\mathbf{v}\}^{\perp_B}$ which is the restriction of the bilinear form B, there exists a basis $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ for $\{\mathbf{v}\}^{\perp_B}$ such that the restriction of B to $\{\mathbf{v}\}^{\perp_B}$ satisfies: $Q(\mathbf{v}_i) = \pm 1$ or 0 and $B(\mathbf{v}_i, \mathbf{v}_j) = 0$ if $i \neq j$. Setting $\mathbf{v} = \mathbf{v}_n$, we have by construction that $Q(\mathbf{v}_n) = \pm 1$ and $B(\mathbf{v}_i, \mathbf{v}_n) = 0$ if $i < n$. Thus, after renumbering to put all of the $\mathbf{v}_i$ with $Q(\mathbf{v}_i) = +1$ first, we have found a basis as in the statement of the theorem. This completes the inductive step.

We still must show that the integers r, s, t are uniquely determined by the bilinear form B. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is as in the statement of the theorem, then $B(\sum_{i=1}^n x_i \mathbf{v}_i, \mathbf{v_j}) = x_i B(\mathbf{v}_j, \mathbf{v_j})$ and hence $B(\mathbf{w}, \mathbf{v_j}) = 0$ for all $j \iff \mathbf{w}$ is in the span of $\mathbf{v}_{r+s+1}, \ldots, \mathbf{v}_n$, i.e. of the last $t = n - r - s$ vectors, in the notation of the statement of the theorem. Thus $\mathbf{v}_{r+s+1}, \ldots, \mathbf{v}_n$ is a basis for the null space, and so $t = n - r - s = \dim N$, where N is the null space of B. It follows that t does not depend on a particular choice of basis.

Now suppose that we have two bases $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{v}'_1, \ldots, \mathbf{v}'_n$ of $\mathbb{R}^n$, with corresponding nonnegative integers r, s and r', s' with $r + s + t = r' + s' + t = n$. Let $V_1 = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ be the span of the first r elements of the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$, so that $Q(\sum_{i=1}^r x_i \mathbf{v}_i) = \sum_{i=1}^r x_i^2$. Thus B is positive definite on V_1: if $\mathbf{v} = \sum_{i=1}^r x_i \mathbf{v}_i \in V_1$, then $Q(\mathbf{v}) \geq 0$ and $Q(\mathbf{v}) = 0 \iff \mathbf{v} = \mathbf{0}$. Let $V_2 = \operatorname{span}\{\mathbf{v}'_{r'}, \ldots, \mathbf{v}'_n\}$ be the span of the *last* $n - r'$ elements of the basis $\mathbf{v}'_1, \ldots, \mathbf{v}'_n$, so that $Q(\sum_{i=r'}^n x_i \mathbf{v}'_i) = -\sum_{i=r'}^{r'+s'} x_i^2$. Thus B is *negative semidefinite* on V_2: if $\mathbf{w} = \sum_{i=r'}^n x_i \mathbf{v}'_i \in V_2$, then $Q(\mathbf{w}) \leq 0$. It follows that $V_1 \cap V_2 = \{\mathbf{0}\}$, since if $\mathbf{v} \in V_1 \cap V_2$, then $Q(\mathbf{v}) \geq 0$ and $Q(\mathbf{v}) \leq 0$, so that $Q(\mathbf{v}) = 0$, and hence, using the fact that B is positive definite on V_1, $\mathbf{v} = \mathbf{0}$. Thus $V_1 + V_2 = V_1 \oplus V_2$, so that

$$\dim V_1 + \dim V_2 = r + n - r' = r + s' + t \leq n = r + s + t.$$

It follows that $s' \leq s$. But we can reverse the roles of V_1 and V_2, which gives $s \leq s'$. Hence $s = s'$ and therefore $r = r'$. This shows that the integers r, s, t are uniquely determined by B. □

To get a feel for what Sylvester's theorem says, let us consider the case $n = 2$. If $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$ is a 2×2 symmetric matrix, then A corresponds to the quadratic form $Q(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$. (Note the factor of 2!) Assume that $a_{11} \neq 0$. (We will leave the analysis of the case $a_{11} = 0$ as an exercise.) Then we can analyze Q by the time-honored method of completing the square:

$$\begin{aligned} Q(x_1, x_2) &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = a_{11}\left(x_1^2 + \frac{2a_{12}}{a_{11}}x_1x_2\right) + a_{22}x_2^2 \\ &= a_{11}\left(x_1^2 + \frac{2a_{12}}{a_{11}}x_1x_2 + \frac{a_{12}^2}{a_{11}^2}x_2^2\right) + \left(a_{22} - \frac{a_{12}^2}{a_{11}}\right)x_2^2 \\ &= a_{11}\left(x_1 + \frac{a_{12}}{a_{11}}x_2\right)^2 + \left(\frac{a_{11}a_{22} - a_{12}^2}{a_{11}}\right)x_2^2. \end{aligned}$$

In particular $Q(x_1, x_2)$ is a sum of squares. If we set $t_1 = x_1 + \frac{a_{12}}{a_{11}}x_2$ and $t_2 = x_2$, then $Q(x_1, x_2) = c_1t_1^2 + c_2t_2^2$. Note that, instead of solving for the t_i in terms of the x_i, we can easily solve for the x_i in terms of the t_i:

$$\begin{aligned} x_1 &= t_1 - \frac{a_{12}}{a_{11}}t_2; \\ x_2 &= t_2, \end{aligned}$$

then $(x_1, x_2) = t_1\mathbf{v}_1 + t_2\mathbf{v}_2$. where $\mathbf{v}_1 = (1, 0) = \mathbf{e}_1$ and $\mathbf{v}_2 = (-a_{12}/a_{11}, 0)$, and in this basis $Q(t_1\mathbf{v}_1 + t_2\mathbf{v}_2) = c_1t_1^2 + c_2t_2^2$. Thus the matrix for Q in the basis $\mathbf{v}_1, \mathbf{v}_2$ is diagonal, and we could then adjust the $\mathbf{v}_i$ by a scalar to make the diagonal entries ± 1 or 0. Clearly Q is positive definite $\iff$ c_1 and c_2 are both positive. Now $c_1 > 0 \iff a_{11} > 0$, and if $a_{11} > 0$, then $c_2 > 0 \iff a_{11}a_{22} - a_{12}^2 > 0$. Now we recognize the expression $a_{11}a_{22} - a_{12}^2$ as $\det A$, the determinant of A. (We shall discuss determinants in general in the next chapter.) So we see:

Proposition 5.27. *If $a_{11} \neq 0$, then the form Q is positive definite $\iff$ $a_{11} > 0$ and $\det A > 0$.* □

A similar argument shows that, again assuming that $a_{11} \neq 0$, then the form Q is *negative* definite $\iff$ $a_{11} < 0$ and $\det A > 0$. Note that the direction of the inequality changes.

We shall generalize the criterion for a form to be positive or negative definite to n dimensions in the next chapter.

Chapter 6

Determinants

6.1 Multilinear forms

Before we begin the discussion of determinants, we digress to discuss multilinear functions from $(\mathbb{R}^n)^d$ to $\mathbb{R}$. We have already discussed bilinear functions in the last chapter.

Definition 6.1. A function $M\colon (\mathbb{R}^n)^d \to \mathbb{R}$ is *d-multilinear* (or simply multilinear if the d is clear from the context) if, for all i, $1 \leq i \leq d$, and for all $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^n$,

$$\begin{gathered} M(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}_1 + \mathbf{w}_2, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_d) \\ = M(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}_1, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_d) + M(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}_2, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_d); \\ M(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, t\mathbf{w}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_d) = tM(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_d) \end{gathered}$$

for all $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w} \in \mathbb{R}^n$ and $t \in \mathbb{R}$. In other words, M is linear in each variable when the other variables are held fixed.

The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are equivalent to a $d \times n$ matrix with entries x_{ij}. In other words, if we write out each $\mathbf{v}_i$ in terms of the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, then $\mathbf{v}_i = \sum_{j=1}^n x_{ji}\mathbf{e}_j$. A very tedious but essentially straightforward computation shows that, in this case, $M(\mathbf{v}_1, \ldots, \mathbf{v}_d)$ can be expanded out in terms of products of the x_{ij} with certain coefficients which only depend on M. To see how this works, we evaluate M on vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$, with $\mathbf{v}_i = \sum_{j=1}^n x_{ji}\mathbf{e}_j$ for every i. So we are confronted with trying to evaluate the expression

$$M(\sum_{j=1}^n x_{j1}\mathbf{e}_j, \ldots, \sum_{j=1}^n x_{jn}\mathbf{e}_j).$$

To understand how to do this, consider first the problem of trying to evaluate a product of n terms, each of which is a sum of n numbers:

$$(x_{11} + x_{21} + \cdots + x_{n1}) \cdot (x_{12} + x_{22} + \cdots + x_{n2}) \cdots (x_{1d} + x_{2d} + \cdots + x_{nd}).$$

The product (by repeated applications of the distributive rule) is the sum of all possible products $x_{i_1,1} \cdot x_{i_2,2} \cdots x_{i_d,d}$ where we choose one index i_1 for the first term, another i_2 for the second, and so on. Notice that in fact there are n^d terms in the sum. A very similar procedure shows that $M(\mathbf{v}_1, \ldots, \mathbf{v}_d)$ is a sum of all possible terms of the form $M(x_{i_1,1}\mathbf{e}_{i_1}, \ldots, x_{i_d,d}\mathbf{e}_{i_d})$. Also, using multilinearity, we have

$$M(x_{i_1,1}\mathbf{e}_{i_1}, \ldots, x_{i_d,d}\mathbf{e}_{i_d}) = x_{i_1,1} \cdots x_{i_d,d} M(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_d}).$$

Thus the multilinear function M is specified by the n^d coefficients

$$a_{i_1,\ldots,i_d} = M(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_d}),$$

where $i_1, \ldots, i_d$ is any sequence whose entries satisfy $1 \leq i_k \leq n$. Note that, if $d = 1$, then we can think of the n coefficients as a *vector*, and if $d = 2$, so that we are in the bilinear case, we can think of the n^2 coefficients a_{i_1,i_2} as a *matrix*. For general d, the data of the n^d coefficients of the multilinear form M is somewhat loosely referred to as a *tensor*.

Finally we give the analogue of the symmetry condition in the bilinear case $d = 2$. We will not however use any of this discussion later. We define a *permutation* of the set $\{1, \ldots, n\}$ to be a one-to-one correspondence $f\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. The set of all permutations of $\{1, \ldots, n\}$ will be denoted S_n. We call M *symmetric* if, for all $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^n$ and for every $f \in S_d$,

$$M(\mathbf{v}_1, \ldots, \mathbf{v}_d) = M(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(d)}).$$

In other words, no matter how we permute the vector variables $\mathbf{v}_1, \ldots, \mathbf{v}_d$, the value of M is unchanged. In case $d = 1$, there is no condition, since there is no nontrivial way to permute a set with one element ($S_1 = \{\mathrm{Id}\}$). For $d = 2$, we recover the symmetry condition for bilinear forms B: B is symmetric $\iff$ for all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$, $B(\mathbf{v}_1, \mathbf{v}_2) = B(\mathbf{v}_2, \mathbf{v}_1)$. For $d = 3$, a 3-multilinear form $M(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is called a *trilinear* form; it is symmetric $\iff$ the six different expressions

$$M(\mathbf{x}, \mathbf{y}, \mathbf{z}), M(\mathbf{y}, \mathbf{z}, \mathbf{x}), M(\mathbf{z}, \mathbf{x}, \mathbf{y}), M(\mathbf{y}, \mathbf{x}, \mathbf{z}), M(\mathbf{x}, \mathbf{z}, \mathbf{y}), M(\mathbf{z}, \mathbf{y}, \mathbf{x})$$

obtained by permuting the variables are all equal.

For a d-multilinear form M, we can form the analogue of the quadratic form associated to a bilinear form by considering $M(\mathbf{x}, \dots, \mathbf{x})$, i.e. all of the $\mathbf{v}_i$ are equa to the same vector $\mathbf{x}$. By the discussion of the explicit form of a multilinear function above, if $\mathbf{x} = (x_1, \dots, x_n)$, then $M(\mathbf{x}, \dots, \mathbf{x})$ is a sum of monomials $x_{i_1} \cdots x_{i_d}$, where some of the indices can be repeated. Such an expression is a polynomial in the variables $x_1, \dots, x_n$. Each monomial $x_{i_1} \cdots x_{i_d}$ has total degree d in the variables $x_{i_1} \cdots x_{i_d}$, and we refer to such a polynomial as a *homogeneous polynomial of degree d*. As in the case of bilinear forms, there is a bijection, in fact an isomorphism, from the vector space of symmetric d-multilinear forms on $\mathbb{R}^n$ to the vector space of homogeneous polynomials in n variables of degree d. In one direction, given a symmetric d-multilinear form $M(\mathbf{v}_1, \dots, \mathbf{v}_d)$, we define the associated homogeneous polynomial $M(\mathbf{x}, \dots, \mathbf{x})$. Of course, we can make this definition even if M is not symmetric. The inverse function is given by a polarization identity which is similar to the one we gave for $d = 2$, but more complicated. The case $d = 3$ is an exercise.

In this chapter, we shall be concerned not with symmetric multilinear functions but with n-multilinear functions M on $\mathbb{R}^n$ which have a related property, called *antisymmetric* or *alternating*. We explain the meaning of this property in the next section.

6.2 Definition and first properties of determinants

Determinants are an important computational and theoretical tool for understanding the following problem: when does an $n \times n$ matrix have an inverse? Equivalently, when are n vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$ linearly independent? The determinant is a function $\det\colon \mathbb{M}_n(\mathbb{R}) \to \mathbb{R}$ with the property that $\det A \neq 0$ if and only if A has an inverse. We can also think of det as a function on n (column) vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$. (While we will mainly stick to the case of $\mathbb{R}$, determinants can be defined for $n \times n$ matrices with coefficients in any field F, and the value of the determinant will then be an element of F. We will use the case $F = \mathbb{C}$ later.)

To define det, we proceed inductively. If $A = (a)$ is a 1×1 matrix, $\det A = a$. If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a 2×2 matrix, then $\det A = ad - bc$. In general, suppose that we have defined the determinant of an $(n-1) \times (n-1)$

matrix. Let

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

be an $n \times n$ matrix. Define

$$\det A = a_{11} \det A_{11} - a_{12} \det A_{12} + \cdots + (-1)^{n+1} a_{1n} \det A_{1n},$$

where A_{1k} is the $(n-1) \times (n-1)$ matrix formed by deleting the first row and k^{th} column of A (and A_{jk} is similarly defined). For example,

$$\det \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ -2 & 4 & 3 \end{pmatrix} = 1(3-4) - 2(0+2) + 3(0+2) = 1.$$

It is possible to write a closed formula for $\det A$. For example, in the 3×3 case, with $A = (a_{ij})$, we have

$$\begin{aligned} \det A = \det & \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} \\ &- a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{23}a_{32}a_{11}. \end{aligned}$$

In general however $\det A$ will involve $n!$ terms (not $2n$). For example, a 4×4 determinant has 24 terms.

Example 6.2. If A is a diagonal matrix ($a_{ij} \neq 0$ if and only if $i = j$, and in this case we let $a_{ii} = d_i$) then $\det A = d_1 \cdots d_n$, as follows from induction. More generally, if A is a lower triangular matrix ($a_{ij} = 0$ if $i < j$), which we write pictorially as

$$A = \begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ * & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ * & * & * & \dots & d_n \end{pmatrix}$$

then $\det A = d_1 \cdots d_n$ is again the product of the diagonal entries.

We will view det as a function on n column vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. Here are some of the basic properties of det:

1. det is a *multilinear* function of n vector variables. In other words, for each i with $1 \leq i \leq n$, and for all $\mathbf{v}_i, \mathbf{w}_i \in \mathbb{R}^n$ and $t \in \mathbb{R}$,

$$\det(\mathbf{v}_1, \ldots, \mathbf{v}_i + \mathbf{w}_i, \ldots, \mathbf{v}_n) = \det(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n) + \det(\mathbf{v}_1, \ldots, \mathbf{w}_i, \ldots, \mathbf{v}_n);$$
$$\det(\mathbf{v}_1, \ldots, t\mathbf{v}_i, \ldots, \mathbf{v}_n) = t \det(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n).$$

2. det satisfies: if $\mathbf{v}_i = \mathbf{v}_{i+1}$ for some i, $1 \leq i \leq n-1$, then

$$\det(\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n) = 0.$$

   We call det an *alternating* function.

3. $\det(\mathrm{Id}) = \det(\mathbf{e}_1, \ldots, \mathbf{e}_n) = 1$.

Proposition 6.3. det *is the unique function defined on n-tuples $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and satisfying (1)—(3) above. More precisely,*

1. det *is an alternating multilinear function from $(\mathbb{R}^n)^n$ to $\mathbb{R}$, such that $\det(\mathbf{e}_1, \ldots, \mathbf{e}_n) = 1$.*

2. *If $D\colon (\mathbb{R}^n)^n \to \mathbb{R}$ is any alternating multilinear function, then $D = c\det$, where $c = D(\mathbf{e}_1, \ldots, \mathbf{e}_n)$.*

Proof. Let us first show that det does satisfy the above properties. We have seen (3) by induction and the definition. For the other properties, we again argue by induction: (1) holds in case $n = 1$, and (2) is easy to check for $n = 2$ and is vacuously true for $n = 1$. To see (1), if say $\mathbf{v}_i = \mathbf{w} + \mathbf{u}$, let $\mathbf{v}_i'$, $\mathbf{w}'$, and $\mathbf{u}'$ be the $(n-1)$-vectors obtained by deleting the first entries of $\mathbf{v}_i$, $\mathbf{w}$, and $\mathbf{u}$, viewed as column vectors, and let x be the first entry of $\mathbf{w}$ and y the first entry of $\mathbf{u}$. Then A_{1k} has a column (either the $(i-1)^{\text{st}}$ or the i^{th}) which is the sum of the vectors $\mathbf{w}'$ and $\mathbf{u}'$, except when $k = i$ in which case we multiply $(-1)^i A_{1i}$ by $(x + y)$. Let B be the matrix where the i^{th} column $\mathbf{v}_i$ of A has been replaced by $\mathbf{w}$, and let C be the matrix where $\mathbf{v}_i$ has been replaced by $\mathbf{u}$. The statement we are trying to show is: $\det A = \det B + \det C$. Note that the $(n-1) \times (n-1)$ matrices A_{1i}, B_{1i}, and C_{1i} are all equal, whereas for $k \neq i$, B_{1k} is obtained from A_{1k} by replacing the column corresponding to $\mathbf{v}'$ by $\mathbf{w}'$ and similarly for C_{1k}. By induction, for $k \neq i$, we have $\det A_{1k} = \det B_{1k} + \det C_{1k}$, and of course $\det A_{1i} = \det B_{1i} = \det C_{1i}$. Using this, it is easy to see that $\det A = \det B + \det C$. The argument for scalar multiplication is similar.

Finally, to see (2), we shall show that $\det(\mathbf{v}_1, \mathbf{v}_1, \ldots) = 0$ (the other cases are similar). For $k \neq 1, 2$, A_{1k} has a repeated column (the first is

equal to the second) and so $\det A_{1k} = 0$ by induction (note the case $n = 2$ is easy, and the case $n = 1$ is in fact vacuously true). So the only remaining terms in the sum are $a_{11} \det A_{11} - a_{12} \det A_{12}$. But by assumption $a_{11} = a_{12}$ (they are the first components of $\mathbf{v}_1$ and $\mathbf{v}_2 = \mathbf{v}_1$ respectively) and likewise $A_{11} = A_{12}$. Thus $a_{11} \det A_{11} - a_{12} \det A_{12} = a_{11} \det A_{11} - a_{11} \det A_{11} = 0$.

Before we show uniqueness, let us make a few remarks on alternating multilinear functions. Suppose that $D(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is an alternating multilinear function of $\mathbf{v}_1, \ldots, \mathbf{v}_n$, in other words $D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_i, \ldots, \mathbf{v}_n) = 0$ whenever there are consecutive repeated values. We claim:

Claim 6.4. *Whenever we switch two consecutive entries, D is replaced by $-D$. In other words*

$$D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n) = -D(\mathbf{v}_1, \ldots, \mathbf{v}_{i+1}, \mathbf{v}_i, \ldots, \mathbf{v}_n).$$

Proof. Consider $0 = D(\mathbf{v}_1, \ldots, \mathbf{v}_i + \mathbf{v}_{i+1}, \mathbf{v}_i + \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n)$. Expanding out by multilinearity and using the alternating property we obtain:

$$\begin{aligned} 0 &= D(\mathbf{v}_1, \ldots, \mathbf{v}_i + \mathbf{v}_{i+1}, \mathbf{v}_i + \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n) \\ &= D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_i, \ldots, \mathbf{v}_n) + D(\mathbf{v}_1, \ldots, \mathbf{v}_{i+1}, \mathbf{v}_i, \ldots, \mathbf{v}_n) \\ &+ D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n) + D(\mathbf{v}_1, \ldots, \mathbf{v}_{i+1}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n) \\ &= D(\mathbf{v}_1, \ldots, \mathbf{v}_{i+1}, \mathbf{v}_i, \ldots, \mathbf{v}_n) + D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n), \end{aligned}$$

as claimed. □

Corollary 6.5. *With D as above, if we switch any two entries, D is replaced by $-D$. In other words*

$$D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_n) = -D(\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n).$$

In particular, if any two vectors of the $\mathbf{v}_1, \ldots, \mathbf{v}_n$, not necessarily consecutive, are equal then $D(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$.

Proof. Let $\ell = j - i$, assuming for simplicity that $j > i$. To switch $\mathbf{v}_j$ and $\mathbf{v}_i$, first switch $\mathbf{v}_i$ with $\mathbf{v}_{i+1}$, then with $\mathbf{v}_{i+2}$, and so on until we reach $\mathbf{v}_j$. There are ℓ switches needed in all, so the result is

$$D(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_n) = (-1)^\ell D(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_j, \mathbf{v}_i, \ldots, \mathbf{v}_n).$$

Now we need to switch $\mathbf{v}_j$ back to the i^{th} spot, by successively switching it with its neighbors to the left. This requires another $(-1)^{\ell-1}$ sign changes. So D changes by $(-1)^\ell (-1)^{\ell-1} = (-1)^{2\ell-1} = -1$. □

To generalize this property still further, recall that we have defined S_n, the set of all permutations of $\{1, \ldots, n\}$. Clearly S_n has $n!$ elements, since to specify an element f, there are n possibilities for $f(1)$, but then $n-1$ possibilities for $f(2)$ once we have chosen $f(1)$, and so on. The composition of two permutations is again a permutation, composition of permutations is an associative operation, the identity function is a permutation, and the inverse function f^{-1} of a permutation is again a permutation. Thus S_n is an example of a *group* (but composition of functions is not commutative if $n \geq 3$). We will often call the composition of two elements of S_n their *product*. A special example of a permutation is a *transposition* $t_{i,j}$, depending on $i, j \in \{1, \ldots, n\}$, which satisfies: $t_{i,j}(i) = j$, $t_{i,j}(j) = i$, and $t_{i,j}(\ell) = \ell$ if $\ell \neq i, j$. We denote the transposition corresponding to i, j by $t_{i,j}$. Note the identity permutation is equal to $t_{i,j} \circ t_{i,j}$, because if we switch i and j and then switch them back again we get the identity function.

Note that we can identify S_{n-1} with the subset $H \subseteq S_n$ defined by

$$H = \{f \in S_n : f(n) = n\}$$

in following sense: If $f \in S_{n-1}$, then we can define $\tilde{f} \in S_n$ by the rule: $\tilde{f}(i) = f(i)$ if $i < n$. and $\tilde{f}(n) = n$. Conversely if $g \in H \subseteq S_n$, then $g(i) \in \{1, \ldots, n-1\}$ if $i \neq n$ so that the restriction $g|\{1, \ldots, n-1\}$ defines a function from $\{1, \ldots, n-1\}$ to itself which is clearly a bijection, hence an element of S_{n-1}. These two constructions give inverse functions from S_{n-1} to H, and we then identify S_{n-1} with H. Clearly, the composition of two elements of H is the same as the composition when we view tham as elements of S_{n-1}.

Lemma 6.6. *Every permutation $f \in S_n$ can be written as a product (composition) of transpositions, although not in a unique way.*

Proof. The proof is by induction on n. We could take the statement as vacuously true for $n = 1$, or begin the induction at $n = 2$. Here S_2 has two elements Id and $t_{1,2}$, with $\mathrm{Id} = t_{1,2} \circ t_{1,2}$. Suppose that we have proved the statement for $n-1$. For the inductive step, if $f \in S_n$, first suppose that $f(n) = n$. Then by definition f is in the subset $H = \{f \in S_n : f(n) = n\}$. Thus f is a product of transpositions in S_{n-1}, and by the remarks before the statement of the lemma, it is a product of transpositions in S_n. Otherwise, $f \notin H$, so that $f(n) = i < n$. Then $t_{n,i} \circ f(n) = t_{n,i}(i) = n$, so that $g = t_{n,i} \circ f \in H$ and hence by induction is a product of transpositions. But then so is $f = t_{n,i} \circ g$, using $t_{n,i} \circ t_{n,i} \circ f = \mathrm{Id} \circ f = f$. □

For a concrete example, the set S_3 has 6 elements. We have the identity Id and three transpositions $t_{1,2}, t_{1,3}, t_{2,3}$. There is also the element r defined by $r(1) = 2, r(2) = 3, r(3) = 1$, and its inverse $r^{-1} = r^2$, which satisfies $r^2(1) = 3, r^2(2) = 1, r^2(3) = 2$. Since we have found 6 different permutations, we have found all of S_6. It is easy to check that $r = t_{1,3} \circ t_{1,2}$ and that $r^2 = t_{1,2} \circ t_{1,3}$.

Next we claim that every permutation $f \in S_n$ defines a linear map P_f from $\mathbb{R}^n$ to itself. To define the linear map P_f, it is enough to define it on the standard basis, and we let $P_f(\mathbf{e}_i) = \mathbf{e}_{f(i)}$. The matrix corresponding to P_f has columns $\mathbf{e}_{f(1)}, \ldots, \mathbf{e}_{f(n)}$ and is called a *permutation matrix*. Note that $P_{\mathrm{Id}} = \mathrm{Id}$. If $n = 2$, the permutation matrix $P_{t_{1,2}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. In case $n = 3$, for example,

$$P_{t_{1,2}} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \qquad P_r = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Lemma 6.7. *In the above notation,*

$$P_{g \circ f} = P_g \cdot P_f,$$

where $P_g \cdot P_f$ denotes matrix product, or equivalently composition of linear maps.

Proof. It is enough to check the equality on the standard basis vectors. But for all i,

$$P_{g \circ f}(\mathbf{e}_i) = \mathbf{e}_{g \circ f(i)} = \mathbf{e}_{g(f(i))},$$

whereas $P_g \cdot P_f(\mathbf{e}_i) = P_g \mathbf{e}_{f(i)} = \mathbf{e}_{g(f(i))}$ as well. Hence $P_{g \circ f} = P_g \cdot P_f$. □

Define the *sign* of f, written $\operatorname{sign} f$, to be $\det(\mathbf{e}_{f(1)}, \ldots, \mathbf{e}_{f(n)})$. Thus if $f = t_{i,j}$, then $\operatorname{sign} f = -1$.

Proposition 6.8. *If $f \in S_n$ is a product of k transpositions, then* $\operatorname{sign} f = (-1)^k$.

Proof. We argue by induction on k, where f is written as a product of k transpositions. We have seen above that the statement is true if $k = 1$. If instead $f = t_{i,j} \circ g$, where g is a product of k transpositions and hence f is a product of $k + 1$ transpositions, then we get the sequence of vectors

$\mathbf{e}_{f(1)}, \dots, \mathbf{e}_{f(n)}$ from the sequence $\mathbf{e}_{g(1)}, \dots, \mathbf{e}_{g(n)}$ by switching the i^{th} entry $\mathbf{e}_{g(i)}$ with the j^{th} entry $\mathbf{e}_{g(j)}$. Thus

$$\det P_f = \det(\mathbf{e}_{f(1)}, \dots, \mathbf{e}_{f(n)}) = -\det(\mathbf{e}_{g(1)}, \dots, \mathbf{e}_{g(n)})$$

and hence by induction $\operatorname{sign} f = -\operatorname{sign} g = (-1)^{k+1}$, completing the inductive step. □

Corollary 6.9. *1.* $\operatorname{sign}(\mathrm{Id}) = 1$.

2. $\operatorname{sign} f_1 f_2 = \operatorname{sign} f_1 \cdot \operatorname{sign} f_2$.

3. $\operatorname{sign}(f)^{-1} = (\operatorname{sign} f)^{-1} = \operatorname{sign} f$.

Proof. The first two statements are clear and the third follows since

$$\operatorname{sign}(f)^{-1} \operatorname{sign} f = \operatorname{sign} \mathrm{Id} = 1.$$

Alternatively, note that if $f = t_{i_1,j_1} \circ t_{i_2,j_2} \circ \cdots \circ t_{i_k,j_k}$ is a product of k transpositions, then

$$f^{-1} = t_{i_k,j_k}^{-1} \circ \cdots \circ t_{i_2,j_2}^{-1} \circ t_{i_1,j_1}^{-1} = t_{i_k,j_k} \circ \cdots \circ t_{i_2,j_2} \circ t_{i_1,j_1}$$

is also a product of k transpositions (where we have used repeatedly $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$ as well as the fact that $t_{i,j}^{-1} = t_{i,j}$). □

It follows that a permutation can never be simultaneously a product of an even and an odd number of transpositions. For example, in S_3 the elements with sign -1 are exactly the transpositions, and the elements with sign $+1$ are the elements Id, r, r^2. In general, one can show that, if $n \geq 2$, there are $n!/2$ elements in S_n with sign $+1$ and $n!/2$ elements with sign -1.

Finally we note that $\det(\mathbf{e}_{f(1)}, \dots, \mathbf{e}_{f(n)}) = \operatorname{sign} f = \operatorname{sign} f \det(\mathbf{e}_1, \dots, \mathbf{e}_n)$, but, using the fact that every permutation is a product of transpositions and the properties of alternating multilinear functions, we see that for **every** alternating multilinear function D, we have

$$D(\mathbf{e}_{f(1)}, \dots, \mathbf{e}_{f(n)}) = (\operatorname{sign} f) D(\mathbf{e}_1, \dots, \mathbf{e}_n).$$

Let us return to the problem of finding a formula for the general alternating multilinear function D. We try to evaluate D on vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, thought of as the columns of the matrix $A = (a_{ij})$. Thus $\mathbf{v}_1 = (a_{11}, \dots, a_{n1}) =$

$\sum_{j=1}^n a_{j1}\mathbf{e}_j$ and similarly $\mathbf{v}_i = \sum_{j=1}^n a_{ji}\mathbf{e}_j$ for every i. So we are confronted with trying to evaluate the expression

$$D(\sum_{j=1}^n a_{j1}\mathbf{e}_j, \ldots, \sum_{j=1}^n a_{jn}\mathbf{e}_j).$$

By the results of the previous section, this expression will be the sum of all possible terms of the form

$$D(a_{i_1 1}\mathbf{e}_{i_1}, \ldots, a_{i_n n}\mathbf{e}_{i_n}) = a_{i_1 1}\cdots a_{i_n n} D(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_n}).$$

Since D is in addition alternating, we can throw out all the terms which we know must be zero, namely those terms where $i_j = i_k$ for some $j \neq k$. So we need only carry through the terms where $i_j \neq i_k$ if $i \neq j$. In this case the function $f(j) = i_j$ is a one-to-one function from the set $\{1, \ldots, n\}$ to itself, which is then necessarily onto since $\{1, \ldots, n\}$ is finite. So the sum reduces to the sum over all $f \in S_n$ of terms

$$a_{f(1),1}\cdots a_{f(n),n} D(\mathbf{e}_{f(1)}, \ldots, \mathbf{e}_{f(n)}) = c(\operatorname{sign} f) a_{f(1),1}\cdots a_{f(n),n},$$

where $c = D(\mathbf{e}_1, \ldots, \mathbf{e}_n)$. Applying this now to $D = \det$, where we know $c = 1$, we see that

$$\det(a_{ij}) = \sum_{f \in S_n} (\operatorname{sign} f) a_{f(1),1}\cdots a_{f(n),n}. \qquad (*)$$

Comparing this with the formula for D, we see that $D = c\det$. □

As an example of the formula for det, we can recover the formula for the determinant of a 3×3 matrix, by working out all of the terms $a_{f(1),1}a_{f(2),2}a_{f(3),3}$ as f runs over the elements of S_3, by using the description of these elements and their signs.

One final comment: we have only considered alternating n-multilinear functions on $\mathbb{R}^n$. We could extend this definition to alternating d-multilinear functions $M(\mathbf{v}_1, \ldots, \mathbf{v}_d)$: by definition, the d-multilinear function M is alternating $\iff$ whenever $\mathbf{v}_i = \mathbf{v}_{i+1}$, then $M(\mathbf{v}_1, \ldots, \mathbf{v}_d) = 0$. It follows as in the case $d = n$ that M is replaced by $-M$ whenever we switch two terms:

$$M(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_d) = -M(\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_d).$$

More generally, for all $f \in S_d$,

$$M(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(d)}) = (\operatorname{sign} f) M(\mathbf{v}_1, \ldots, \mathbf{v}_d).$$

For $d = n$, we have seen that there is a unique alternating n-multilinear function up to a scalar multiple. For $d > n$, a similar argument shows that the only alternating d-multilinear function is the zero function $M(\mathbf{v}_1, \ldots, \mathbf{v}_d) = 0$ for all $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^n$. This is because M can be expanded out with coefficients of the form $M(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_d})$. For $d > n$, at least two of the indices $i_1, \ldots i_d$ must be equal, and so all coefficients are zero by the alternating property so that $M = 0$. One can check that, for $1 \leq d \leq n$, the vector space of all alternating d-multilinear functions has dimension given by the binomial coefficient $\binom{n}{d} = \dfrac{n!}{d!(n-d)!}$ (and every alternating d-multilinear function is built out of $d \times d$ determinants).

6.3 Further properties of the determinant

Let us give some corollaries of Proposition 6.3 and the formula $(*)$.

Proposition 6.10. *Let A be an $n \times n$ matrix.*

1. $\det A = \det({}^tA)$, *where, if $A = (a_{ij})$, then ${}^tA$ is the matrix with $(i,j)^{th}$ entry a_{ji}.*

2. $\det A$ *can be evaluated by expanding about the i^{th} row for any i:*

$$\det A = \sum_k (-1)^{i+k} a_{ik} \det A_{ik}.$$

   Here $(a_{i1}, \ldots, a_{in})$ is the i^{th} row of A and A_{ik} is the $(n-1) \times (n-1)$ matrix obtained by deleting the i^{th} row and k^{th} column of A.

3. $\det A$ *can also be evaluated by expanding about the j^{th} column of A:*

$$\det A = \sum_k (-1)^{k+j} a_{kj} \det A_{kj}.$$

Proof. The first statement follows since by $(*)$

$$\det A = \sum_{f \in S_n} (\operatorname{sign} f) a_{f(1),1} \cdots a_{f(n),n}$$

whereas

$$\det {}^tA = \sum_{f \in S_n} (\operatorname{sign} f) a_{1,f(1)} \cdots a_{n,f(n)}.$$

On the other hand it is easy to see that $a_{1,f(1)} \cdots a_{n,f(n)} = a_{f^{-1}(1),1} \cdots a_{f^{-1}(n),n}$ and we have shown that $\operatorname{sign} f = \operatorname{sign} f^{-1}$. Thus

$$\begin{aligned} \det {}^t A &= \sum_{f \in S_n} (\operatorname{sign} f^{-1}) a_{f^{-1}(1),1} \cdots a_{f^{-1}(n),n} \\ &= \sum_{f \in S_n} (\operatorname{sign} f) a_{f(1),1} \cdots a_{f(n),n}, \end{aligned}$$

since summing over all terms $f \in S_n$ is the same as summing over all terms f^{-1}. Thus $\det A = \det {}^t A$. The second statement follows since the argument showing that det has the properties (1)—(3) also works by expanding about any row, not just the first, and by applying the uniqueness part of the proposition. The main point of the signs $(-1)^{i+k}$ is that the sign for the diagonal term $i = k$ is $+1$, which insures the correct sign $\det \mathrm{Id} = +1$. One way to remember the signs is to visualize a checkerboard pattern, starting with $+$ in the upper left hand corner and alternating (then the signs along the diagonal are always $+$):

$$\begin{pmatrix} + & - & + & \dots \\ - & + & - & \dots \\ + & - & + & \dots \end{pmatrix}.$$

Thus we have (2), and (3) follows because taking transposes switches rows and columns. □

For example, if A is an upper triangular matrix ($a_{ij} = 0$ if $i > j$). then $\det A$ is the product of the diagonal entries of A, since in this case ${}^t A$ is a lower triangular matrix with the same diagonal entries and vice versa.

As another application, we have the important result:

Proposition 6.11. *For all $n \times n$ matrices A and B,* $\det AB = \det A \det B$.

Proof. Fix A, and consider the function $D(\mathbf{v}_1, \dots, \mathbf{v}_n) = \det(A\mathbf{v}_1, \dots, A\mathbf{v}_n)$. By definition of matrix multiplication, if B is the matrix with columns $\mathbf{v}_1, \dots, \mathbf{v}_n$, then AB has columns $A\mathbf{v}_1, \dots, A\mathbf{v}_n$, so that we can also think of D as the function on $n \times n$ matrices given by $D(B) = \det AB$. Now it is easy to check that D is alternating and multilinear. It follows that $D = c \det$ for some constant c; in fact $c = D(I)$. But $D(I) = \det(AI) = \det A$. It follows that, for all B, $\det AB = \det A \det B$. □

Corollary 6.12. *Let A be invertible, with inverse A^{-1}. Then* $\det A \neq 0$, *and*

$$\det(A^{-1}) = \frac{1}{\det A}.$$

Proof. Apply Proposition 6.11 to the product $A \cdot A^{-1} = \mathrm{Id}$, we see that $\det A \cdot \det(A^{-1}) = 1$. □

Corollary 6.13. *If C is an invertible $n \times n$ matrix and A is any $n \times n$ matrix, then $\det A = \det(CAC^{-1})$. Thus, the determinant of a linear map can be read off from its matrix with respect to any basis.*

Proof. Immediate from $\det(CAC^{-1}) = \det C \cdot \det A \cdot \det(C^{-1}) = \det C \cdot \det A \cdot (\det C)^{-1} = \det A$. □

Thus for example if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathbb{R}^n$ and, for every i, there exists a $d_i \in \mathbb{R}$ such that $A\mathbf{v}_i = d_i\mathbf{v}_i$, then $\det A = d_1 \cdots d_n$.

Corollary 6.14. *The matrix A has an inverse if and only if $\det A \neq 0$. Equivalently, the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent if and only if $\det(\mathbf{v}_1, \ldots, \mathbf{v}_n) \neq 0$.*

Proof. We have seen that A has an inverse if and only if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent. The above corollary says that if A has an inverse, then $\det A \neq 0$. Conversely, if A does not have an inverse, then the columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of A are linearly dependent. Thus one of the $\mathbf{v}_i$, say for simplicity of notation $\mathbf{v}_n$, is a linear combination of the others: $\mathbf{v}_n = \sum_{i=1}^{n-1} t_i\mathbf{v}_i$. Expanding out $\det(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ by multilinearity we get a sum of terms involving $\det(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_i) = 0$. Thus $\det(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$. □

In fact, we can use the determinant to find a closed form for A^{-1} in case $\det A \neq 0$.

Proposition 6.15 (Cramer's rule I)**.** *Let $A = (a_{ij})$ be an $n \times n$ matrix, and set $\tilde{A} = ((-1)^{i+j} \det A_{ji})$. Here as usual A_{ij} denotes the $(n-1) \times (n-1)$ matrix formed from A be deleting the i^{th} row and j^{th} column; note the switch in indices in the definition of $\tilde{A}$. Then $A \cdot \tilde{A} = (\det A) \cdot \mathrm{Id}$. In particular, if $\det A \neq 0$, then*

$$A^{-1} = \frac{1}{\det A}\tilde{A}.$$

Proof. Recall that the $(i,j)^{\text{th}}$ entry of $A \cdot \tilde{A}$ is equal to

$$\sum_{\ell=1}^{n} a_{i\ell}(-1)^{\ell+j} \det A_{j\ell}.$$

If $i = j$, this sum is just the expansion of $\det A$ along the i^{th} row, and so all the diagonal entries of $A\tilde{A}$ are equal to $\det A$. For $i \neq j$, let $A(i,j)$ be

the $n \times n$ matrix obtained by replacing the j^{th} row $\mathbf{r}_j$ of of A by the i^{th} row $\mathbf{r}_i$. Thus $A(i,j)$ has a repeated row and so (since ${}^tA(i,j)$ has a repeated column) $\det A(i,j) = \det {}^tA(i,j) = 0$. But it is easy to see that, for $i \neq j$, $\sum_{\ell=1}^n a_{i\ell}(-1)^{\ell+j} \det A_{j\ell}$ is the expansion of $\det A(i,j)$ along the j^{th} row and thus is zero. Hence $A\tilde{A}$ has diagonal terms equal to $\det A$ and has all other terms zero, so that $A\tilde{A} = (\det A)\,\mathrm{Id}$. □

In case A is invertible, Cramer's rule gives an explicit solution for the system $A \cdot \mathbf{x} = \mathbf{b}$. However, there is a simple formula, also called Cramer's rule, for this solution:

Proposition 6.16 (Cramer's rule II)**.** *Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ are linearly independent. If $\sum_i x_i \mathbf{v}_i = \mathbf{b}$, then, for all i,*

$$x_i = \frac{\det(\mathbf{v}_1, \ldots, \mathbf{b}, \ldots, \mathbf{v}_n)}{\det(\mathbf{v}_1, \ldots, \mathbf{v}_n)},$$

where for x_i the term $\mathbf{b}$ occurs in the i^{th} place.

Proof. Consider the expression $\det(\mathbf{v}_1, \ldots, \mathbf{b}, \ldots, \mathbf{v}_n)$, where $\mathbf{b}$ replaces the term $\mathbf{v}_i$ in the i^{th} place. Substitute $\mathbf{b} = \sum_j x_j \mathbf{v}_j$ and expand out. Using the alternating property, all the terms drop out because of repeats except for the term $x_i \mathbf{v}_i$. This says that

$$\det(\mathbf{v}_1, \ldots, \mathbf{b}, \ldots, \mathbf{v}_n) = x_i \det(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n)$$

whether or not $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent. If they are, then we can divide by $\det(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ to obtain the formula for x_i. □

It is easy to see that the second form of Cramer's rule is the same as the first (for example, expand out $\det(\mathbf{v}_1, \ldots, \mathbf{b}, \ldots, \mathbf{v}_n)$ along the i^{th} column).

In general the formula for the inverse is too unwieldy to apply in practice. The one useful formula is that of the inverse of a 2×2 matrix: if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $\det A = ad - bc \neq 0$, then

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

If one really wants to find an inverse matrix or solve a system of linear equations, the most efficient method in practice is to use row reduction with bookkeeping as previously described.

We can also use row reduction to find the determinant of a matrix. In fact, the actual formula $(*)$ involves too many arithmetic operations to be practical to implement on a computer. The idea here is to use only those operations involved in row reduction which preserve the determinant, to bring the matrix into upper triangular form. Here we are allowed to subtract off a multiple of $\mathbf{v}_j$ from $\mathbf{v}_i$, if $i \neq j$, and to switch two vectors $\mathbf{v}_i, \mathbf{v}_j$ if we keep track of the sign change, but not to multiply $\mathbf{v}_i$ by a scalar. For example, to find $\det \begin{pmatrix} 2 & 3 \\ 5 & -2 \end{pmatrix}$, do the following:

$$\begin{pmatrix} 2 & 3 \\ 5 & -2 \end{pmatrix} \to \begin{pmatrix} 2 & 3 \\ 0 & -\frac{19}{2} \end{pmatrix},$$

where we subtract $\frac{5}{2}(2,3)$ from the second row to get $(0, -\frac{19}{2})$. From the upper triangular matrix we can read off: $\det = -19$.

What is the meaning of the determinant? The above discussion gave many properties of the determinant, all of which hold for a general field F and matrices with coefficients in F. Here we discuss some properties which are only meaningful for the field $\mathbb{R}$. For a 2×2 matrix A, the determinant is connected with area as follows. Suppose that $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with columns $\mathbf{v} = (a, c), \mathbf{w} = (b, d)$ which are linearly independent. Then $\mathbf{v}$ and $\mathbf{w}$ define a parallelogram P in $\mathbb{R}^2$, which is the image under A of the unit square $S = \{(x_1, x_2) : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$. Thus

$$P = \{x_1\mathbf{v} + x_2\mathbf{w} : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}.$$

What is the area of P?

Claim 6.17. *The area of P is $|\det A|$.*

To see this, the area of P is its base times the height (see Figure 6.1). The base has length $\|\mathbf{v}\|$ and the height has length $\|\mathbf{w}\||\sin\theta|$, where θ is the angle between $\mathbf{v}$ and $\mathbf{w}$. Thus the area is

$$\begin{aligned} \|\mathbf{v}\|\|\mathbf{w}\||\sin\theta| &= \|\mathbf{v}\|\|\mathbf{w}\|\sqrt{1-\cos^2\theta} \\ &= \|\mathbf{v}\|\|\mathbf{w}\|\sqrt{1 - \frac{(\langle \mathbf{v}, \mathbf{w}\rangle)^2}{\|\mathbf{v}\|^2\|\mathbf{w}\|^2}} \\ &= \sqrt{\|\mathbf{v}\|^2\|\mathbf{w}\|^2 - (\langle \mathbf{v}, \mathbf{w}\rangle)^2}. \end{aligned}$$

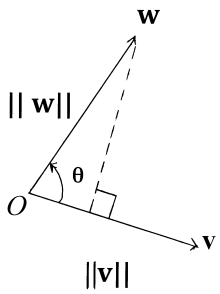

Figure 6.1: The area of a parallelogram

Now plugging in for $\mathbf{v}$ and $\mathbf{w}$, the expression in the square root is

$$\begin{gathered}(a^2+c^2)(b^2+d^2)-(ab+cd)^2= \\ =a^2b^2+a^2d^2+b^2c^2+c^2d^2-a^2b^2-2abcd-c^2d^2 \\ =a^2d^2-2abcd+b^2c^2=(ad-bc)^2.\end{gathered}$$

Thus the area of P is $|ad-bc|=|\det A|$ as claimed.

Note that, if the vectors $\mathbf{v}$ and $\mathbf{w}$ are not linearly independent, then the set P defined above is the origin or a line segment, and so its area, in any reasonable sense, is zero. This agrees with the fact that $\det A=0$ in this case as well.

A similar result holds in $\mathbb{R}^3$. Suppose that A is a 3×3 matrix with linearly independent columns $\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3$. Then the vectors $\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3$ complete to a *parallelpiped* P, which by defintion is the image of the unit cube in $\mathbb{R}^3$ under A, or equivalently the set

$$P=\{x_1\mathbf{v}_1+x_2\mathbf{v}_2+x_3\mathbf{v}_3 : 0\le x_1\le 1, 0\le x_2\le 1, 0\le x_3\le 1\}.$$

Here, P is analogous to a cube: it has 6 "faces", each of which is a parallelogram. In this case, one can argue that the volume of P is $|\det A|$ again.

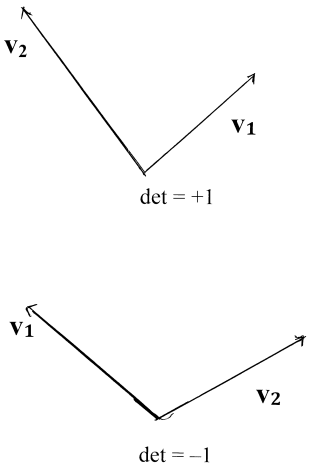

Figure 6.2: The two possible signs for $\det A$

A very similar result holds in $\mathbb{R}^n$, once we have defined the "n-volume" of certain subsets.

The remaining question is the meaning of the sign of $\det A$, in case $\det A \neq 0$. This is a subtle point, and we can only explain it in dimension 2 and 3. First suppose that A is a 2×2 matrix, with columns $\mathbf{v}_1, \mathbf{v}_2$. It turns out that $\det A > 0$ if the shorter angle from $\mathbf{v}_1$ to $\mathbf{v}_2$ is in the counterclockwise direction and that $\det A > 0$ if the shorter angle from $\mathbf{v}_1$ to $\mathbf{v}_2$ is in the clockwise direction. (See Figure 6.2.) Thus positive determinant preserves the idea of counterclockwise/clockwise direction and negative determinant reverses it. In dimension three, the sign of the determinant is connected with "right-handedness" or the "right-hand rule." Given A a 3×3 matrix with linearly independent columns $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, we say that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are a *right-handed triple* if the following holds: take your right hand, and curl your fingers from $\mathbf{v}_1$ to $\mathbf{v}_2$ (using the angle between them which is less than π). If you thumb is pointing out from the same side of the plane spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$ as $\mathbf{v}_3$, then $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are a right-handed triple; otherwise they are a left-handed triple. In the first case $\det A > 0$, and in the second case $\det A < 0$. (See Figure 6.3.)

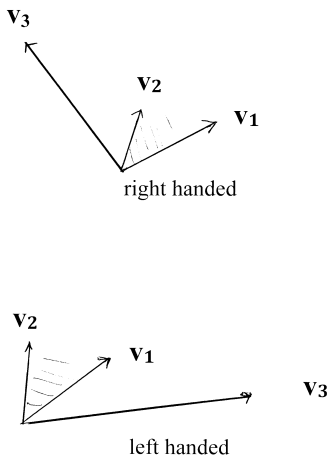

Figure 6.3: Right and left handed triples

6.4 Eigenvalues and eigenvectors

A problem which arises frequently in mathematics as well as in many questions in physics or engineering is the following: Given an $n \times n$ matrix A (or equivalently a linear map $F\colon \mathbb{R}^n \to \mathbb{R}^n$), we seek a nonzero vector $\mathbf{v}$ such that the effect of applying A to $\mathbf{v}$ is the same as multiplying $\mathbf{v}$ by a scalar. We formalize this in a definition:

Definition 6.18. Let A be an $n \times n$ matrix A (equivalently, let $F\colon \mathbb{R}^n \to \mathbb{R}^n$ be a linear map). An *eigenvector* $\mathbf{v}$ for A (or F) is a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ such that $A \cdot \mathbf{v} = \lambda \mathbf{v}$ (equivalently, $F(\mathbf{v}) = \lambda \cdot \mathbf{v}$) for some (in fact unique) scalar $\lambda \in \mathbb{R}$). In this case, we say that λ is an *eigenvalue of A (or F) belonging to the eigenvector* $\mathbf{v}$.

Example 6.19. Here are some examples of eigenvectors and eigenvalues:

1. A nonzero vector $\mathbf{v}$ is an eigenvector of A with eigenvalue zero $\iff$ $\mathbf{v}$ is a nonzero element of $\operatorname{Ker} A$.

2. If $A = \mathrm{Id}$, then every nonzero $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector with eigenvalue 1. Similarly, if $A = \lambda \cdot \mathrm{Id}$, then every nonzero $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector

with eigenvalue λ.

3. If A is a diagonal matrix with diagonal entries $a_{ii} = \lambda_i$, then, for every i, $\mathbf{e}_i$ is an eigenvector of A with eigenvalue λ_i. Conversely, if, for every i, $\mathbf{e}_i$ is an eigenvector of A with eigenvalue λ_i, then A is a diagonal matrix with diagonal entries λ_i.

The next proposition tells us how to find eigenvalues:

Proposition 6.20. *The real number λ is an eigenvalue of $A \iff \det(\lambda \operatorname{Id} - A) = 0$.*

Proof. λ is an eigenvalue of $A \iff$ there exists a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ such that $A \cdot \mathbf{v} = \lambda \cdot \mathbf{v} \iff$ there exists a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ such that $(\lambda \operatorname{Id} - A) \cdot \mathbf{v} = \mathbf{0} \iff \operatorname{Ker}(\lambda \operatorname{Id} - A) \neq \{\mathbf{0}\} \iff \det(\lambda \operatorname{Id} - A) = 0$. □

Definition 6.21. Let A be an $n \times n$ matrix. Then the *characteristic polynomial* $p_A(t)$ is the function $\det(t \operatorname{Id} - A)$.

For example, suppose that $\mathbb{R}^n$ has a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ consisting of eigenvectors of A, with eigenvalues λ_i. Then the matrix $t \operatorname{Id} - A$ with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a diagonal matrix with diagonal entries $t - \lambda_i$. Thus, since we can compute determinants by taking the determinant of the matrix with respect to **any** basis, it follows that $p_A(t) = (t - \lambda_1) \cdots (t - \lambda_n)$.

Lemma 6.22. *If A is an $n \times n$ matrix, then the characteristic polynomial $p_A(t)$ is a polynomial in t of degree n and leading coefficient* 1.

Proof. Examining the formula $(*)$ for the determinant given at the end of the proof of Proposition 6.3, we see that

$$p_A(t) = \sum_{f \in S_n} (\operatorname{sign} f) A_{f(1),1}(t) \cdots A_{f(n),n}(t),$$

where $A_{ij}(t)$ is either the constant $-a_{ij}$, if $i \neq j$, or the degree one polynomial $t - a_{ii}$, if $i = j$. From this it is clear that a product $A_{f(1),1}(t) \cdots A_{f(n),n}(t)$ has degree at most n, and has degree $n \iff f = \operatorname{Id}$, in which case it is of the form $(t - a_{ii}) \cdots (t - a_{nn})$. Thus the degree of $p_A(t)$ is n and (since $\operatorname{sign} \operatorname{Id} = 1$) its leading coefficient is 1. □

Corollary 6.23. *The eigenvalues of A are the roots of the characteristic polynomial $p_A(t)$. In particular, an $n \times n$ matrix has at most n different eigenvalues.*

If we are given a matrix A and want to find its *eigenvectors* we proceed as follows:

1. First compute the characteristic polynomial $p_A(t)$.
2. Next, find the roots of $p_A(t)$.
3. Finally, if λ is a root of $p_A(t)$, then determine $\mathrm{Ker}(\lambda \,\mathrm{Id} - A)$.

Here Step (1) is computationally effective but tedious. Step (2) may not be possible to carry out if n is large, at least in an explicit way. But, once we have found a root λ of $p_A(t)$, then it is always computationally effective via row reduction to determine $\mathrm{Ker}(\lambda \,\mathrm{Id} - A)$, i.e. to carry out Step (3).

Example 6.24. Take $A = \begin{pmatrix} -4 & 6 \\ 9 & 11 \end{pmatrix}$ so that $t \,\mathrm{Id} - A = \begin{pmatrix} t+4 & -6 \\ -9 & t-11 \end{pmatrix}$. Hence

$$p_A(t) = (t+4)(t-11) - 54 = t^2 - 7t - 98 = (t-14)(t+7).$$

Thus there is an eigenvector $\mathbf{v}$ with eigenvalue 14 of A corresponding to $\mathrm{Ker}(14 \,\mathrm{Id} - A) = \begin{pmatrix} 18 & -6 \\ -9 & 3 \end{pmatrix}$. Clearly we must take $\mathbf{v}$ to be any nonzero scalar multiple of $(1, 3)$; then $A \cdot (1, 3) = (14, 42) = 14(1, 3)$. Similarly, there is an eigenvector $\mathbf{w}$ with eigenvalue -7 of A corresponding to $\mathrm{Ker}(-7 \,\mathrm{Id} - A) = \begin{pmatrix} -3 & -6 \\ -9 & -18 \end{pmatrix}$, and we can take $\mathbf{w}$ to be any nonzero scalar multiple of $(2, -1)$, so that $A \cdot (2, -1) = (-14, 7) = (-7)(2, -1)$.

Lest the above example make one overly optimistic about finding eigenvectors, there are many matrices which have **no** eigenvectors, at least in $\mathbb{R}^n$. For example, if A is any matrix of the form $A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ with $b \neq 0$, then $p_A(t) = (t-a)^2 + b^2 = t^2 - 2at + (a^2 + b^2)$, and this polynomial has no real roots as its discriminant is $4a^2 - 4a^2 - 4b^2 = -4b^2 < 0$. However, $p_A(t)$ has the two **complex** (conjugate) roots $a \pm bi$. We can then apply the above procedure to find **complex** eigenvectors of A, i.e. nonzero vectors in $\mathbb{C}^2$, belonging to the eigenvalues $a \pm bi$. For example, $(a + bi) \,\mathrm{Id} - A = \begin{pmatrix} bi & -b \\ b & bi \end{pmatrix}$, and so $\mathrm{Ker}((a + bi) \,\mathrm{Id} - A)$ is the (complex) span of the vector $(1, i)$. It is then easy to check that $A \cdot (1, i) = (a + bi)(1, i)$. Similarly, $(1, -i)$ is an eigenvector of A with eigenvalue $a - bi$.

A deep fact about the complex numbers is the following: Recall that a (complex) polynomial is an expression of the form $f(t) = \sum_{i=0}^{n} a_i t^i$ where the a_i, i.e. the coefficients of $f(t)$, are complex numbers. We say that $f(t)$ has degree n if $a_n \neq 0$. In the usual way, by evaluation, $f(t)$ defines a function from $\mathbb{C}$ to $\mathbb{C}$, also denoted by $f(t)$. The complex number λ is a (complex) *root* of $f(t)$ if $f(\lambda) = 0$. In the usual way, roots of $f(t)$ correspond to linear factors: λ is a root of $f(t) \iff$ there exists a polynomial $g(t)$ of degree $n-1$ such that $f(t) = (t-\lambda)g(t)$.

Theorem 6.25 (Fundamental Theorem of Algebra)**.** *Let $f(t)$ be a complex polynomial of degree $n > 0$. Then there exists a complex root λ of $f(t)$. Hence, by induction, $f(t)$ can be factored into linear factors: there exists a sequence of complex numbers $\lambda_1, \ldots, \lambda_n$, unique up to order, but not necessarily distinct, such that $f(t) = c(t-\lambda_1)\cdots(t-\lambda_n)$.*

As a corollary we see:

Proposition 6.26. *Let $A \in \mathbb{M}_n(\mathbb{C})$ be an $n \times n$ matrix with complex coefficients. Then there exists a complex eigenvector $\mathbf{v} \in \mathbb{C}^n$ for A, in other words there exists a nonzero element $\mathbf{v} \in \mathbb{C}^n$ and a $\lambda \in \mathbb{C}$ such that $A \cdot \mathbf{v} = \lambda \mathbf{v}$.*

If $A \in \mathbb{M}_n = \mathbb{M}_n(\mathbb{R})$ and n is odd, then there exists a (real) eigenvector $\mathbf{v} \in \mathbb{R}^n$ for A.

Proof. The first statement follows directly from the Fundamental Theorem of Algebra, and the second from the standard calculus fact that every (real) polynomial of odd degree has at least one (real) root. □

One can say a little more in the real case. Recall that, if $\lambda = a + bi$ is a complex number, then its *complex conjugate* $\bar{\lambda}$ is by definition the complex number $a - bi$. Note that $\bar{\lambda} = \lambda \iff \lambda$ is real. A standard result about polynomials with real coefficients says that, if $f(t)$ is a polynomial with real coefficients and λ is a (complex) root of $f(t)$, then $\bar{\lambda}$ is also a root. Of course, if λ is real, this doesn't say anything new. In particular, if A is a real $n \times n$ matrix and λ is a complex eigenvalue of A, then so is $\bar{\lambda}$. In fact, if, for a vector $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{C}^n$, let $\bar{\mathbf{v}}$ denote the vector $(\bar{v}_1, \ldots, \bar{v}_n)$. Suppose that A is a real $n \times n$ matrix such that $\mathbf{v} \in \mathbb{C}^n$ is a complex eigenvector for A with eigenvalue λ. Let $\bar{\mathbf{v}}$ be the vector obtained by taking the complex conjugate of all of the entries of $\mathbf{v}$. Then it is easy to check that, if A is a **real** matrix, then $\overline{A\mathbf{v}} = A\bar{\mathbf{v}}$. Thus, if $A\mathbf{v} = \lambda\mathbf{v}$, then

$$A\bar{\mathbf{v}} = \overline{A\mathbf{v}} = \overline{\lambda\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}.$$

Hence $\bar{\mathbf{v}}$ is a complex eigenvector for A with eigenvalue $\bar{\lambda}$. We have seen this directly in the 2×2 example given above.

As we have seen, part of the problem with finding eigenvectors lies in the possibility that $p_A(t)$ has complex, non-real roots. However, there is another issue connected with finding eigenvectors. For example, let A be any matrix of the form $A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$, then $p_A(t) = (t-\lambda)^2$, and thus has one repeated root. Here A is a real matrix $\iff \lambda \in \mathbb{R}$, and in fact already the case $\lambda = 0$ illustrates the main issue. Here $\lambda \operatorname{Id} - A = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}$, and hence $\operatorname{Ker}(\lambda \operatorname{Id} - A)$ is just the span of $\mathbf{e}_1$.

We will just show here that the problem arises only when the characteristic polynomial has repeated roots, although one can say a great deal about matrices in general, both real and complex.

Lemma 6.27. *Let A be a real or complex $n \times n$ matrix, and let $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{C}^n$ be eigenvectors for A with eigenvalues $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$. Suppose that the λ_i are distinct complex numbers, i.e. for $i \neq j$, $\lambda_i \neq \lambda_j$. Then $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent.*

Proof. Suppose that $t_1\mathbf{v}_1 + \cdots + t_r\mathbf{v}_r = \mathbf{0}$ with not all of the t_i equal to 0. After renumbering the $\mathbf{v}_i$ and eliminating those with $t_i = 0$, we can asume that none of the t_i is equal to 0. Clearly we must have $r > 1$, since by definition an eigenvector is not equal to $\mathbf{0}$. Applying A to the above sum gives

$$t_1\lambda_1\mathbf{v}_1 + \cdots + t_r\lambda_r\mathbf{v}_r = \mathbf{0};$$
$$t_1\lambda_r\mathbf{v}_1 + \cdots + t_r\lambda_r\mathbf{v}_r = \mathbf{0},$$

where the second line is simply the result of multiplying the equality $t_1\lambda_1\mathbf{v}_1 + \cdots + t_r\lambda_r\mathbf{v}_r = \mathbf{0}$ by the scalar λ_r. Subtracting gives an equality

$$t_1(\lambda_1 - \lambda_r)\mathbf{v}_1 + \cdots + t_{r-1}(\lambda_{r-1} - \lambda_r)\mathbf{v}_{r-1} = \mathbf{0},$$

involving only $r-1$ of the $\mathbf{v}_i$. By assumption $\lambda_i \neq \lambda_r$ for all $i \leq r-1$, and hence none of the coefficients $t_i(\lambda_i - \lambda_r)$ is equal to 0. Continuing in this way, we eventually reach an equation of the form $s_1\mathbf{v}_1 = \mathbf{0}$ with $s_1 \neq 0$, and as we have seen this is a contradiction. □

Corollary 6.28. *Let A be a real $n \times n$ matrix and suppose that the characteristic polynomial $p_A(t)$ has n distinct real roots. Then there exists a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ of eigenvectors for A.*

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the distinct real roots of $p_A(t)$ and let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ be corresponding eigenvectors for A. Then the $\mathbf{v}_i$ are linearly independent (with complex and hence real coefficients). Since there are n of them, they are a basis for $\mathbb{R}^n$. □

Finding a basis of eigenvectors for a matrix A is called *diagonalizing* A. Why does one care about this? Eigenvectors and their eigenvalues give a lot of useful information about a matrix A and its overall behavior. It is also computationally very helpful to diagonalize A. For example, if $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ is a basis of eigenvectors for A, then the matrix for A with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a diagonal matrix with diagonal entries λ_i. If B is the change of basis matrix ($B \cdot \mathbf{e}_i = \mathbf{v}_i$), then

$$A = B \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} B^{-1},$$

and so it is very easy to compute the powers of A: using the fact that

$$\begin{aligned} A^k = (BDB^{-1})^k &= \underbrace{(BDB^{-1})(BDB^{-1}) \cdots (BDB^{-1})}_{k} \\ &= BD(B^{-1}B)D(B^{-1}B) \cdots (B^{-1}B)DB^{-1} = BD^kB^{-1}, \end{aligned}$$

we see that

$$A^k = B \begin{pmatrix} \lambda_1^k & 0 & \ldots & 0 \\ 0 & \lambda_2^k & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n^k \end{pmatrix} B^{-1}.$$

For example, for systems of linear differential equations with constant coefficients one needs to compute a matrix power series $e^{tA} = \sum_{k=0}^{\infty} (tA)^k/k!$. If A can be diagonalized, the above shows that

$$e^{tA} = B \begin{pmatrix} e^{\lambda_1 t} & 0 & \ldots & 0 \\ 0 & e^{\lambda_2 t} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e^{\lambda_n t} \end{pmatrix} B^{-1}.$$

Finally, we prove a weak result about existence of eigenvalues in the real case, which we shall sue in the next section in the proof of the spectral theorem.

Proposition 6.29. *Let $A \in \mathbb{M}_n(\mathbb{R})$. Then either A has a real eigenvector $\mathbf{v}$ or there exists a two-dimensional vector subspace W of $\mathbb{R}^n$ such that $A(W) \subseteq W$.*

Proof. By the Fundamental Theorem of Algebra, the characteristic polynomial $p_A(t)$ has a complex root $\lambda = a + bi$. If $\lambda \in \mathbb{R}$, we are done. So assume $\lambda \notin \mathbb{R}$. As $\lambda \in \mathbb{C}$, there is a complex eigenvector $\mathbf{v}$, i.e. a nonzero vector $\mathbf{v} \in \mathbb{C}^n$ such that $A\mathbf{v} = \lambda\mathbf{v}$. Then, as we have seen, $\bar{\mathbf{v}}$ is also a complex eigenvector of A with eigenvalue $\bar{\lambda}$. We can define $\mathbf{w}_1 = \operatorname{Re}\mathbf{v} = \dfrac{1}{2}(\mathbf{v} + \bar{\mathbf{v}})$ and $\mathbf{w}_2 = \operatorname{Im}\mathbf{v} = \dfrac{1}{2i}(\mathbf{v} - \bar{\mathbf{v}})$, so that $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$ (i.e. their components are real). A straightforward calculation then shows that the vector subspace $W = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2\}$ of $\mathbb{R}^n$ satisfies: $A(W) \subseteq W$. In fact, we have the equalities (recall that, if $\lambda = a + bi$, then $\operatorname{Re}\lambda = a = \frac{1}{2}(\lambda + \bar{\lambda})$ and $\operatorname{Im}\lambda = b = \frac{1}{2i}(\lambda - \bar{\lambda})$):

$$a\mathbf{w}_1 = \frac{1}{4}(\lambda + \bar{\lambda})(\mathbf{v} + \bar{\mathbf{v}}) = \frac{1}{4}((\lambda\mathbf{v} + \bar{\lambda}\bar{\mathbf{v}}) + (\lambda\bar{\mathbf{v}} + \bar{\lambda}\mathbf{v})) = \frac{1}{2}\left(A\mathbf{w}_1 + \operatorname{Re}(\lambda\bar{\mathbf{v}})\right);$$
$$b\mathbf{w}_1 = \frac{1}{4i}(\lambda - \bar{\lambda})(\mathbf{v} + \bar{\mathbf{v}}) = \frac{1}{4i}((\lambda\mathbf{v} - \bar{\lambda}\bar{\mathbf{v}}) + (\lambda\bar{\mathbf{v}} - \bar{\lambda}\mathbf{v})) = \frac{1}{2}\left(A\mathbf{w}_2 + \operatorname{Im}(\lambda\bar{\mathbf{v}})\right);$$
$$a\mathbf{w}_2 = \frac{1}{4i}(\lambda + \bar{\lambda})(\mathbf{v} - \bar{\mathbf{v}}) = \frac{1}{4i}((\lambda\mathbf{v} - \bar{\lambda}\bar{\mathbf{v}}) - (\lambda\bar{\mathbf{v}} - \bar{\lambda}\mathbf{v})) = \frac{1}{2}\left(A\mathbf{w}_2 - \operatorname{Im}(\lambda\bar{\mathbf{v}})\right);$$
$$b\mathbf{w}_2 = -\frac{1}{4}(\lambda - \bar{\lambda})(\mathbf{v} - \bar{\mathbf{v}}) = -\frac{1}{4}((\lambda\mathbf{v} + \bar{\lambda}\bar{\mathbf{v}}) - (\lambda\bar{\mathbf{v}} + \bar{\lambda}\mathbf{v})) = -\frac{1}{2}\left(A\mathbf{w}_1 - \operatorname{Re}(\lambda\bar{\mathbf{v}})\right).$$

Hence $A\mathbf{w}_1 = a\mathbf{w}_1 - b\mathbf{w}_2$ and $A\mathbf{w}_2 = b\mathbf{w}_1 + a\mathbf{w}_2$, showing that $A(W) \subseteq W$. (This tedious calculation can be redone by rewriting A, which in the basis $\{\mathbf{v}, \bar{\mathbf{v}}\}$ is the diagonal matrix $\begin{pmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{pmatrix}$, in terms of the basis $\{\mathbf{w}_1, \mathbf{w}_2\}$: in terms of this basis, its matrix is $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$.) In any case, we see directly that $\dim W \leq 2$ and that $A(W) \subseteq W$. If $\dim W = 1$, then A has a real eigenvalue, and so in either case we are done. □

A vector subspace W of $\mathbb{R}^n$ such that $A(W) \subseteq W$ is called an *invariant subspace* of A. The proposition then says that every $n \times n$ matrix has a nonzero invariant subspace of dimension at most two.

6.5 Applications to symmetric matrices

We want to apply our general results on determinants to prove two theorems about symmetric matrices. The first is an explicit condition for a symmetric

matrix A to correspond to a *positive definite* quadratic form $Q(\mathbf{x})$. In fact, the aim is to show the following: if A is an $n \times n$ symmetric matrix, let $A_1 = (a_{11})$, $A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$, and more generally let A_d be the $d \times d$ symmetric matrix whose $(i,j)^{\text{th}}$ entry is a_{ij} for $1 \leq i, j \leq d$ (so that $A_n = A$). Then:

Theorem 6.30. *The quadratic form $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$ is positive definite $\iff$ $\det A_d > 0$ for all d, $1 \leq d \leq n$.*

Proof. The basic tool is the fact that, if A is the matrix corresponding to Q and we choose a new basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{R}^n$, then the new quadratic form $Q(\sum_{i=1}^n x_i \mathbf{v}_i)$ corresponds to the matrix ${}^tC \cdot A \cdot C$. Taking determinants and using

$$\det({}^tC \cdot A \cdot C) = \det({}^tC) \det A \det C = (\det C)^2 \det A,$$

since $\det({}^tC) = \det C$, it follows that the sign of $\det A$ is independent of the choice of basis. In particular, if Q is positive definite, then there exists a diagonal basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{R}^n$ such that $Q(\sum_{i=1}^n x_i \mathbf{v}_i) = \sum_{i=1}^n x_i^2$, as in Sylvester's theorem. In this case the determinant of the matrix corresponding to Q in this new basis is the identity and hence its determinant is 1, so that $\det A > 0$. Of course, this would also hold if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is just **some** diagonal basis, i.e. such that $Q(\sum_{i=1}^n x_i \mathbf{v}_i) = \sum_{i=1}^n d_i x_i^2$ with $d_i > 0$. In particular, if Q is positive definite, then $\det A > 0$, and since the restriction of A to the span of $\mathbf{e}_1, \ldots, \mathbf{e}_d$ is also positive definite, in fact $\det A_d > 0$ for every d. This proves the "if" direction of the theorem.

To prove the converse, we begin with the following useful formula. Let A be an $n \times n$ matrix of the form $\begin{pmatrix} A_1 & O \\ O & A_2 \end{pmatrix}$, where A_1 is a $d \times d$ matrix, A_2 is an $(n-d) \times (n-d)$ matrix, and the remaining entries are 0. In other words, $a_{ij} = 0$ unless $i, j \leq d$ or $i, j \geq d+1$. Such a matrix is said to be in *block form*. Then:

Lemma 6.31. *With notation as above,*

$$\det \begin{pmatrix} A_1 & O \\ O & A_2 \end{pmatrix} = \det A_1 \cdot \det A_2.$$

Proof. If $d = n - 1$, so that the last row consist of all zero entries except for a_{nn}, then the lemma follows by expanding about the last row. (This is the only case that we need.) In general, one way to prove this is as follows: thinking of A_2 as a fixed matrix and A_1 as a variable matrix with

d columns $\mathbf{v}_1, \ldots, \mathbf{v}_d$, where each $\mathbf{v}_i \in \mathbb{R}^d$ defines an alternating multilinear function of $\mathbf{v}_1, \ldots, \mathbf{v}_d$, necessarily of the form $\det A_1 c_2(A_2)$, where $c_2(A_2)$ is a constant, i.e. independent of A_1 but clearly depending on A_2. By symmetry $\det A = c_1(A_1) \det A_2$. Taking $A_1 = I$ shows that $c_2(A_2) = \det \begin{pmatrix} I & O \\ O & A_2 \end{pmatrix} = c_1(I) \det A_2$ and taking $A_1 = A_2 = I$ shows that $c_1(I) = 1$. Thus $c_2(A_2) = \det A_2$ and so $\det \begin{pmatrix} A_1 & O \\ O & A_2 \end{pmatrix} = \det A_1 c_2(A_2) = \det A_1 \cdot \det A_2$. □

Returning to the proof of the converse statement, that $\det A_d > 0$ for every $d \implies A$ corresponds to a positive definite quadratic form, we can argue by induction on n. The case $n = 1$ is clear. For the inductive step, assume that the theorem has been proved for $\mathbb{R}^{n-1}$, i.e. for $(n-1) \times (n-1)$ symmetric matrices. Then if A is an $n \times n$ symmetric matrix with $\det A_d > 0$ for every d, by induction we can assume that the restriction of A to the span of $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}$ is positive definite. Let $Q(\mathbf{x})$ be the quadratic form corresponding to A and $B(\mathbf{x}, \mathbf{y})$ the associated bilinear form. Then (by analogy with the proof of Sylvester's theorem) we can define

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}^{\perp_B} = \{\mathbf{v} \in \mathbb{R}^n : B(\mathbf{v}, \mathbf{e}_i) = 0, i = 1, \ldots, n-1\}.$$

We claim that there exists a $\mathbf{v} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}^{\perp_B}$ such that $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}, \mathbf{v}$ is a basis of $\mathbb{R}^n$. To see this, we argue as in the proof of Sylvester's theorem: the kernel of the linear map $F \colon \mathbb{R}^n \to \mathbb{R}^{n-1}$ defined by

$$F(\mathbf{x}) = (B(\mathbf{x}, \mathbf{e}_1), \ldots, B(\mathbf{x}, \mathbf{e}_{n-1})$$

is by definition $\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}^{\perp_B}$, and from $\dim \operatorname{Ker} F + \dim \operatorname{Im} F = n$, $\dim \operatorname{Im} F \leq n-1$ we see that $\dim \operatorname{Ker} F \geq 1$. But if $\mathbf{v}$ is a nonzero vector in $\operatorname{Ker} F$, then since Q is positive definite on $\operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}$, it follows that

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}^{\perp_B} \cap \operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\} = \{\mathbf{0}\},$$

hence that $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}, \mathbf{v}$ are linearly independent and so a basis of $\mathbb{R}^n$.

Using the basis $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}, \mathbf{v}$ above, the the matrix A' of Q with respect to this basis is of the form $\begin{pmatrix} A_{n-1} & O \\ O & \lambda \end{pmatrix}$, where $\lambda = Q(\mathbf{v})$. Moreover, $\det A > 0 \implies \lambda > 0$ and since $Q(x_1\mathbf{e}_1 + \cdots + x_{n-1}\mathbf{e}_{n-1} + x_n\mathbf{v}) = Q(x_1\mathbf{e}_1 + \cdots + x_{n-1}\mathbf{e}_{n-1}) + x_n^2\lambda$ it follows that Q is positive definite. □

Corollary 6.32. *The quadratic form $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$ is negative definite $\iff$ $\det A_1 < 0$, $\det A_2 > 0$, $\det A_3 < 0$, $\ldots$, i.e. $(-1)^d \det A_d > 0$.*

Proof. Clearly Q is negative definite $\iff$ $-Q$ is positive definite. On the other hand, the matrix corresponding to $-Q$ is $-A$. For a $k \times k$ matrix B, we have the formula $\det tB) = t^k \det B$. Applying this to $B = A_d$ and $t = -1$ gives: Q is negative definite $\iff$ $-Q$ is positive definite $\iff$ for all $d, 1 \leq d \leq n$, $(-1)^d \det A_d > 0$. □

Our second result is the spectral theorem:

Theorem 6.33 (Spectral Theorem)**.** *Let A be a symmetric $n \times n$ matrix. Then there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A. In particular, all eigenvalues of A are real, and there exists a basis of $\mathbb{R}^n$ consisting of eigenvetors of A.*

Note: If $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is an orthonormal basis for $\mathbb{R}^n$ such that each $\mathbf{u}_i$ is an eigenvector for A, with eigenvalue λ_i, then

$$\begin{aligned} Q(\sum_{i=1}^n x_i \mathbf{u}_i) &= \left\langle \sum_{i=1}^n x_i \mathbf{u}_i, A(\sum_{i=1}^n x_i \mathbf{u}_i) \right\rangle \\ &= \left\langle \sum_{i=1}^n x_i \mathbf{u}_i, \sum_{i=1}^n x_i \lambda_i \mathbf{u}_i) \right\rangle \\ &= \sum_{i=1}^n \lambda_i x_i^2, \end{aligned}$$

by the orthonormality of the basis. Thus Q is in diagonal form (a sum of squares of the coefficients times some constants). Moreover, an orthonormal basis is a diagonal basis for the usual quadratic form $Q_0(\mathbf{x}) = x_1^2 + \cdots x_n^2$; We say that the two quadratic forms Q_0 (the standard form) and Q (coming from A) can be *simultaneously diagonalized*.

Proof of the spectral theorem. We argue by induction on n, the case $n = 1$ being clear since every vector in that case is an eigenvector. So assume the theorem is true for every symmetric $(n-1) \times (n-1)$ matrix, and let A be a symmetric $n \times n$ matrix. The key part of the argument is to show that A has at least **one** real eigenvalue. Assuming this, let us prove the theorem. If d is a real eigenvalue of A and $\mathbf{u}_n$ is a nonzero eigenvector with eigenvalue d, then after dividing by the length we can assume that $\mathbf{u}_n$ has unit length. Now let B be the bilinear form corresponding to A, i.e. $B(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, A\mathbf{y} \rangle$. We can define $\{\mathbf{u}_n\}^{\perp_B}$ as in the proof of Sylvester's theorem:

$$\{\mathbf{u}_n\}^{\perp_B} = \{\mathbf{v} \in \mathbb{R}^n : B(\mathbf{u}_n, \mathbf{v}) = 0\}.$$

We also have the usual perpendicular space

$$\{\mathbf{u}_n\}^\perp = \{\mathbf{v} \in \mathbb{R}^n : \langle \mathbf{u}_n, \mathbf{v}\rangle = 0\}.$$

Note that

$$B(\mathbf{u}_n, \mathbf{v}) = B(\mathbf{v}, \mathbf{u}_n) = \langle \mathbf{v}, A\mathbf{u}_n\rangle = \langle \mathbf{v}, d\mathbf{u}_n\rangle = d\langle \mathbf{v}, \mathbf{u}_n\rangle.$$

Hence, if $\langle \mathbf{v}, \mathbf{u}_n\rangle = 0$, then $B(\mathbf{v}, \mathbf{u}_n) = \langle \mathbf{v}, A\mathbf{u}_n\rangle = 0$ as well, so that $\{\mathbf{u}_n\}^\perp \subseteq \{\mathbf{u}_n\}^{\perp_B}$. Let $V = \{\mathbf{u}_n\}^\perp$, so that $\dim V = n-1$, and let $\mathbf{f}_1, \ldots, \mathbf{f}_{n-1}$ be an orthonormal basis for V. We claim that $A(V) \subseteq V$, and that, in the basis $\mathbf{f}_1, \ldots, \mathbf{f}_{n-1}$, the matrix corresponding to the restriction of A to V is a symmetric matrix. The statement that $A(V) \subseteq V$ is just the statement that $\mathbf{v} \in V = \{\mathbf{u}_n\}^\perp \implies A\mathbf{v} \in \{\mathbf{u}_n\}^\perp$, which we have seen.

Let A_{n-1} be the matrix corresponding to the linear map $V \to V$ induced by A with respect to the orthonormal basis $\mathbf{f}_1, \ldots, \mathbf{f}_{n-1}$. Since $\mathbf{f}_1, \ldots, \mathbf{f}_{n-1}$is an orthonormal basis, the $(i,j)^{\text{th}}$ entry of A_{n-1} is $\langle \mathbf{f}_i, A_{n-1}\mathbf{f}_j\rangle$ in the usual way (see Chapter 4). But since A_{n-1} is just the restriction of A, we have

$$\langle \mathbf{f}_i, A_{n-1}\mathbf{f}_j\rangle = \langle \mathbf{f}_i, A\mathbf{f}_j\rangle = \langle {}^tA\mathbf{f}_i, A\mathbf{f}_j\rangle = \langle A\mathbf{f}_i, A\mathbf{f}_j\rangle = \langle A_{n-1}\mathbf{f}_i, A\mathbf{f}_j\rangle = \langle \mathbf{f}_j, A_{n-1}\mathbf{f}_i\rangle.$$

It follows that A_{n-1} is symmetric in the basis $\mathbf{f}_1, \ldots, \mathbf{f}_{n-1}$. We can now apply induction to see that there is a basis of eigenvectors of V for A_{n-1}, say $\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$. Then $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is an orthonormal basis of $\mathbb{R}^n$ and the $\mathbf{u}_i$ are eigenvectors for A.

So we must show that A has at least one real eigenvalue. By Proposition 6.29, either A has a real eigenvector or there exists a two-dimensional vector subspace of $\mathbb{R}^n$ such that $A(W) \subseteq W$. Let $B = A|W \colon W \to W$ be the linear map from the vector subspace W to itself. We claim that B has a real eigenvector. Since W is spanned by $\mathbf{w}_1, \mathbf{w}_2$, either $\dim W = 1$ or $\dim W = 2$. If $\dim W = 1$, then a nonzero vector of W is an eigenvector of B. Otherwise, let $\mathbf{f}_1, \mathbf{f}_2$ be an orthonormal basis of W. Then since $\langle \mathbf{f}_i, B\mathbf{f}_j\rangle = \langle B\mathbf{f}_i, \mathbf{f}_j\rangle$, the usual arguments show that the matrix for B in the basis $\mathbf{f}_1, \mathbf{f}_2$ is a symmetric matrix $\begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix}$. But now a straightforward calculation (exercise) shows that, in this case, the characteristic polynomial $p_B(t)$ of B has two real roots, or possibly a repeated real root. In any case, $p_B(t)$ has a real root and hence B has a real eigenvalue. Thus, A has a real eigenvalue as well. □

Corollary 6.34. *Let A be a symmetric $n \times n$ matrix. Then there exists an orthogonal matrix C, i.e. a matrix C such that ${}^tC = C^{-1}$, and a diagonal matrix D such that $A = {}^tCDC = C^{-1}DC$.*

Proof. If C is the change of basis matrix for the orthonormal basis given in the statement of the spectral theorem, then, by the change of basis formula for quadratic forms, $A = {}^tCDC$ for some diagonal matrix D. But since the columns of C are orthonormal, ${}^tC = C^{-1}$ (exercise). Thus $A = {}^tCDC = C^{-1}DC$. □

Let B be any $n \times n$ matrix and let $A = {}^tB \cdot B$. Then A is symmetric, because

$$ {}^tA = {}^t({}^tB \cdot B) = {}^tB \cdot {}^t({}^tB) = {}^tB \cdot B = A.$$

Moreover, if Q is the quadratic form corresponding to A, so that $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$, then

$$Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, {}^tB \cdot B\mathbf{x} \rangle = \langle B\mathbf{x}, B\mathbf{x} \rangle = \|B\mathbf{x}\|^2.$$

Thus Q is positive semi-definite. The next result says that in fact every positive semi-definite quadratic form arises in this way.

Corollary 6.35. *Let $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$ be a quadratic form. Then the following are equivalent:*

1. *There exists an $n \times n$ matrix B such that ${}^tB \cdot B = A$.*
2. *The form Q is positive semi-definite.*
3. *Every eigenvalue of A is non-negative.*
4. *There exists an $n \times n$ symmetric matrix B such that $B^2 = A$.*

Proof. We have seen that (1) $\implies$ (2) above.
(2) $\implies$ (3): Suppose that (2) holds, i.e. that Q is positive semi-definite. If λ is an eigenvalue of A, then there exists a $\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{v} = \lambda\mathbf{v}$. But

$$Q(\mathbf{v}) = \langle \mathbf{v}, A\mathbf{v} \rangle = \langle \mathbf{v}, \lambda\mathbf{v} \rangle = \lambda\|\mathbf{v}\|^2,$$

and since $Q(\mathbf{v}) \geq 0$ and $\|\mathbf{v}\|^2 > 0$, we must have $\lambda \geq 0$.
(3) $\implies$ (4): By Corollary 6.34, $A = C^{-1}DC$ for an orthogonal matrix C and a diagonal matrix D, whose diagonal entries $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. By the assumption of (3), $\lambda_i \geq 0$. Let D_1 be the diagonal matrix whose entries are $\sqrt{\lambda_i}$, $i = 1, \ldots, n$. Then clearly $D_1 \cdot D_1 = D$, and so, if $B = C^{-1}D_1C$, then

$$\begin{aligned} B^2 &= (C^{-1}D_1C)(C^{-1}D_1C) = C^{-1}D_1(CC^{-1})D_1C \\ &= C^{-1}D_1(\mathrm{Id})D_1C = C^{-1}D_1^2C = C^{-1}DC = A. \end{aligned}$$

Furthermore, B is symmetric since

$$ {}^tB = {}^t(C^{-1}D_1C) = {}^tC{}^tD_1{}^t(C^{-1}) = C^{-1}D_1C, $$

as C is orthogonal (${}^tC = C^{-1}$ and hence ${}^t(C^{-1}) = {}^{tt}C = C$) and D_1 is diagional, hence symmetric.

(4) $\Longrightarrow$ (1): Taking B to be an $n \times n$ symmetric matrix such that $B^2 = A$, we have ${}^tB = B$ and hence ${}^tB \cdot B = B^2 = A$. $\square$

Appendix A

Sets and functions

The language of sets and functions pervades mathematics, and most of the important operations in mathematics turn out to be functions or to be expressible in terms of functions. We will not define what a set is, but take as a basic term the idea of a set X and of membership $x \in X$ (x is an element of X). Two sets X and Y are by definition equal if they have the same elements: $X = Y \iff$ the following holds: $x \in X \iff x \in Y$. If, for every $x \in X, x \in Y$, then X is a *subset* of Y, written $X \subseteq Y$. Note that we always have $X \subseteq X$ and $\emptyset \subseteq X$. A subset A of X is called a *proper* subset if $A \neq \emptyset$ and $A \neq X$. By the definition of equality of sets, $X = Y \iff X \subseteq Y$ and $Y \subseteq X$. If $X \subseteq Y$ and $Y \subseteq Z$, then $X \subseteq Z$; this is called the *transitivity property*. The notation $X \subset Y$ is sometimes used to mean that $X \subseteq Y$ but $X \neq Y$.

Recall the standard operations on sets:

Definition A.1. If X_1 and X_2 are two sets, then:

1. The *union* of X_1 and X_2 is the set

$$X_1 \cup X_2 = \{x : x \in X_1 \text{ or } x \in X_2\}.$$

   Thus $X_1 \subseteq (X_1 \cup X_2)$ and $X_2 \subseteq (X_1 \cup X_2)$.

2. The *intersection* of X_1 and X_2 is:

$$X_1 \cap X_2 = \{x : x \in X_1 \text{ and } x \in X_2\}.$$

   Thus $(X_1 \cap X_2) \subseteq X_1$ and $(X_1 \cap X_2) \subseteq X_2$.

3. If $A \subseteq X$, the *complement* of A in X, written $X - A$, is the set

$$\{x \in X : x \notin A\}.$$

Thus $A \cap (X - A) = \emptyset$ and $A \cup (X - A) = X$. For example, $X - X = \emptyset$ and $X - \emptyset = X$.

By logic (deMorgan's laws), $X - (A_1 \cap A_2) = (X - A_1) \cup (X - A_2)$ and $X - (A_1 \cup A_2) = (X - A_1) \cap (X - A_2)$. (For example, if $x \notin A_1 \cap A_2$, then either $x \notin A_1$ or $x \notin A_2$ and conversely.)

The set of all subsets of X is also a set, and is called the *power set* of X, often denoted $\mathcal{P}(X)$:

$$\mathcal{P}(X) = \{A : A \subseteq X\}.$$

By the transitivity property, if $Y \subseteq X$, then $\mathcal{P}(Y)$ is a subset of $\mathcal{P}(X)$.

Definition A.2. Given X and Y, we define $X \times Y$, the *Cartesian product* of X and Y, to be the set of ordered pairs (x, y) with $x \in X$ and $y \in Y$. Here x is the *first component* of the ordered pair (x, y) and y is the *second component*. Clearly, if $A \subseteq X$ and $B \subseteq Y$, then $A \times B \subseteq X \times Y$.

If $X = Y$, we abbreviate $X \times X$ by X^2. Likewise, if we have n sets $X_1, \ldots, X_n$, then $X_1 \times \cdots \times X_n$ is the set of ordered n-tuples $(x_1, \ldots, x_n)$ with $x_i \in X_i$ for every i, the i^{th} *component* of $(x_1, \ldots, x_n)$ is x_i, and we again abbreviate $X \times \cdots \times X$ (n times) by X^n.

Remark A.3. The operative properties of an ordered pair (x, y) are: 1) For all $x \in X$ and $y \in Y$, there exists an ordered pair $(x, y) \in X \times Y$, and 2) two ordered pairs (x_1, y_2) and (x_2, y_2) are equal $\iff$ they have the same first components and the same second components, i.e. $\iff$ $x_1 = x_2$ and $y_1 = y_2$; it is not enough to require that the sets $\{x_1, y_1\}$ and $\{x_2, y_2\}$ be equal. It is possible to give a formal definition of an ordered pair just using set theory. In fact, one can define $(x, y) = \{\{x\}, \{x, y\}\}$. (In other words, ordered pair does not have to be an undefined term.) However, we shall not really care what the precise definition is, but only that an ordered pair has the operative properties 1) and 2) above. Using functions, though, we can give a careful definition of an ordered n-tuple; we shall describe this later.

Next we define a function $f\colon X \to Y$. Although we can think of a function as a "rule" which assigns to every $x \in X$ a unique $y \in Y$, it is easier to define this precisely by identifying the function f with its graph in $X \times Y$. Thus a function f is the same thing as a subset G_f of $X \times Y$ with

the following property: for all $x \in X$, there is a unique $y \in Y$ such that $(x, y) \in G_f$ and we set $y = f(x)$. To say that there is a unique $y \in Y$ says that $f(x)$ is uniquely determined by x, and to say that for every $x \in X$ there exists an $(x, y) \in G_f$ says that in fact $f(x)$ is defined for all $x \in X$. (This is in contrast to the practice in some calculus courses where f is allowed to be not everywhere defined.) Two functions f_1 and f_2 are equal if and only if their graphs are equal, if and only if, for all $x \in X$, $f_1(x) = f_2(x)$. (Thus a function is specified by its values.) We shall use the word *map* or *mapping* as a synonym for function; typically maps are functions in some kind of geometric setting.

Definition A.4. In the above notation, we call X the *domain* of f and Y the *range*. The set

$$\{y \in Y : \text{ there exists } x \in X \text{ such that } f(x) = y\}$$

is called the *image* of f and is sometimes written $f(X)$. More generally, if A is a subset of X, then we set

$$f(A) = \{y \in Y : \text{ there exists an } x \in A \text{ such that } f(x) = y\}.$$

We say that f is *onto* or *surjective* if $f(X) = Y$, in other words if the image of f is Y. Thus the domain and range are a part of the information of a function. Note that a function *must be defined at all points of its domain*; thus for example the function $f(x) = 1/x$ cannot have domain $\mathbb{R}$ without assigning some value to $f(0)$. Also, in general the image of a function $f(x)$ will be a subset of the range, but *need not* equal the range.

We also define, for B a subset of Y, the subset

$$f^{-1}(B) = \{x \in X : f(x) \in B\}$$

of X, called the *preimage* of B. If $B = \{y\}$ has just one point we write $f^{-1}(y)$ instead of $f^{-1}(\{y\})$. For example, $f^{-1}(Y) = X$ and $f^{-1}(y) \neq \emptyset$ if and only if $y \in f(X)$.

Note: if Y is a subset of another set Y', then a function $f\colon X \to Y$ defines (in an obvious way) a function from X to Y'. Technically these are two *different* functions, although we will occasionally (and incorrectly) blur the distinction. Also, given a function $f\colon X \to Y$, we can always replace it by a function from X to $f(X) \subseteq Y$, and this new function is always onto.

To be careful, we make the following definition:

Definition A.5. If $f\colon X \to Y$ is a function and $A \subseteq X$, then we define the *restriction* $(f|A)$ of f to A to be the function $(f|A)\colon A \to Y$ defined by $(A \times Y) \cap G_f$, where G_f is the graph of f. In other words, $(f|A)(a) = f(a)$ for all $a \in A$, and the domain of $(f|A)$ is exactly A. If moreover $f(A) \subseteq B$, there is the *induced* function $g\colon A \to B$, which technically is different from $(f|A)$. However we shall sometimes be a little careless.

Here are some basic examples of functions:

1. For any set X, the *identity function* $\mathrm{Id}\colon X \to X$ satisfies: $\mathrm{Id}_X(x) = x$ for every $x \in X$. Thus its graph in X^2 is the set $\{(x, x) : x \in X\}$, which we can think of as the "diagonal" viewed as a subset of X^2. (Does the diagonal satisfy the test of being the graph of a function?) The preimage of $A \subseteq X$ is just A.

2. A related example is inclusion: if $X \subseteq Y$, then $\{(x, x) : x \in X\}$ is a subset of $X \times Y$ which is the graph of the inclusion function from X to Y. Recall that technically this is not the same as the identity function.) The preimage of $B \subseteq Y$ is then $B \cap X$.

3. Another example, if $Y \neq \emptyset$, is a *constant function*: choose $c \in Y$ and define $f(x) = c$ for all $x \in X$. (What is the graph of this function and why is it a function?) In this case, the preimage of a subset B of Y is $\emptyset$ if $c \notin B$ and is X if $c \in B$.

4. Of course, all of the standard functions of calculus give examples of functions from $\mathbb{R}$ to $\mathbb{R}$. (If $f\colon \mathbb{R} \to \mathbb{R}$ is the function $f(x) = x^2$, what is the image of f? What is $f^{-1}(a)$?)

5. Another example is the Cartesian product $X \times X$, which we can identify with the set of functions $f\colon \{1, 2\} \to X$. In fact, in the notation for an ordered pair (x_1, x_2), we can think of x_i as a function which assigns the value x_1 to 1 and x_2 to 2. If we wanted to define the Cartesian product of two possibly different sets in this way, we could define $X \times Y$ to be the set of all functions $f\colon \{1, 2\} \to X \cup Y$ such that $f(1) \in X$ and $f(2) \in Y$. Likewise X^n is the set of functions $f\colon \{1, 2, \ldots, n\} \to X$. A sequence $x_1, x_2, \ldots$ of real numbers is then the same as a function $\mathbb{N} \to \mathbb{R}$.

6. If X and Y are two sets, then the set of all functions from X to Y is a new set, sometimes denoted by Y^X:

$$Y^X = \{f : \ f \text{ is a function from } X \text{ to } Y\}.$$

Then, given $x \in X$, we get a function ev_x from Y^X to Y by evaluating at x:

$$\mathrm{ev}_x(f) = f(x).$$

Thus, when we write $f(x)$ above, the symbol f has become the "variable." There is a similar function of two variables, $e\colon X \times Y^X \to Y$, defined by

$$e(x, f) = f(x).$$

Given functions $f\colon X \to Y$ and $g\colon Y \to Z$, we have the composed function $g \circ f\colon X \to Z$ defined by

$$g \circ f(x) = g(f(x))$$

for all $x \in X$. For example, given $f\colon X \to Y$, $\mathrm{Id}_Y \circ f = f \circ \mathrm{Id}_X = f$. Function composition has the important property that it is associative where defined:

Proposition A.6. *Suppose given functions $f\colon X \to Y$, $g\colon Y \to Z$, and $h\colon Z \to W$. Then*

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

Proof. For all $x \in X$, $(h \circ (g \circ f))(x) = h((g \circ f)(x)) = h(g(f(x)))$, and likewise $((h \circ g) \circ f)(x) = (h \circ g)(f(x)) = h(g(f(x)))$. Thus $(h \circ (g \circ f))(x) = ((h \circ g) \circ f)(x)$ for all $x \in X$ and so $h \circ (g \circ f) = (h \circ g) \circ f$. □

The operation of function composition is somewhat like an algebraic operation, in that we can sometimes "combine" two functions and get a third. But we can't always do so: we can only define $g \circ f$ when the range of f is equal to the domain of g. In general function composition is not commutative. For example, given $f\colon X \to Y$, we can only compose it with $g\colon Y \to Z$ in both orders when $X = Y$. In this case $g \circ f\colon X \to X$ and $f \circ g\colon Y \to Y$, and we can only compare these when $X = Y$. Finally, very simple examples show that the random composition of two functions, whose domain and range are both equal to a fixed set X will depend on the order (for example, $X = \mathbb{R}$, $f(x) = e^x, g(x) = x^2 + 1$).

Note that the identity function behaves very much like an identity element under composition, as long as we are careful about the domains of the relevant identity functions: if $f\colon X \to Y$ is a function, then $f \circ \mathrm{Id}_X = f$ and $\mathrm{Id}_Y \circ f = f$.

Definition A.7. A function $f\colon X \to Y$ is *one-to-one* or *injective* if, for all $x_1, x_2 \in X$, $f(x_1) = f(x_2)$ if and only if $x_1 = x_2$. Equivalently, for all $y \in Y$, the set $f^{-1}(y)$ has at most one element. Thus f is injective if, for all $y \in Y$, the equation $f(x) = y$ has at most one solution, or in other words if a solution exists, then it is unique. By contrast, f is onto if the equation $f(x) = y$ has a solution (not necessarily unique) for every $y \in Y$. A function $f\colon X \to Y$ which is one-to-one and onto is called a *one-to-one correspondence* or a *bijection*.

For example, taking $X = \mathbb{R}$, the function $f(x) = x^2$ is neither one-to-one nor onto. (When is $x_1^2 = x_2^2$? What is the image of f?) The function $f(x) = e^x$ is one-to-one but not onto. (What is the image of f?) The function $f(x) = x^3 + 1$ is a bijection. The identity function $\mathrm{Id}_X\colon X \to X$ is always a bijection.

One-to-one correspondences express the idea that two sets have the same number of elements, and also that two sets might be essentially the same even if technically different. For example, the sets $X \times Y$ and $Y \times X$ are different sets if $X \neq Y$, but there is a natural function $F\colon X \times Y \to Y \times X$ defined by $F(x, y) = (y, x)$. This function is a one-to-one correspondence: if $F(x_1, y_1) = F(x_2, y_2)$, then by definition $(y_1, x_1) = (y_2, x_2)$ as ordered pairs in $Y \times X$. Hence by the operative property of equality of ordered pairs, $y_1 = y_2$ and $x_1 = x_2$, and thus $(x_1, y_1) = (x_2, y_2)$. Hence F is injective. To see that it is surjective, let (y, x) be an arbitrary element of $Y \times X$. Then $(y, x) = F(x, y)$. Thus F is surjective and hence a bijection. Likewise, there is a one-to-one correspondence from $X_1 \times (X_2 \times X_3)$ to $(X_1 \times X_2) \times X_3$, and from either of these sets to $X_1 \times X_2 \times X_3$. In fact, these one-to-one correspondences are so obvious that most of the time we don't bother to make them explicit unless it is absolutely necessary. On the other hand, some one-to-one correspondences are very non-obvious. For example, one can show that there is a one-to-one correspondence from $\mathbb{R}$ to $\mathbb{R}^2$, or in fact to $\mathbb{R}^n$ for any $n > 0$, but such a one-to-one correspondence has no geometric properties. Its existence says that $\mathbb{R}$ and $\mathbb{R}^2$ have the same number of elements from a purely quantitative point of view, but in no other real sense do $\mathbb{R}$ and $\mathbb{R}^2$ resemble each other.

One-to-one correspondences are related to inverse functions. Let $f\colon X \to Y$ be a function. An *inverse function* $g\colon Y \to X$ is a function g such that $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$. As we will show soon, if an inverse function exists it is unique and is denoted f^{-1}. This should not be confused with the preimage which can be defined for any function, and it should never be confused with $1/f$, which can only be defined for a real-valued function

which is never zero. Similarly, a *left inverse* for f is a function g such that $g \circ f = \mathrm{Id}_X$, and a *right inverse* for f is a function g such that $f \circ g = \mathrm{Id}_Y$. Note that g is a left inverse for $f \iff f$ is a right inverse for g. It is possible for a function to have a right inverse but not a left inverse, and vice-versa. However, if a function has both a right and a left inverse they are equal:

Proposition A.8. *Suppose that $f\colon X \to Y$ is a function, and that $g\colon Y \to X$ and $h\colon Y \to X$ are functions such that $g \circ f = \mathrm{Id}_X$ and $f \circ h = \mathrm{Id}_Y$. Then $g = h$ and so g is an inverse function for f.*

Proof. Consider $g \circ f \circ h$. Since function composition is associative, this is

$$(g \circ f) \circ h = \mathrm{Id}_X \circ h = h$$

but associating the other way says that it is also equal to

$$g \circ (f \circ h) = g \circ \mathrm{Id}_Y = g.$$

Hence $g = h$. □

Corollary A.9. *If g_1 and g_2 are two inverse functions for f, then $g_1 = g_2$. In other words, an inverse function, if it exists, is unique.*

Proof. Since an inverse function is both a left and a right inverse, we can apply the previous proposition, viewing, say, g_1 as a right inverse and g_2 as a left inverse, to conclude that $g_1 = g_2$. □

The relation between left and right inverses and one-to-one, onto is given by the following:

Proposition A.10. *Let $f\colon X \to Y$ be a function.*

1. *Suppose that $X \neq \emptyset$. Then f has a left inverse if and only if f is one-to-one.*
2. *f has a right inverse if and only if f is onto.*
3. *f has an inverse if and only if f is a one-to-one correspondence.*

Proof. (1) If f has a left inverse g, suppose that $f(x_1) = f(x_2)$. Then $g \circ f(x_1) = g \circ f(x_2)$. But $g \circ f(x) = x$ for all $x \in X$, so that $x_1 = x_2$. It follows that f is one-to-one. Conversely suppose that f is one-to-one. Choose $c \in X$, which is possible since $X \neq \emptyset$. Define $g\colon Y \to X$ as follows.

Given $y \in Y$, if $y = f(x)$ for some $x \in X$, then x is unique and let $g(y) = x$. If there is no x with $f(x) = y$, set $g(x) = c$. Then $g \circ f(x) = x$ for all $x \in X$, so that g is a left inverse for f.

(2) If f has a right inverse h, given $y \in Y$, by definition $f(h(y)) = y$. Hence f is onto. The proof of the converse implication is less straightforward. The idea is the following: for every $f \in Y$, choose some x such that $f(x) = y$, and define $h(y) = x$. The problem here is that it is not immediately clear that there really is a subset of $Y \times X$ which will have the required property, in other words that we can make these kind of unspecified choices in set theory. Granting this however it follows by construction that $f \circ h(y) = y$, so that h is a right inverse.

(3) If f has both a right and a left inverse, it is one-to-one and onto and so a one-to-one correspondence, by (1) and (2). Conversely, if f is one-to-one and onto, then by (1) f has a left inverse and by (2) f has a right inverse, so they are both equal and are an inverse function for f. However, without using (1) and (2), it is easy to see directly that the subset

$$\{(y, x) : \ (x, y) \text{ belongs to the graph of } f\}$$

is the graph of a function $g\colon Y \to X$, such that $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$. Thus g is an inverse function. □

Appendix B

Integers and induction

What are the basic properties of the natural numbers $\mathbb{N}$? First, we need the number 1. Second, given a number $n \in \mathbb{N}$, we can always find a next number which we will write as $s(n)$ and think of as the successor of n. Note that 1 is not the successor of anything. Finally, we exhaust all the natural numbers by taking 1 and all of its successors. We may summarize these properties by saying the following: There is a function $s\colon \mathbb{N} \to \mathbb{N}$ (the *successor function*) with the following properties:

1. The function s is 1–1. In other words, if $s(n) = s(m)$, then $n = m$.
2. There is an element $1 \in \mathbb{N}$, and 1 is not in the image of s, in other words there is no $n \in \mathbb{N}$ with $s(n) = 1$.
3. Let X be a subset of $\mathbb{N}$ with the following properties: $1 \in X$, and, if $n \in X$, then $s(n) \in \mathbb{N}$. Then $X = \mathbb{N}$.

Here (3) is referred to as the *principle of mathematical induction*.

We can make a model for $\mathbb{N}$ as follows: for each natural number n, we shall construct an exemplary set with n elements, and then call that set n. Starting out with $0 = \emptyset$ (which is not in fact an natural number), let $1 = \{\emptyset\} = \{0\}$, let $2 = \{\emptyset, \{\emptyset\}\} = \{0, 1\}$, and in general let $n = \{0, 1, \ldots, n-1\}$. Note the inductive nature of this definition: given all the numbers up through n, we let $s(n) = \{0, \ldots, n\} = n \cup \{n\}$. It follows (confusingly) that $n \subseteq s(n)$. The set $\mathbb{N}$ is then the set of all sets so obtained (we will not worry about whether or not this set exists). We will then say that a set S has n *elements*, written $\#(X) = n$, if there exists a bijection from the set n to X. A set X is *finite* if it has n elements for some $n \in \mathbb{N}$, or $X = \emptyset$ (in which case $\#(X)$ is defined to be 0).

Now we can define the basic operations of addition and order as follows:

Addition. Define $n+1$ (naturally enough) to be $s(n)$. Now we can define addition inductively: suppose that $n+m$ has already been defined. Then define $n+s(m)$ to be $s(n+m)$. Next we claim that addition is associative.

Proposition B.1. *For all $n, m, p \in \mathbb{N}$, $n+(m+p) = (n+m)+p$.*

Proof. For a given n, m, we will prove the statement by induction on p. For $p = 1$, the statement reads

$$n + (m+1) = (n+m) + 1,$$

which we can rewrite as $n + s(m) = s(n+m)$. Note that this is true by the very definition of addition, which if you like has been defined to force associativity at this stage. Now assume that, for some $p \in \mathbb{N}$, we have shown that $n+(m+p) = (n+m)+p$. We must show that $n+(m+(p+1)) = (n+m)+(p+1)$. On the other hand, using repeatedly the associativity when the number on the right is 1, we see that

$$\begin{aligned} n+(m+(p+1)) &= n+((m+p)+1)) \\ &= (n+(m+p))+1 \\ &= ((n+m)+p)+1 \\ &= (n+m)+(p+1), \end{aligned}$$

where the third line uses the inductive hypothesis and the fourth is again the statement when one of the numbers is 1. So we have completed the inductive step, and showed that the statement is true for all p. □

Likewise, in this way we can by induction prove the following (addition is commutative):

Proposition B.2. *For all $n, m \in \mathbb{N}$, $n+m = m+n$.*

Proof. First we claim that $1+m = s(m) = m+1$ for all $m \in \mathbb{N}$. If $m = 1$, then clearly $1+1 = 1+1$. Now suppose that $1+m = s(m) = m+1$. Then $1+s(m) = 1+(m+1) = (1+m)+1 = (m+1)+1 = s(m)+1$, where we have used associativity and the inductive hypothesis. Thus the statement is true for $s(m)$ as well, completing the inductive step. So it holds for all $m \in \mathbb{N}$.

Now fix m and ask if $n+m = m+n$. We will prove this statement by induction as well. We have just showed above that it is true if $n = 1$.

Suppose that it holds for n, and let us prove it for $s(n)$. We have

$$\begin{aligned}(n+1)+m &= n+(1+m) && \text{by associativity,}\\ &= n+(m+1) && \text{by the first part of the proof,}\\ &= (n+m)+1 && \text{by associativity,}\\ &= (m+n)+1 && \text{by the inductive hypothesis,}\\ &= m+(n+1) && \text{again by associativity.}\end{aligned}$$

Thus $(n+1)+m = m+(n+1)$, completing the inductive step. □

Similarly one can show the general cancellation law:

Proposition B.3. *For all $n, m, k \in \mathbb{N}$, if $n+k = m+k$, then $n = m$.*

Some remarks: first, in the above arguments, we don't always need to go back to first principles to prove something. Once we have proved a result, we are free to use it in the proof of another result later on. The only point is to be careful of circular reasoning (using the result that we want to prove). The second comment is that it is hard to prove **everything**! There are dozens of statements about the arithmetic of natural numbers, all proved by using the definitions, induction, and previous results, and it would both overwhelming and boring to have to write down a proof of every such statement.

Multiplication. The natural inductive definition is the following: $n \cdot 1 = n$ for all $n \in \mathbb{N}$, and $n \cdot (m+1) = (n \cdot m)+1$. With this definition multiplication is commutative and associative and distributes over addition.

Exponents. Define $a^1 = a$ and $a^{n+1} = a^n \cdot a$. With this definition we have the usual rules of exponents:

$$\begin{aligned}a^n \cdot a^m &= a^{n+m}\\ (a^n)^m &= a^{nm}.\end{aligned}$$

Proof by induction!

A final inductive definition is factorials: we define $1! = 1$ and

$$(n+1)! = n!(n+1).$$

Order. We define $\leq$ as follows: $n \leq m$ if either $n = m$ or there exists some $p \in \mathbb{N}$ such that $m = n+p$. With our model construction of the natural numbers, we have the following:

Proposition B.4. *For all $n, m \in \mathbb{N}$,$n \leq m$ if and only if $n \subseteq m$.*

Of course, the proof is by induction! For example, informally if $n \leq m$ and $n \neq m$, then $m = s(s(\cdots s(n)\cdots))$, where we apply the successor function p times for some $p \in \mathbb{N}$. But

$$n \subseteq s(n) \subseteq s(s(n)) \cdots \subseteq s(s(\cdots s(n)\cdots)).$$

Of course, a formal proof replaces the ellipsis '$\cdots$' by induction. The other direction, that if $n \subseteq m$ then $n \leq m$, also follows by induction. (But it is a little involved.) From the definition it is easy to see that, if $n \leq m$, then for every $p \in \mathbb{N}$, $n+p \leq m+p$ and $np \leq mp$. More generally, if $n \leq m$ and $a \leq b$ then $n + a \leq m + b$ and $na \leq mb$. There are other easy properties of '$\leq$' which can be established by induction. For example $1 \leq n$ for every $n \in \mathbb{N}$ (this will be a homework problem), and for every $n \in \mathbb{N}$, if $n \leq m \leq n$, then $m = n$. In other words, there is no number strictly between n and $n + 1$. Similarly, if $n \leq m \leq n+1$, then either $m = n$ or $m = n+1$. A fundamental property is:

Proposition B.5 (Trichotomy property)**.** *If $n, m \in \mathbb{N}$, then either $n \leq m$ or $m \leq n$. Moreover if $n \leq m$ and $m \leq n$ then $m = n$.*

Here trichotomy refers to the following: as usual, we define $n < m$ if $n \leq m$ and $n \neq m$ (in other words, $m = n + p$ for some $p \in \mathbb{N}$). The trichotomy property is equivalent to the statement that, given $n, m \in \mathbb{N}$, exactly one of the following three alternatives holds: either $n < m$, $m < n$, or $n = m$.

Proof. Fix m and argue by induction on n. If $n = 1$, then as we have seen above, $1 \leq m$ for every $m \in \mathbb{N}$, so the property holds for n. Now suppose that it is true for a given n that either $n \leq m$ or $m \leq n$. If $m \leq n$, then either $m = n$ or $n = m + p$ for some $p \in \mathbb{N}$. Thus either $n + 1 = m + 1$ and so $m \leq n + 1$, or $n + 1 = m + (p + 1)$ and again $m \leq n + 1$. Suppose on the other hand that $n \leq m$. We may assume that $n \neq m$, since we have already discussed this case above. Then $m = n + p$. If $p = 1$, then $n + 1 = m$. Thus $n + 1 \leq m$. Otherwise, $p = q + 1$ for some $q \in \mathbb{N}$, so that $m = n + p = (n + 1) + q$, and so again $n + 1 \leq m$. □

Aside: Quite generally, a *ordering* on a set X is a subset O of $X \times X$ with the following properties (here we abbreviate $(x, y) \in O$ by $x \leq y$):

1. $x \leq x$;

2. If $x \leq y$ and $y \leq x$, then $x = y$;

3. If $x \leq y$ and $y \leq z$, then $x \leq z$;

4. For all $x, y \in X$, either $x \leq y$ or $y \leq x$.

In general, a *relation* on a set X is a subset of $X \times X$. Thus an order is a special kind of relation. Another example is the graph of a function.

There are natural orderings on $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ (although not on $\mathbb{C}$). What is the graph of the corresponding subset of $\mathbb{R} \times \mathbb{R}$? On the other hand, many relations denoted '$\leq$' are not really orderings. For example, if $f, g \colon \mathbb{R} \to \mathbb{R}$ are real valued functions, define $f \leq g$ if $f(x) \leq g(x)$ for all $x \in \mathbb{R}$. Which of the above properties fails here? Another example is the following: for X a set and $\mathcal{P}(X)$ is the power set of X, we can define a subset O of $\mathcal{P}(X) \times \mathcal{P}(X)$ by: $(A, B) \in O$ if and only if $A \subseteq B$. This is not in general an order (which property above fails to hold?) It is called a *partial order*.

One importance of mathematical induction is that it is related to the following ordering property of $\mathbb{N}$:

Proposition B.6 (Well-Ordering Principle)**.** *Let A be a nonempty subset of $\mathbb{N}$. Then A has a least element, i.e. an element $x \in A$ such that for every $a \in A$, $x \leq a$.*

We say that $\mathbb{N}$ is *well-ordered*. Note that $\mathbb{Q}, \mathbb{R}, \mathbb{Z}$ are **not** well-ordered.

Proof. Let A be a subset of $\mathbb{N}$ which does not have a least element. We shall show that $A = \emptyset$. To do so, let B be the set of all elements n of $\mathbb{N}$ such that, if $k \leq n$, then $k \notin A$. In particular, if $n \in B$ then $n \notin A$. We shall show that $B = \mathbb{N}$ by induction. Thus since every $n \in \mathbb{N}$ lies in B, $n \notin A$, so $A = \emptyset$.

First note that $1 \in B$. For, if $1 \notin B$, then there exists a $k \leq 1$ such that $k \in A$. But the only natural number $k \leq 1$ is 1 itself (why?), so this is equivalent to saying that $1 \in A$. But 1 is a least element of $\mathbb{N}$ and therefore of A, so A would have a least element, contrary to our assumption. Thus $1 \notin A$ and so $1 \in B$.

Now suppose that $n \in B$. Thus for all $k \leq n$, $k \notin A$. We claim that $n+1 \in B$. Otherwise there is some $k \leq n+1$ with $k \in A$. Now if $k \neq n+1$, then $k \leq n$, so by the inductive hypothesis $k \notin A$. Thus $n+1 \in A$. But then $n+1 \in A$ and $k \notin A$ for $k \leq n$, so that $n+1$ is a smallest element of A, contrary to assumption. Thus $n+1 \notin A$. So for all $k \leq n+1$, $k \notin A$. By definition, then, $n+1 \in B$. This completes the inductive step and shows that $B = \mathbb{N}$. $\square$

We also have the strong principle of induction (or complete induction):

Proposition B.7 (Principle of Complete Induction)**.** *Let $A \subseteq \mathbb{N}$, and suppose that:*

1. *$1 \in A$;*
2. *For all $n \in \mathbb{N}$, if $k \in A$ for all $k \leq n$, then $n+1 \in A$.*

Then $A = \mathbb{N}$.

Proof. Suppose that A satisfies (1) and (2) above. Let B be the set of elements of $\mathbb{N}$ not in A. We must show that $B = \emptyset$. Otherwise B has a smallest element. Since $1 \in A$, by (1) above, the smallest element of B is bigger than 1, and thus can be written as $n+1$ for some $n \in \mathbb{N}$. Thus $k \in A$ for all $k < n+1$, or equivalently for all $k \leq n$. But by (2) above it follows that $n+1 \in A$. Thus $n+1$ cannot belong to B, and so $B = \emptyset$. □

As an interesting application of the principle of complete induction, let us prove the following. Recall that a *prime number* n is a number $n \in \mathbb{N}$ such that $n \neq 1$ and such that if $n = ab$, then either a or b is 1.

Theorem B.8. *Every natural number $n > 1$ is a product of primes.*

Here by convention a single prime is a product of (one) prime, namely itself.

Proof. Let A be the set

$$\{n \in \mathbb{N} \mid n = 1 \text{ or } n \text{ is a product of primes}\}.$$

Clearly $1 \in A$. Now suppose that $n \in \mathbb{N}$ and that $k \in A$ for all $k \leq n$. We must show that $n+1 \in \mathbb{N}$. If $n+1$ is prime, then it is a product of primes so we are done. Otherwise we can write $n+1 = ab$, $a, b \in \mathbb{N}$, with $1 < a < n+1$ and $1 < b < n+1$. By the induction hypothesis, a and b are products of primes. Thus, so is $n+1$. This completes the inductive step and thus the proof. □

The harder statement, which we shall not prove here, is that the product of primes above is unique up to order, i.e. every number can be **uniquely** factored.

An important inductive definition is that of the sum operator. Let $f\colon \mathbb{N} \to \mathbb{R}$ be a function (which we can think of as a sequence). We define

$$\sum_{i=1}^{n} f(i)$$

inductively as follows: $\sum_{i=1}^{1} f(i) = f(1)$, and

$$\sum_{i=1}^{n+1} f(i) = \sum_{i=1}^{n} f(i) + f(n+1).$$

It should be clear by now that inductive definitions are a way to avoid writing '$\cdots$' in mathematical expressions. For example,

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n,$$

where we write $f(i) = i$ as shorthand for the identity function. Expressions such as $\sum_{i=5}^{n} f(i)$ are similarly defined. Note that the letter i in the expression plays the role of a "dummy variable:" we could call it anything we want to (except n) and the expression has the same meaning. In fact, a better notation would probably be $\sum_{1}^{n} f$, but putting in some name for the variable is a well-ingrained habit based on describing the function f by corresponding rule.

You are probably familiar with evaluating expressions such as $\sum_{i=1}^{n} i = 1 + 2 + \cdots + n$ by induction. For example, we have the famous formula that

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{1}{2} n(n+1).$$

To prove this formula, note that it is true for $n = 1$. For the inductive step, assume that the formula is true for n. Then

$$\begin{aligned}
\sum_{i=1}^{n+1} i &= \sum_{i=1}^{n} i + (n+1), \\
&= \frac{1}{2} n(n+1) + (n+1), \qquad \text{by the inductive step,} \\
&= \frac{1}{2}(n(n+1) + 2(n+1)), \\
&= \frac{1}{2}(n+2)(n+1) = \frac{1}{2}(n+1)(n+2),
\end{aligned}$$

which is the statement with $n+1$ replacing n throughout. Thus we have completed the inductive step and proved the formula. To see the relation of this method with the principle of mathematical induction as we have stated it above, let

$$X = \{n \in \mathbb{N} : \sum_{i=1}^{n} i = \frac{1}{2}n(n+1)\}.$$

Then the above shows that $1 \in X$ and, if $n \in X$, then $n+1 \in X$. Hence $X = \mathbb{N}$, i.e. the formula $\sum_{i=1}^{n} i = \frac{1}{2}n(n+1)$ holds for all natural numbers.

There is a direct argument for the formula for the sum of the first n integers. Suppose that this sum is denoted S. The trick is to compute S in two different ways:

$$\begin{aligned} S &= 1 + 2 + \cdots + (n-1) + n; \\ &= n + (n-1) + \cdots + 2 + 1. \end{aligned}$$

Adding up these expressions by pairing up the terms above each other, we notice that $2S$ is the sum of n terms all equal to $n+1$. Thus $2S = n(n+1)$ so that $S = \frac{1}{2}n(n+1)$.

Given a function f, we can define a new function $F(n) = \sum_{i=1}^{n} f(i)$. Conversely, given a function $F\colon \mathbb{N} \to \mathbb{R}$, we can define the *difference operator* ΔF by the formula

$$\Delta F(n) = F(n+1) - F(n).$$

Then clearly, if $F(n) = \sum_{i=1}^{n} f(i)$, then $\Delta F(n) = f(n+1)$. Likewise,

$$\begin{aligned} \sum_{i=1}^{n} \Delta F(n) &= (F(2) - F(1)) + (F(3) - F(2)) + \cdots + (F(n+1) - F(n)) \\ &= F(n+1) - F(1). \end{aligned}$$

Of course, a careful argument would be by induction. Thus the operations Δ and $\sum$ are almost inverse to each other.

Appendix C

Equivalence relations

We begin with the problem of constructing the rational numbers. We assume that we have constructed the integers $\mathbb{Z}$. The idea will be to find new numbers $\mathbb{Q}$, the set of *rational numbers* so that the equation $bx = a$ always has a (unique) solution $x \in \mathbb{Q}$ for $a, b \in \mathbb{Z}$ as long as $b \neq 0$. (If $b = 0$, and if the usual rules of arithmetic are to apply, then the equation $bx = a$ has no solution if $a \neq 0$ and every $x \in \mathbb{Q}$ is a solution if $a = 0$.) Note that the integers will be contained in $\mathbb{Q}$, because if we take $a = bn$ for $n \in \mathbb{Z}$ then the unique solution must be n. Also, we will want to be able to add, multiply and compare rational numbers just as we do integers. Note that this process is similar to how we construct the integers from the natural numbers. In fact, the integers are the numbers x we need to allow in order to be able to solve the equation $a + x = b$ where a and b are natural numbers (with no restriction on a and b).

The basic idea is to note that the rational number x which solves $bx = a$ is specified by the ordered pair $(a, b) \in \mathbb{Z} \times (\mathbb{Z} - \{0\})$. We will think of this as the ratio a/b and write it as such. Thus we could try to define a rational number to be such an ordered pair. But if x solves $bx = a$, then it also solves $(kb)x = ka$ for every integer k, so that different ordered pairs (a, b) can define the same rational number a/b. In fact, we know from experience in grade school that a/b and c/d define the same rational number if and only if $ad = bc$. In fact, we can also see this from the viewpoint of solving equations: if x is such that $bx = a$ and $dx = c$, then $ad = d(bx) = b(dx) = bc$. One way to solve this problem is to agree that we shall only look at those pairs (a, b) "in lowest terms," in other words such that $b > 0$ is as small as possible, which happens exactly when a and b have no common factor. But this leads into questions about factoring which we haven't discussed, and it

is more convenient to let (a, b) be any element of $\mathbb{Z} \times (\mathbb{Z} - \{0\})$ and then describe a general mechanism for treating certain such pairs as equal.

Let X be a set. Recall that a *relation* R is just a subset of $X \times X$. Choose some symbol such as $\sim$ and denote by $x \sim y$ the statement that $(x, y) \in R$. There are three important types of relations in mathematics: functions $f\colon X \to X$ (we denote by $y = f(x)$ the condition that $(x, y) \in R$), order (we use $x \leq y$ or $x < y$ for $(x, y) \in R$), and equivalence relations for relations that are "like" equality. These are usually denoted by some special symbol such as $\sim$, $\cong$, or $\equiv$. Here is the formal definition:

Definition C.1. An *equivalence relation* on a set X is a subset $R \subseteq X \times X$ with the following properties: denoting $(x, y) \in R$ by $x \sim y$, we have

1. For all $x \in X$, $x \sim x$. (We say $\sim$ is *reflexive*.)
2. For all $x, y \in X$, if $x \sim y$ then $y \sim x$. (We say $\sim$ is *symmetric*.)
3. For all $x, y, z \in X$, if $x \sim y$ and $y \sim z$ then $x \sim z$. (We say $\sim$ is *transitive*.)

Here (1) says that the diagonal is a subset of R. (2) says that the set R is symmetric about the diagonal (if X is the real numbers, say) and (3) is not so easy to describe geometrically.

Examples. (1) The graph of a function $f\colon X \to X$ is an equivalence relation only if it is the identity, i.e. the graph is the diagonal. (This follows since we must have (x, x) in the graph for every $x \in X$.)

(2) Order is not an equivalence relation on a set with at least two elements: $\leq$ is not symmetric and $<$ is neither reflexive nor symmetric.

(3) The relation x loves y is neither reflexive, symmetric nor transitive.

On the other hand, here are some equivalence relations:

1. Equality.
2. The set $R = X \times X$. Here $x \sim y$ for all $x, y \in X$.
3. The relation of congruence on the set of all plane triangles (or all plane figures); likewise similarity of triangles.
4. Let $\ell_1 = \overrightarrow{\mathbf{p}_1\mathbf{q}_1}$ and $\ell_2 = \overrightarrow{\mathbf{p}_2\mathbf{q}_2}$ be two directed line segments in the plane (i.e. ℓ_i is the line segment starting at $\mathbf{p}_i$ and ending at $\mathbf{q}_i$). Then we can define ℓ_1 and ℓ_2 to be *equivalent* if they have the same magnitude and direction, or equivalently if they define the same vectors; as we

have seen, this is the same as requiring that $\mathbf{q}_1 - \mathbf{p}_1 = \mathbf{q}_2 - \mathbf{p}_2$. It is straightforward to check that this defines an equivalence relation on the set of directed line segments.

5. We can define when two sets A and B have the same number of elements by saying that there is a one-to-one correspondence from A to B. This is an equivalence relation, provided we restrict to a set of sets (we cannot just define this as an equivalence relation on the set of all sets, since this set is too big). For example, we could define this relation on a set such as $\mathcal{P}(\mathbb{R})$, the set of all subsets of the real numbers. The content of this statement is as follows: (1) given a set A, $A \sim A$, i.e. there is a one-to-one correspondence from A to itself (the identity function Id_A); (2) If $A \sim B$, i.e. there is a one-to-one correspondence from A to B, say $f\colon A \to B$, then there is a one-to-one correspondence from B to A, in fact f^{-1} exists by general properties of one-to-one correspondences, and $f^{-1}\colon B \to A$ is a one-to-one correspondence from B to A since f is its inverse; (3) If $A \sim B$ and $B \sim C$, then $A \sim C$. In fact, given a one-to-one correspondence from A to B, say f, and a one-to-one correspondence from B to C, say g, then we have seen that the composition $g \circ f$ is a one-to-one correspondence from A to C.

6. Consider the following equivalence relation on integers: n and m are equivalent (write this as $n \equiv m$) if they are both even or both odd. Another way to say this is to say that $n \equiv m$ if and only if $n - m$ is even, if and only if 2 divides $n - m$.

7. Suppose that X is a set and that $f\colon X \to Y$ is a fixed function from X to some set Y. Define $a \sim b$ if $f(a) = f(b)$. The fact that $\sim$ is an equivalence relation follows from the basic properties of equality on Y. For example, $a \sim a$ just says that $f(a) = f(a)$.

Warning: A relation R is a subset of $X \times X$, but equivalence relations say something about elements of X, **not** ordered pairs of elements of X. The ordered pair part comes in because the relation R is the set of all (x, y) such that $x \sim y$. It is accidental (but confusing) that our original example of an equivalence relation involved a set X which itself happened to be a set of ordered pairs.

Equivalence relations are a way to break up a set X into a union of disjoint subsets. Given an equivalence relation $\sim$ and $a \in X$, define $[a]$, the

equivalence class of a, as follows:

$$[a] = \{x \in X : x \sim a\}.$$

Thus we have $a \in [a]$. Given an equivalence class $[a]$, a *representative* for $[a]$ is an element of $[a]$, in other words it is a $b \in X$ such that $b \sim a$.

For example:

1. If $\sim$ is equality $=$, then $[a] = \{a\}$.
2. If $\sim$ corresponds to $R = X \times X$, in other words $x \sim y$ for all $x, y \in X$, then $[a] = X$ for every $a \in X$.
3. If $\sim$ is congruence $\cong$ of triangles, then the equivalence class of a triangle T is the set of all triangles which are congruent to T.
4. For the equivalence relation on $\mathbb{Z}$, $a \equiv b$ if $a - b$ is divisible by 2, there are two equivalence classes, the set of even integers and the set of odd integers.
5. Given $f : X \to Y$ a function and the equivalence relation $x \sim y$ if $f(x) = f(y)$, the equivalence classes are the sets of preimages $f^{-1}(z)$ for z in the image of f. (Why?)

In the above examples, two equivalence classes which are not equal are disjoint. In fact, this is a general property:

Proposition C.2. *Let $\sim$ be an equivalence relation on a set X, and let $[a]$ be the equivalence class of a. If $[a] \cap [b] \neq \emptyset$, then $[a] = [b]$. Thus a is contained in exactly one equivalence class.*

Proof. Suppose that there is some $c \in [a] \cap [b]$. We will show that $[a] \subseteq [b]$. By symmetry $[b] \subseteq [a]$ (there is nothing special about either one) so that $[a] = [b]$. By definition $c \sim a$ and $c \sim b$. Using symmetry of $\sim$, $a \sim c$ as well. Given $x \in [a]$, by definition $x \sim a$ also. Now $x \sim a$ and $a \sim c$, so by transitivity $x \sim c$. Since $c \sim b$, again by transitivity $x \sim b$. Thus by definition $x \in [b]$. □

For an equivalence relation $\sim$, we have the set of all equivalence classes $\{[a] : a \in X\}$. This set is sometimes written $X/\sim$. It is a subset of the power set $\mathcal{P}(X)$ with the following property: two subsets of X which lie in $X/\sim$ are either equal or disjoint (this is the statement of the proposition above) and every element of X lies in some (hence exactly one) set in $X/\sim$. Put another way: X is a **disjoint union** of the equivalence classes.

One can also define an equivalence relation by reversing this procedure: suppose that X is the union of of disjoint subsets, and define $x \sim y$ if x and y are in the same subset. Then one can check that $\sim$ is an equivalence relation whose equivalence classes are exactly the subsets we started with.

Sometimes equivalence classes can have a "best" representative. For example, for the rational number example below, a good choice of representative is to take (a, b) with $b > 0$ and as small as possible. For the relation on $\mathbb{Z}$, $a \equiv b$ if 2 divides $a - b$, there are two equivalence classes, the even and the odd integers, and good choices are 0 for the equivalence class of even integers and 1 for the equivalence class of odd integers. However, it is often useful to be able to speak of all equivalence classes equally.

We return to the example that comes from trying to construct the rational numbers. Let $\mathbb{Z}$ denote the set of integers, and let $X = \mathbb{Z} \times (\mathbb{Z} - \{0\})$ (the set of ordered pairs (a, b) with $a, b \in \mathbb{Z}$ and $b \neq 0$). Define $(a, b) \sim (c, d)$ if $ad = bc$. Then we have the following:

Proposition C.3. *$\sim$ is an equivalence relation on X.*

Proof. First $(a, b) \sim (a, b)$: this is just the statement $ab = ba$, which holds since multiplication is commutative. Also, if $(a, b) \sim (c, d)$, then by definition $ad = bc$. But then $(c, d) \sim (a, b)$ since $cb = da$. Finally to check transitivity, suppose that $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then by definition $ad = bc$ and $cf = de$. We must show that $af = be$. Start with $ad = bc$ and multiply by f to get

$$adf = bcf = bde.$$

Since $d \neq 0$, we can cancel it to get $af = be$, as claimed. □

With this notation, we can define addition and multiplication of rational numbers as follows: define, for ordered pairs (a, b),

$$(a, b) + (c, d) = (ad + bc, bd)$$
$$(a, b) \cdot (c, d) = (ac, bd).$$

We would like to define this operation on equivalence classes $[(a, b)]$, as follows: define

$$[(a, b)] + [(c, d)] = [(ad + bc, bd)],$$

and similarly for multiplication. The meaning of this is as follows: given two equivalence classes, pick representatives for each and add according to the formula for adding ordered pairs above. Then take the equivalence class

of this sum. The problem is to show that this operation is **well-defined**. In other words, the equivalence class of the sum doesn't depend on which representatives for the equivalence classes you use This amounts to showing for example that if $(a,b) \sim (a',b')$ then $(a,b)+(c,d) \sim (a',b')+(c,d)$. This is a calculation: we have assumed that $ab' = a'b$ and want to show that $(ad+bc, bd) \sim (a'd+b'c, b'd)$. This in turn boils down to checking if

$$(b'd)(ad+bc) = (bd)(a'd+b'c).$$

However the left hand side is $d(b'ad+b'bc) = d(a'bd+b'bc) = (bd)(a'd+b'c)$, so that $(a,b)+(c,d) \sim (a',b')+(c,d)$. An easier calculation shows that, if $(a,b) \sim (a',b')$ then $(a,b)\cdot(c,d) \sim (a',b')\cdot(c,d)$.

We can now define the rational numbers $\mathbb{Q}$ to be the set of all equivalence classes. The integers are contained in the natural numbers if we identify $n \in \mathbb{Z}$ with $[(n,1)]$. Note that $(n,1) \sim (m,1)$ means that $n \cdot 1 = 1 \cdot m$, i.e. $n = m$. Thus each equivalence class can contain at most one integer—another way to say this is to say that the function $f\colon \mathbb{Z} \to \mathbb{Q}$ defined by $f(n) = [(n,1)]$ is 1–1. Also note that $(n,1)+(m,1) = (n+m,1)$ and that $(n,1)\cdot(m,1) = (nm,1)$. So addition of integers is the same as adding them when viewed as rational numbers. Another way to say this is that the function f has the following properties: $f(n+m) = f(n)+f(m)$ and $f(nm) = f(n)\cdot f(m)$.

Addition and multiplication of rational numbers have all the usual properties (commutative, associative, multiplication distributes over addition). The additive identity is $(0,1)$ and the additive inverse to (a,b) is $(-a,b)$. The multiplicative identity is $(1,1)$. Finally, it is easy to see that $(a,b) \sim (0,1)$ if and only if $a = 0$. Thus if $[(a,b)] \neq [(0,1)]$, then (b,a) is an element of $\mathbb{Z} \times (\mathbb{Z} - \{0\})$ and $(a,b)\cdot(b,a) = (ab,ab) \sim (1,1)$. So every nonzero element has a multiplicative inverse. In particular, for $b \neq 0$, $[(1,b)]$ is the multiplicative inverse to b.

Of course, we usually write the equivalence class $[(a,b)]$ as a/b. If we identify the integer a with $[(a,1)]$ as above, then $[(1,b)] = 1/b$ is the multiplicative inverse to b and $a/b = a(1/b)$ in the usual way.

Finally, let us give some equivalence relations connected to linear algebra:

1. Let $A = \mathbb{R}^2 - \{0\}$, and define $\mathbf{v} \sim \mathbf{w}$ if the line through the origin and $\mathbf{v}$ is equal to the line through the origin and $\mathbf{w}$, or in other words if $\text{span}\{\mathbf{v}\} = \text{span}\{\mathbf{w}\}$. Another way to say this is: $\mathbf{v} \sim \mathbf{w}$ if there exists $t \in \mathbb{R}, t \neq 0$ such that $\mathbf{w} = t\mathbf{v}$. One checks easily that this is an equivalence relation. More generally, for $A = \mathbb{R}^n - \{0\}$, we can similarly define $\mathbf{v} \sim \mathbf{w}$ if the line through the origin and $\mathbf{v}$ is equal to

the line through the origin and $\mathbf{w}$. The equivalence class containing $\mathbf{v}$ is the set of all nonzero vectors in the line $\text{span}\{\mathbf{v}\}$, and the set of all equivalence classes is the set of all lines in $\mathbb{R}^n$.

2. For $A = \mathbb{R}^2$, define $\mathbf{v} = (v_1, v_2) \simeq \mathbf{w} = (w_1, w_2)$ if $v_2 = w_2$. This is an equivalence relation: the equivalence class containing (v_1, v_2) is the set of all (x, v_2), in other words the horizontal line passing through $(0, v_2)$. The set of all equivalence classes is the set of all horizontal lines.

   More generally, take L to be any line (through the origin) in $\mathbb{R}^2$, and, for $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$, define $\mathbf{v} \sim \mathbf{w}$ if $\mathbf{v} - \mathbf{w} \in L$. In this case, the equivalence class containing $\mathbf{v}$ is the unique line in $\mathbb{R}^2$, not necessarily through the origin, parallel to L and containing $\mathbf{v}$, and the set of all equivalence classes is the set of all lines in $\mathbb{R}^2$ parallel to L.

   We can generalize this still further: for V a vector subspace in $\mathbb{R}^n$, given $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, define $\mathbf{v} \sim \mathbf{w}$ if $\mathbf{v} - \mathbf{w} \in V$. The equivalence class containing $\mathbf{v}$ is the affine subspace of $\mathbb{R}^n$ parallel to V and containing $\mathbf{v}$. The set of all equivalence classes is the set of all affine subspaces of $\mathbb{R}^n$ parallel to V. It is a nice exercise to identify this set with $V^\perp$.

Appendix D

Construction of the real numbers

Our goal is to give a discussion of the set of real numbers $\mathbb{R}$. Whatever the real numbers are, they should serve to model the physical realities of distance, space and time as we perceive them (hence the term *real*). Implicit in this idea is the idea of measurement and order (one distance is shorter than another; one time comes before another). Thus a number should be defined operationally, as something that can be measured. In particular an infinitesimal number (a number smaller than any given physical quantity) does not have any operational significance, although one might want to define such a quantity mathematically. Similarly a measurement cannot be made to an arbitrarily high degree of accuracy, because of practical necessity and also perhaps for theoretical reasons (e.g. quantum mechanics). Thus there is a limit to the extent to which an approximation is physically meaningful; similarly, one cannot draw a perfectly straight line, say, or an exact right triangle or perfectly round circle. From this perspective, there is no need to introduce real numbers; rational numbers with some large but bounded denominator would work just as well. Moreover, the assumption that space and time are infinitely divisible and form some kind of continuous entity might not be physically justified, and indeed this assumption may not hold over very small scales. Nonetheless, real numbers are an extremely powerful and useful one, even in social or biological sciences where it is clearly not true that we are dealing with a continuous and infinitely divisible quantity. One basic reason is that it is very hard to do mathematical manipulations in a universe that is made up of a very large number of discrete chunks; the assumption of a continuous universe leads to a much simpler model.

Experience and observation tell us that the functions describing motion or change for physical (and even some biological and social) quantities tend to behave in an essentially continuous way. That is, if we have enough observations we can predict the general behavior. This even applies to change that seems sudden or abrupt: if we have enough data on a small enough time scale, we expect this behavior to appear continuous as well. (A similar statement applies to social or biological phenomena they it concerns large enough numbers.) Such an assumption is important in a world where all measurement is approximate: it guarantees that a small error or uncertainty in our measurement will lead to only a small error in the model's predictions. This suggests that a defining property of the real numbers, along with the ordering, is that 1) they can be approximated by known quantities and 2) they have "no holes" which would allow sudden breaks or discontinuities.

There are also many mathematical reasons for introducing real numbers. One example is that we can solve certain equations or define certain numbers using the reals: $\sqrt{2}$, $\sin 1$, π (but not $\sqrt{-1}$). In particular, real numbers arise as solutions to equations or mathematical problems which can be defined via limiting processes, or processes involving successive approximation. Related to this is the fact that functions defined on the reals under certain assumptions (continuity, say) behave much better than functions on the rationals (for example, in finding maximum values or inverse functions). These properties are also related to the fact that the reals have "no holes". Finally, geometrically the real numbers should describe a line, which in turn can be used to measure space or time, and an ideal line would again have no holes (whatever this means).

How then should we define the real numbers? If we think of the real numbers as filling in the gaps between the rational numbers, then we should be able to define the real numbers via the rational numbers. One way is to try to define a real number directly as a "hole" in $\mathbb{Q}$. For example, the hole corresponding to $\sqrt{2}$ can be detected as follows: the two sets

$$I_1 = \{x \in \mathbb{Q} : x^2 \leq 2\} \qquad \text{and} \qquad I_2 = \{x \in \mathbb{Q} : x^2 \geq 2\}$$

divide $\mathbb{Q}$ into two disjoint intervals. What's missing to glue the two intervals together is the point $\sqrt{2}$.

Another way to describe the reals, which is the one we shall pursue here, is the following: if the reals fill in the gaps between the rationals, then one should be able to describe the reals by trying to approximate them by rational numbers. In general there will be many ways to approximate a number, real or rational. For example I could approximate the number 1

by itself. Thus we will allow the exact value to be an approximate value. This usage of approximate differs from the standard English usage, where approximate connotes (if it does not actually mean) that the approximation is close to the object being approximated, but is not equal to it. However it is much more convenient to make a definition allowing an approximation to be actually equal (just as the mathematical use of "or" differs from the usual one). In general every rational number can be approximated by itself. Of course there are many other approximations to 1: .999993, for example, or 33768/33769. In practical applications, an approximation is only as good as an understanding of the error involved. For a rational number, it may seem rather uninteresting in general to try to approximate it, since we already know the exact value. But given an irrational real number, we will not be able to write it down exactly as a rational number and so it is important to understand what we mean by approximating it. For example, 1, 1.4, 1.41, 1.414, are better and better approximations to $\sqrt{2}$. What we seek is a first guess a_1 to $\sqrt{2}$, a second guess a_2 which is hopefully a better guess than a_1, and so on. Since this process never stops we arrive at an a_n for every n. Formally, we make the following definition:

Definition D.1. A *sequence* in a set X is a function from $\mathbb{N}$ to X.

We will for the moment only be concerned with sequences in $\mathbb{Q}$. We denote the function by $\{a_n\}$ or $\{x_n\}$ or any other convenient choice of letter and not by the usual functional notation. Nonetheless, it is important not to confuse the *sequence* $\{a_n\}$, which is a function from $\mathbb{N}$ to X, with its *image*, which by definition is the subset $\{a_n : n \in \mathbb{N}\} \subseteq X$.

Here are some examples of sequences:

1. The constant sequence $x_n = 0$ for all n, or more generally $x_n = c$ for all n for some $c \in \mathbb{Q}$.

2. The sequence $x_n = n$.

3. The sequence $x_n = 1/n$.

4. The sequence $x_n = 0$ if n is odd and 1 if n is even; in closed form this sequence is given by $x_n = 1 + (-1)^n$.

5. The sequence $x_n = \sum_{i=0}^{n} 9/10^n$. Thus $x_1 = .9, x_2 = .99, x_3 = .999, \ldots$.

6. The sequence $a_1 = 1, a_2 = 1.4, a_3 = 1.41, \ldots$ of successive decimal approximations to $\sqrt{2}$.

Let us elaborate on the definition of the sequence $\{a_n\}$ in (6) above. In general, we seek successively longer decimal expansions, which can be done inductively, by making sure that a_n is "correct to n decimal places": define $a_1 = 1.4$. Inductively suppose we are given

$$a_n = 1 + \sum_{j=1}^{n} e_j 10^{-j},$$

where e_j is an integer, $0 \le e_j \le 9$, $a_n^2 \le 2$ and $(a_n + 10^{-n})^2 \ge 2$, (so that if we replace the last digit e_n by $e_n + 1$, at least if $e_n < 9$, then the square becomes larger than 2). Then we claim that there is a unique integer e_{n+1} with

$$\begin{gathered} 0 \le e_{n+1} \le 9; \\ (a_n + e_{n+1} 10^{-(n+1)})^2 \le 2; \\ (a_n + (e_{n+1} + 1) 10^{-(n+1)})^2 \ge 2. \end{gathered}$$

To see this claim, note that

$$a_n^2 = (a_n + 0 \cdot 10^{-(n+1)})^2 \le 2$$

and that

$$(a_n + 10 \cdot 10^{-(n+1)})^2 = (a_n + 10^{-n})^2 \ge 2.$$

So there is a largest integer e_{n+1} (possibly 9) between 0 and 9 for which $(a_n + e_{n+1} 10^{-(n+1)})^2 \le 2$ and then necessarily $(a_n + (e_{n+1}+1) 10^{-(n+1)})^2 \ge 2$. Then we can define $a_{n+1} = 1 + \sum_{j=1}^{n+1} e_j 10^{-j}$, and a_{n+1} satisfies: $a_{n+1}^2 \le 2$ and $(a_{n+1} + 10^{-(n+1)})^2 \ge 2$. Instead of saying that a_n approaches $\sqrt{2}$ (which is a number whose existence and meaning are unknown to the rational numbers), we could ask if a_n^2 approaches 2. We can estimate how close a_n^2 is to 2, without knowing the exact value of a_n: since

$$a_n^2 \le 2 \le (a_n + 10^{-n})^2,$$

the distance from a_n^2 to 2 is at most the distance between a_n^2 and $(a_n + 10^{-n})^2$. This last distance is

$$(a_n + 10^{-n})^2 - a_n^2 = a_n^2 + 2a_n 10^{-n} + 10^{-2n} - a_n^2 = 2a_n 10^{-n} + 10^{-2n}.$$

Now clearly $a_n \le 1.5$ for all n and so

$$|2 - a_n^2| \le 2a_n 10^{-n} + 10^{-2n} \le 3 \cdot 10^{-n} + 10^{-2n} < 4/10^n.$$

There are many other ways to approximate $\sqrt{2}$. For example, a much more efficient way is to "divide and average": start out with a good first guess, say $b_1 = 1$. Now define the sequence b_n inductively: set

$$b_{n+1} = \frac{1}{2}\left(b_n + \frac{2}{b_n}\right).$$

For example, with $b_1 = 1$ we get $b_2 = 1.5$, $b_3 = 1.41666...$, $b_4 = 1.4142158....$ In general $|2 - b_n^2|$ is considerably smaller than $|2 - a_n^2|$.

Our goal now is to make more of this discussion precise. Let $\{x_n\}$ be a sequence of rational numbers. How should we describe when $\{x_n\}$ gets closer and closer to some fixed rational number L? One way is to describe how well we want the numbers x_n to approximate L, for example to say that the difference $|x_n - L|$ is no larger than some fixed number e.g. 10^{-n}. Thus the decimal expansions of x_n and L must agree up to a certain number of places. You can think of this as a sort of challenge: I give you a (small) number, traditionally denoted ϵ, and challenge you to show me that x_n is eventually within distance ϵ of L. Another way to think of this is to think of giving yourself an allowable error, which can be as small as you need it to be. Thus for example for the sequence a_n^2 above, I want to know that a_n^2 is as close to 2 as I want. So if I want $|2 - a_n^2|$ to be smaller than 10^{-6}, it suffices to take n so that $4/10^n \leq 10^{-6}$, i.e. $10^{n-6} \geq 4$. For example, this is satisfied if $n \geq 7$. The precise definition of a limit is as follows:

Definition D.2. Let $\{x_n\}$ be a sequence of rational numbers, and let L be a rational number. We say that *the limit of the sequence* $\{x_n\}$ *is* L (written $\lim_{n\to\infty} x_n = L$) if, for every rational number $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that, for all $n \geq N$, $|x_n - L| < \epsilon$. (We will see later—it is not at all obvious—that a sequence can have at most one limit.) WE also say that $\{x_n\}$ *converges to* L or *has* L *as a limit*. A sequence which converges to some limit is *convergent*; otherwise the sequence is *divergent* or the limit *does not exist*.

For example, if $x_n = c$ for all c is the constant sequence, then $\lim_{n\to\infty} x_n = c$. Here are two more sequences with limit 1:

$$\tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}, \ldots, \tfrac{n}{n+1}, \ldots$$
$$.9, .99, .999, \ldots.$$

The general term of the first sequence is $n/(n+1)$. To see that its limit is 1, we apply the definition: $|n/(n+1) - 1| = 1/(n+1)$. So we must show

that, given $\epsilon > 0$, there exists a positive integer N such that, for all $k \geq N$, $1/(k+1) < \epsilon$. Clearly it is enough to choose N to be any integer $\geq 1/\epsilon$, for then $1/(k+1) \leq 1/(N+1) < 1/N \leq \epsilon$. The proof for the second sequence is similar: here we can take the n^{th} term of the sequence to be $1 - 10^{-n}$, and so $|1 - 10^{-n} - 1| = 10^{-n}$. What we must show here is that, for every $\epsilon > 0$, there exists an N such that, if $k \geq N$, then $10^{-k} < \epsilon$. We can rewrite this last statement as $10^k > 1/\epsilon$, and then it follows from for example the inequality $10^k > k$ if $k \geq 1$ (proof by induction). We will revisit both of the above arguments shortly.

Looking at the constant sequence, we see that there is no requirement that the sequence x_n never attain its limit. We also do not require that the sequence give better and better approximations to the limit, or equivalently that $|x_{n+1} - L| \leq |x_n - L|$ for all n. Thus for example the sequence $1, \frac{1}{2}, 1, \frac{2}{3}, 1, \frac{3}{4}, \ldots$ also has limit 1.

On the other hand, the sequences $\{x_n\}$ defined by $x_n = n$ or by $x_n = 1 + (-1)^n$ do not converge. For example, for $x_n = 1 + (-1)^n$, supppose that $\lim_{n\to\infty} x_n = L$. Now we can choose $\epsilon = 1/2$. The definition says that, for this choice, there would exist an N such that, for all $n \geq N$, $|x_n - L| < 1/2$. But whatever N is, there are choices for $n \geq N$ so that $x_n = 1$ and thus $x_{n+1} = 0$. But then we must have $|1 - L| < 1/2$ and $|0 - L| = |L| < 1/2$, and this is impossible. Similar arguments handle the sequence $x_n = n$. Explicitly, the sequence $\{x_n\}$ does **not** converge to a rational number if, for every $L \in \mathbb{Q}$, there exists an ϵ such that, for every $N \in \mathbb{N}$, there exists an $n \geq N$ such that $|x_n - L| \geq \epsilon$. Of course, the possibility is still open that the sequence is still trying to converge to some **real** number. But in our examples above of $x_n = 1 + (-1)^n$ or $x_n = n$. this will not be the case.

One final general point about the definition of a limit. Since the definition is phrased in terms of all n sufficiently large (this means all $n \geq N$ for some fixed N), in general we are always free to ignore **finitely many** terms of a sequence. For example, if $\{x_n\}$ and $\{y_n\}$ are two sequences which agree except for finitely many terms, i.e. there exists an N such that, for all $n \geq N$, $x_n = y_n$, then $\{x_n\}$ and $\{y_n\}$ either both converge or both diverge. For another example, we can speak of the convergence of $\{x_n\}$ even if x_n is undefined for finitely many n.

To be able to work with the definitions of limits, we need to recall some basic properties of absolute values and of rational numbers. Viewing a rational number a/b as a ratio of two integers $a, b, b \neq 0$, we say that a/b is positive (written $a/b > 0$) if either $a, b > 0$ or $a, b < 0$. This is independent of the choice of a, b used to define the rational number a/b. Likewise $a/b < 0$ if one of a, b is positive and the other is negative. As we have seen, $a/b = 0$

if and only if $a = 0$. With this definition, we have the usual properties of positive and negative numbers: if $r, s > 0$, then $rs > 0$ and $r + s > 0$. If $r, s < 0$, then $rs > 0$ and $r + s < 0$. If $r > 0$ and $s < 0$, then $rs < 0$. We define $r > s$ if $r - s > 0$, and have the usual rules for inequalities (which we will not write down) which generalize the above.

Recall the absolute value function

$$|x| = \begin{cases} x, & \text{if } x \geq 0; \\ -x, & \text{if } x < 0. \end{cases}$$

Note that $|-x| = |x|$. We think of $|x|$ as the basic measure of the distance from x to the origin: thus it is large if x is very big (positive) or very small (negative). Likewise $|x_1 - x_2|$ is the distance from x_1 to x_2. We have the following basic properties of the absolute value:

Proposition D.3. *For all $x, y \in \mathbb{Q}$,*

1. *$|x| \geq 0$, and $|x| = 0$ if and only if $x = 0$.*
2. *$x \leq |x|$, with equality if and only if $x \geq 0$.*
3. *(The triangle inequality.) $|x + y| \leq |x| + |y|$, with equality if and only if x and y have the same sign.*
4. *$||x| - |y|| \leq |x - y|$.*

Proof. (1): If $x \geq 0$, then $|x| = x \geq 0$, with equality only if $x = 0$. If $x < 0$, then $|x| = -x > 0$.

(2): If $x \geq 0$, then $|x| = x$. Otherwise, $x < 0 < |x|$.

(3): If $x, y \geq 0$, then $|x + y| = x + y = |x| + |y|$. If $x, y \leq 0$, then $|x+y| = -(x+y) = -x+(-y) = |x|+|y|$. Otherwise, say $x > 0$ and $y < 0$. Thus $y < |y|$. Now either $x + y \geq 0$ or $y + x \geq 0$. If for example $x + y \geq 0$, then $|x + y| = x + y < x + |y| = |x| + |y|$. The other cases are similar.

(4): Apply the triangle inequality as follows:

$$|x| = |(x - y) + y| \leq |x - y| + |y|.$$

Thus $|x|-|y| \leq |x-y|$. Likewise $|y|-|x| \leq |y-x| = |x-y|$. Since $||x| - |y||$ is either $|x|-|y|$ or $|y|-|x|$, we see that in either case $||x| - |y|| \leq |x-y|$. □

Next we have a basic property of rational numbers (this is sometimes described as the *Archimedean property*.)

Proposition D.4. *If c and A are positive rational numbers, then there exists a positive integer N such that for all $n \geq N$, $nc > A$.*

Here we think of c as possibly very small and A as potentially very large.

Proof. First suppose that $c, A \in \mathbb{N}$. Then since $c \geq 1$, we can take $N = A+1$. In this case, if $n \geq N$, then

$$nc \geq Nc > (A+1) \cdot 1 = A + 1 > A.$$

Next, we have the following lemma:

Lemma D.5. *If $r \in \mathbb{Q}$ and $r > 0$, there exists an $N_0 \in \mathbb{N}$ such that $N_0 r \in \mathbb{N}$. Moreover, $N_0 r \geq r$. Thus, for every positive rational number r, there exists a natural number $M \geq r$.*

Proof. If $r = a/b$ with $a, b > 0$, we can just take $N_0 = b$. Since $N_0 \geq 1$ and $r > 0$, $N_0 r \geq r$, and the last statement follows since we can take $M = N_0 r$. □

Returning to the proof of the proposition, first find $M \geq A$ and find N_0 such that $N_0 c \in \mathbb{N}$. Applying the first part of the proof to the natural numbers $N_0 c$ and M, we see that there exists an $N_1 \in \mathbb{N}$ such that $N_1 N_0 r > M$. Now choose $N = N_0 N_1$. Then if $n \geq N$, we have

$$nc \geq N_0 N_1 c = N_1 N_0 c > M \geq A,$$

as claimed. □

Corollary D.6. *If c is a rational number > 1 and A is a positive rational number, then there exists a positive integer N such that for all $n \geq N$, $c^n > A$.*

Proof. We can write $c = 1 + h$ with $h \in \mathbb{Q}$ and $h > 0$. By induction,

$$c^n = (1+h)^n \geq 1 + nh.$$

(It is clearly true if $n = 0$, and, for the inductive step,

$(1+h)^{n+1} = (1+h)^n(1+h) \geq (1+nh)(1+h) = 1+(n+1)h+nh^2 > 1+(n+1)h.$

The inequality also follows from the binomial theorem.) Now apply the proposition to the numbers h and A, to find an N such that, for all $n \geq N$, $nh > A$. It follows that, for all $n \geq N$,

$$c^n = (1+h)^n \geq 1 + nh > 1 + A > A,$$

and we are done. □

The argument above is easier if $c \geq 2$, for then an easy induction shows that $c^n \geq nc$.

Corollary D.7. *We have* $\lim_{n\to\infty} 1/n = 0$. *If* $0 \leq r < 1$, *then* $\lim_{n\to\infty} r^n = 0$.

Proof. We shall just write out the second statement. We need to show: given $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that, for all $n \geq N$, $|0 - r^n| < \epsilon$. Now $|0 - r^n| = r^n$, and we want to make this smaller than ϵ. Equivalently, we want

$$\frac{1}{r^n} = \left(\frac{1}{r}\right)^n > \frac{1}{\epsilon}.$$

But $1/r > 1$, and we can apply the above corollary, taking $c = 1/r$ and $A = 1/\epsilon$, to find the appropriate N. □

Using properties of absolute values, let us show that a sequence can have at most one limit:

Proposition D.8. *Suppose that* $\{x_n\}$ *is a sequence and* L_1 *and* L_2 *are both limits of the sequence* $\{x_n\}$. *Then* $L_1 = L_2$.

Proof. For $\epsilon > 0$, there exists N_1 such that, if $n \geq N_1$, then $|x_n - L_1| < \epsilon/2$. (We will see the reason for this tricky choice of $\epsilon/2$ in a minute.) Likewise, there exists N_2 such that, if $n \geq N_2$, then $|x_n - L_2| < \epsilon/2$. Now choose n larger than both N_1 and N_2. Then $|x_n - L_1| < \epsilon/2$ and $|x_n - L_2| < \epsilon/2$. Thus

$$|L_1 - L_2| = |L_1 - x_n + x_n - L_2| \leq |L_1 - x_n| + |x_n - L_2| < \epsilon/2 + \epsilon/2 = \epsilon.$$

So we will be done once we prove the following lemma, which says that two numbers cannot be arbitrarily close to one another: □

Lemma D.9. *Let* $L_1, L_2 \in \mathbb{Q}$ *be such that, for every* $\epsilon > 0$, $|L_1 - L_2| < \epsilon$. *Then* $L_1 = L_2$.

Proof. Either $L_1 = L_2$ or $|L_1 - L_2| > 0$. So suppose that $L_1 \neq L_2$. Then we can choose $\epsilon = |L_1 - L_2|$. In this case we would have

$$|L_1 - L_2| < |L_1 - L_2|,$$

which is a contradiction. Thus we must have $L_1 = L_2$. □

Here are some basic properties of limits:

Proposition D.10. *1. If the sequence $\{x_n\}$ has limit L and c is a rational number, then the sequence $\{cx_n\}$ has limit cL.*

2. If the sequence $\{x_n\}$ has limit L and the sequence $\{y_n\}$ has limit M, then the sequence $\{x_n + y_n\}$ has limit $L + M$.

3. If the sequence $\{x_n\}$ has limit L and the sequence $\{y_n\}$ has limit M, then the sequence $\{x_n y_n\}$ has limit LM.

4. Suppose that $x_n \neq 0$ for all n and that the sequence $\{x_n\}$ has limit L with $L \neq 0$. Then the sequence $\{1/x_n\}$ has limit $1/L$.

Proof. (1) If $c = 0$ then $cx_n = 0$ for all n and trivially $0 = 0 \cdot L$ is the limit of this sequence. Otherwise we can divide by c. By definition, given $\epsilon > 0$ there is an N such that, if $n \geq N$, then $|x_n - L| < \epsilon/|c|$. (Note that we have changed the ϵ to which we are applying the definition of a limit! This is another tricky choice of ϵ. The reason for this will become clear in a minute.) Thus

$$|cx_n - cL| = |c||x_n - L| < |c|\epsilon/|c| = \epsilon.$$

It follows that the limit of the sequence $\{cx_n\}$ is cL.

(2) Given $\epsilon > 0$, there is an N_1 such that, for $n \geq N_1$, $|x_n - L| < \epsilon/2$ and likewise an N_2 such that, for $n \geq N_2$, $|y_n - L| < \epsilon/2$. Choosing N to be the larger of N_1, N_2, we see that, if $n \geq N$, then

$$|x_n+y_n-(L+M)| = |(x_n-L)+(y_n-M)| \leq |x_n-L|+|y_n-M| < \epsilon/2+\epsilon/2 = \epsilon.$$

(3) First we claim:

Lemma D.11. *If $\lim_{n\to\infty} x_n = L$, then there exists an N_1 such that, if $n \geq N_1$, then $|x_n| \leq |L|+1$. Hence every convergent sequence is bounded: if $\{x_n\}$ is convergent, then there exists a C such that $|x_n| \leq C$ for all $n \in \mathbb{N}$.*

Proof. The first statement follows by taking $\epsilon = 1$ in the definition of a limit: there exists an N_1 such that, if $n \geq N_1$, then $|x_n - L| < 1$ and hence $|x_n| - |L| \leq |x_n - L| < 1$, so that $|x_n| \leq |L| + 1$. To see the second, choose N_1 as in the first statement and let C be the maximum value of $|x_1|, \ldots, |x_{N_1-1}|, |L| + 1$. Then for all n, either $n < N_1$ and $|x_n| \leq C$ by construction, or $n \geq N_1$ and $|x_n| \leq |L| + 1 \leq C$ as well. □

Returning to the proof of (3), with $\{x_n\}$ and $\{y_n\}$ as in (3), write

$$|x_n y_n - LM| = |x_n y_n - x_n M + x_n M - LM| \leq |x_n||y_n - M| + |M||x_n - L|.$$

Now choose N_1 so that, if $n \geq N_1$, then $|x_n| \leq |L| + 1$. Given $\epsilon > 0$, there exists an N_2 such that, if $n \geq N_2$, then $|y_n - M| < \epsilon/2(|L| + 1)$ and, if $M \neq 0$, an N_3 such that, if $n \geq N_3$, then $|x_n - L| < \epsilon/2|M|$. (If $M = 0$ then we just ignore the term $|M||x_n - L| = 0$.) Then, for n greater than or equal to the maximum value of N_1, N_2, N_3,

$$\begin{aligned} |x_n y_n - LM| &\leq (|L| + 1)|y_n - M| + |M||x_n - L| \\ &< (|L| + 1) \cdot \epsilon/2(|L| + 1) + |M| \cdot \epsilon/2|M| = \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

(4) First we show the following:

Lemma D.12. *If $\lim_{n\to\infty} x_n = L$ and if $L \neq 0$, then there is an N such that, for all $n \geq N$, $|x_n| > |L|/2 > 0$. In other words, the sequence $\{x_n\}$ is eventually bounded away from 0. Moreover, for all $n \geq N$, x_n and L have the same sign.*

Proof. Apply the definition of limit with $\epsilon = |L|/2$: there is an N_1 such that, for all $n \geq N_1$, $|x_n - L| < |L|/2$. Now one consequence of the triangle inequality is that $|L| - |x_n| \leq |L - x_n| = |x_n - L| < |L|/2$. So

$$|x_n| > |L| - |L|/2 = |L|/2.$$

The last statement is clear since if x_n and L have different signs then $|L - x_n| \geq |L|$. □

Returning to the proof of (4), given the sequence sequence $1/x_n$, we have

$$\left|\frac{1}{x_n} - \frac{1}{L}\right| = \frac{|L - x_n|}{|x_n L|}.$$

If we use the above lemma, the denominator is at least $|L| \cdot |L|/2 = L^2/2$. Thus

$$\frac{|L - x_n|}{|x_n L|} < 2|x_n - L|/L^2.$$

So choose N_2 such that, if $n \geq N_2$, then $|x_n - L| < \epsilon L^2/2$. Then for $k \geq N = \max(N_1, N_2)$,

$$\left|\frac{1}{x_n} - \frac{1}{L}\right| = \frac{|L - x_n|}{|x_n L|} < \epsilon,$$

as desired. □

Note that the proof above also shows that, if $\lim_{n\to\infty} x_n = L \neq 0$, then there is an N such that, for all $n \geq N$, $x_n \neq 0$ and hence $1/x_n$ is defined. In other words, if the sequence $\{x_n\}$ converges to a limit other than zero, eventually all of the terms must be nonzero.

Note: the devices we used to make the estimates may have seemed somewhat devious and artificial. But they should be familiar from a laboratory science class, from estimating the error in a product or quotient, say, of two measurements given the errors in each.

Next we define subsequences of a sequence:

Definition D.13. Let $\{x_n\}$ be a sequence. A *subsequence* of $\{x_n\}$ is defined as follows. Let $f\colon \mathbb{N} \to \mathbb{N}$ be a function which satisfies $f(k+1) > f(k)$. Such a function is called *strictly increasing*. As f is a function on the natural numbers, we can also think of f as a sequence of natural numbers $n_1, n_2, n_3, \ldots$. Here we set $n_k = f(k)$ and the n_j are natural numbers such that $n_1 < n_2 < n_3 < \ldots$. By definition a *subsequence* of $\{x_n\}$ is a sequence of the form $\{x_{f(k)}\}$ for a strictly increasing function f. Usually we denote $x_{f(k)}$ instead by the somewhat confusing notation x_{n_k}.

In particular, a subsequence is itself a sequence, whose k^{th} term is denoted by x_{n_k}.

Example D.14. (0) For a fixed natural number A, the function $f(n) = n + A$ is strictly increasing. Given a sequence $\{x_n\}$, the subsequence corresponding to this is the sequence whose terms are $x_{A+1}, x_{A+2}, \ldots$, in other words all the terms in the original sequence after x_A. It is easy to see that the subsequence converges if and only if the original sequence converges.

(1) The functions $f(n) = 2n$ and $f(n) = 2n+1$ are both increasing functions on $\mathbb{N}$. Thus the sequence $(-1)^n$ has two constant subsequences: the constant sequence 1 (from the subsequence $(-1)^{2n}$) and the constant sequence -1 (from the subsequence $(-1)^{2n+1}$). The example of the sequence $x_n = 1 + (-1)^n$ is similar.

(2) For the sequence $x_n = (-1)^n + 1/n$, which begins

$$0, \tfrac{3}{2}, -\tfrac{2}{3}, \tfrac{5}{4}, -\tfrac{4}{5}, \tfrac{7}{6}, -\tfrac{6}{7}, \tfrac{9}{8}, -\tfrac{8}{9}, \ldots,$$

the subsequence $\{x_{2n}\}$ is the sequence $1 + 1/2n = (2n+1)/2n$— this is the n^{th} term of this new subsequence. Likewise the subsequence $\{x_{2n+1}\}$ is the new sequence $-1 + 1/(2n+1) = -2n/(2n+1)$. The first sequence converges to 1 and the second to -1. In particular, it is possible for a subsequence of a sequence to converge even if the original sequence is divergent.

(3) Consider the sequence of rational numbers between 0 and 1, written down in lowest terms, ordered by increasing denominator, and ordered within a fixed denominator by increasing numerator. Thus the sequence begins as follows:

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \frac{1}{7}, \dots$$

This sequence does not converge either. For example it contains the terms $1/n$ which approach 0 as well as the terms $(n-1)/n$ which approach 1. In fact, given any rational number in $[0,1]$, you can extract from this sequence a subsequence of terms which converge to that rational number, and in fact the same is true for all irrational numbers in $[0,1]$ as well: this sequence is trying to converge to all of them at the same time!

Now suppose that $\{x_n\}$ is a *convergent* sequence. Then every subsequence of $\{x_n\}$ converges to the same limit:

Proposition D.15. *Let $\{x_n\}$ be a sequence converging to the limit L. Then every subsequence $\{x_{n_k}\}$ converges to L as well.*

Proof. Given $\epsilon > 0$, there is by definition an N such that, if $k > N$, then $|x_k - L| < \epsilon$. Now given the increasing function n_k, by induction $n_k \geq k$. (Clearly n_1 is some natural number, thus at least 1, and since $n_{k+1} > n_k$, we must have $n_{k+1} \geq n_k + 1$. So if $n_k \geq k$, then $n_{k+1} \geq k+1$.) Hence, for all $k \geq N$, where N is chosen as above, $n_k \geq k \geq N$. Thus $|x_{n_k} - L| < \epsilon$. □

The problem we now want to consider is the following one: we are trying to define real numbers by taking sequences of rational numbers. However, some sequences of rational numbers have a chance of converging to a rational number, and some (such as $x_n = n$) do not. We need to isolate a property of a sequence of rational numbers, which can be formulated just in terms of rational numbers, which says that it is a candidate for a sequence which converges to a real number. One possibility is to say that consecutive terms of $\{x_n\}$ eventually get close together, in other words that $\lim_{n\to\infty}(x_{n+1} - x_n) = 0$. However, this is not strong enough. For example, the sequence

$$1, 1\tfrac{1}{2}, 2, 2\tfrac{1}{4}, 2\tfrac{1}{2}, 2\tfrac{3}{4}, 3, 3\tfrac{1}{8}, 3\tfrac{1}{4}, 3\tfrac{3}{8}, 3\tfrac{1}{2}, \dots$$

is unbounded and so cannot converge, but the difference between two consecutive terms is eventually less than $1/2^n$. A similar example is the standard example of the divergent series $\sum_{n=1}^{\infty} 1/n$: here the relevant sequence is the sequence of partial sums $s_n = \sum_{i=1}^{n} 1/i$, and this sequence diverges. But

$s_{n+1} - s_n = 1/n + 1$ and thus the limit of $s_{n+1} - s_n$ is 0, even though $\{s_n\}$ diverges.

The correct notion in defining a sequence which can converge to a real number is to say that **all** terms which are sufficiently large eventually get close to each other, not just consecutive terms. Such a sequence has a strong "focussing" property. This leads to the following definition:

Definition D.16. A *Cauchy sequence* $\{x_n\}$ is a sequence $\{x_n\}$ such that for every $\epsilon > 0$, there is a positive integer N, such that, if n and m are both $\geq N$, then $|x_n - x_m| < \epsilon$.

Next we show that, if $\{x_n\}$ is a convergent sequence, then $\{x_n\}$ is Cauchy:

Proposition D.17. *Let $\{x_n\}$ be a convergent sequence. Then $\{x_n\}$ is a Cauchy sequence, in other words for every $\epsilon > 0$, there is a positive integer N, such that, if n and m are both $\geq N$, then $|x_n - x_m| < \epsilon$.*

Proof. Let x_n converge to L. By definition, given ϵ there is N such that $|x_k - L| < \epsilon/2$ if $k \geq N$. Now if both n and m are $\geq N$, then

$$\begin{aligned} |x_n - x_m| &= |x_n - L + L - x_m| \leq |x_n - L| + |L - x_m| \\ &< \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

□

Examples of Cauchy sequences, and sequences which are not Cauchy:

1. A convergent sequence is a Cauchy sequence, by the proposition.
2. The sequence $x_n = n$ is not a Cauchy sequence since, for every n and m, if $n \neq m$, then $|x_n - x_m| \geq 1$.
3. The sequence $x_n = (-1)^n$ is not a Cauchy sequence, since for every n there exists an $m \geq n$ (for example $n = n + 1$) with $|x_n - x_m| = 2$.
4. Decimal expansions are Cauchy sequences:

Proposition D.18. *Let d_n be a sequence of integers with $0 \leq d_n \leq 9$. Form the sequence*

$$x_n = \sum_{j=1}^{n} \frac{d_j}{10^j}.$$

Then $\{x_n\}$ is a Cauchy sequence.

Proof. We estimate $|x_n - x_m|$. We may suppose that $n \geq m$ by symmetry. By definition

$$|x_n - x_m| = \sum_{j=m+1}^{n} \frac{d_j}{10^j}.$$

Now $d_j \leq 9$ for all j, so this sum is at most

$$9 \sum_{j=m+1}^{n} 10^{-j} = 9 \cdot 10^{-m} \sum_{j=1}^{n-m} 10^{-j}.$$

On the other hand we have the formula for the sum $r + r^2 + \cdots + r^k$: it is

$$r + r^2 + \cdots + r^k = r\frac{1 - r^k}{1 - r}.$$

In our case $r = 10^{-1}$ and $k = n - m$, so we obtain

$$|x_n - x_m| \leq 9 \cdot 10^{-m} \cdot 10^{-1} \cdot \frac{1 - 10^{-(n-m)}}{1 - 10^{-1}}$$
$$= 9 \cdot 10^{-m} \cdot \frac{1 - 10^{-(n-m)}}{10 - 1} = 10^{-m} \cdot (1 - 10^{-(n-m)}) < 10^{-m}.$$

So, given $\epsilon > 0$, it suffices to take N large enough so that $10^{-N} \leq \epsilon$. Then, if n and m are $\geq N$ and m, say, is $\leq n$, we have $|x_n - x_m| < 10^{-m} \leq 10^{-N} \leq \epsilon$. □

Cauchy sequences have a lot of good properties. For example, we have the following:

Proposition D.19. *Let $\{x_n\}$ be a Cauchy sequence. Then*

1. *$\{x_n\}$ is bounded, i.e. there is a rational number C such that $|x_n| \leq C$ for all n.*
2. *If $\{x_n\}$ has a convergent subsequence $\{x_{n_k}\}$ with limit L, then $\{x_n\}$ also converges to L.*
3. *If $\{y_n\}$ is another sequence such that the sequence $\{y_n - x_n\}$ converges to 0, then $\{y_n\}$ is a Cauchy sequence too.*

Proof. (1) Choosing $\epsilon = 1$, there is an integer N such that, for all $n, m \geq N$, $|x_n - x_m| < 1$. In particular, choosing $m = N$, this says that for all $n \geq N$, $|x_n| < |x_N| + 1$. So choose C to be the maximum of the numbers

$|x_1|, |x_2|, \ldots, |x_{N-1}|, |x_N| + 1$. Then for every n, either $n < N$ in which case $|x_n| \leq C$ by the choice of C or $n \geq N$ in which case $|x_n| < |x_N| + 1 \leq C$.

(2) For every $\epsilon > 0$, there is an N_1 such that, if $k \geq N_1$, then $|x_{n_k} - L| < \epsilon/2$. Likewise, there is an N_2 such that, if $n, m \geq N_2$, then $|x_n - x_m| < \epsilon/2$. Now if N is the larger of N_1, N_2, then for $k \geq N$, we have $n_k \geq k \geq N_2$ and $k \geq N_1$. Thus

$$\begin{aligned} |x_k - L| &= |x_k - x_{n_k} + x_{n_k} - L| \leq |x_k - x_{n_k}| + |x_{n_k} - L| \\ &< \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

Hence $\lim_{n\to\infty} x_n = L$.

(3) Given $\epsilon > 0$, there is an N_1 such that, for all $n, m \geq N_1$, $|x_n - x_m| < \epsilon/3$. Also, there is an N_2 such that, for all $n \geq N_2$, $|x_n - y_n| < \epsilon/3$. Choose N to be the larger of N_1, N_2. Then if $n, m \geq N$

$$\begin{aligned} |y_n - y_m| &= |y_n - x_n + x_n - x_m + x_m - y_m| \\ &\leq |y_n - x_n| + |x_n - x_m| + |x_m - y_m| \\ &< \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon. \end{aligned}$$

□

Cauchy sequences also satisfy the same kinds of rules that convergent sequences do:

Proposition D.20. *Let $\{x_n\}$ and $\{y_n\}$ be Cauchy sequences. Then:*

1. *For every rational number c, the sequence $\{cx_n\}$ is a Cauchy sequence.*
2. *The sequences $\{x_n + y_n\}$ and $\{x_n y_n\}$ are Cauchy sequences.*
3. *If $\{x_n\}$ does not converge to 0, then there is a positive rational number D and a natural number N such that, for all $n \geq N$, $|x_n| \geq D$ and all of the numbers x_n for $n \geq N$ have the same sign. Finally the sequence $\{1/x_n\}$, which is defined as long as $n \geq N$, is again a Cauchy sequence.*
4. *If $\{x_n\}$ does not converge to 0, then the sequence $\{y_n/x_n\}$, which is defined for all n sufficiently large by (3), is again a Cauchy sequence.*

Proof. Most of these are easy reworkings of the corresponding proofs given before for limits, replacing x_k, L by x_n, x_m everywhere and similarly for y_k, M. One place where we must argue more carefully is (3). The contrapositive of the first sentence of (3) is the statement that, if there is no positive rational number D and no natural number N such that, for all $n \geq N$, $|x_n| \geq D$,

then $\{x_n\}$ converges to 0. If no such D and N exist, then, for every $\epsilon > 0$, and for every N, there exists an $n \geq N$ with $|x_n| < \epsilon$. For example, taking $\epsilon = 1$, there exists a natural number n_1 such that $|x_{n_1}| < 1$. Suppose now that we have inductively found natural numbers $n_1 < n_2 < \cdots < n_k$ such that $|x_{n_j}| < 1/j$ for every j with $1 \leq j \leq k$. Applying the above with $\epsilon = 1/(k+1)$ and $N = n_k + 1$, we see that there is an $n_{k+1} > n_k$ such that $|x_{n_{k+1}}| < 1/(k+1)$. In this way we construct a subsequence $\{x_{n_k}\}$ with $|x_{n_k}| < 1/k$ for every k, which says that $\{x_{n_k}\}$ converges to 0. But then by part (2) of the previous proposition, $\{x_n\}$ also converges to 0. This contradicts our assumptions on $\{x_n\}$. Thus a D as asserted in (3) exists. Now let us show that, possibly after taking N larger, we can assume that all of the x_n have the same sign as long as $n \geq N$. Choose N large enough so that $|x_n| \geq D$ and so that $|x_n - x_m| < D$ as long as $n, m \geq N$. If for some pair of natural numbers $n, m \geq N$, x_n and x_m have opposite signs, then $|x_n - x_m| \geq D + D = 2D > D$, which contradicts our choice of N. Thus x_n and x_m have the same sign.

Given the inequality $|x_n| \geq D > 0$ for $n \geq N$, the rest of the argument for (3) is as in the discussion on limits. □

Now we can define real numbers. Let $\mathcal{S}$ be the set of all sequences of rational numbers. Let $\mathcal{R}$ be the set of all Cauchy sequences of rational numbers, so that $\mathcal{R} \subseteq \mathcal{S}$. Thus one element of $\mathcal{R}$ is the sequence $x_n = 3$ for all n. Another is the sequence $x_n = 1 + (-1)^n/n$. Still another is the sequence a_n defined in the previous handout whose terms begin $1, 1.4, 1.41, 1.414, \ldots$. Hence $\mathcal{R}$ contains sequences converging to rational numbers as well as many more sequences, but it does **not** contain the sequence $x_n = n$ for all n or $x_n = -1 + (-1)^n$. Of course, the same number can be described in many different ways by a sequence and we shall have to say when two sequences describe the same number. This is best done by defining a suitable equivalence relation on $\mathcal{R}$.

Definition D.21. Let $\{x_n\}$ and $\{y_n\}$ be two Cauchy sequences. We write $\{x_n\} \sim \{y_n\}$ if the sequence $\{x_n - y_n\}$ converges to 0.

Proposition D.22. *The relation $\sim$ is an equivalence relation on $\mathcal{R}$.*

Proof. We must show that $\sim$ is reflexive, symmetric and transitive. To see that $\{x_n\} \sim \{x_n\}$, we must show that $\lim_{n\to\infty}(x_n - x_n) = 0$, which is clear since $\{x_n - x_n\}$ is the constant sequence 0. If $\{x_n\} \sim \{y_n\}$, i.e. if $\lim_{n\to\infty}(x_n - y_n) = 0$, then we must show that $\{y_n\} \sim \{x_n\}$, i.e. that $\lim_{n\to\infty}(y_n - x_n) = 0$. But this is clear since $\lim_{n\to\infty}(y_n - x_n) =$

$(-1)\lim_{n\to\infty}(x_n - y_n) = 0$. Finally suppose that $\{x_n\} \sim \{y_n\}$ and that $\{y_n\} \sim \{z_n\}$. Thus $\lim_{n\to\infty}(x_n - y_n) = 0$ and $\lim_{n\to\infty}(y_n - z_n) = 0$. Then

$$\begin{aligned}\lim_{n\to\infty}(x_n - z_n) &= \lim_{n\to\infty}[(x_n - y_n) + (y_n - z_n)] \\ &= \lim_{n\to\infty}(x_n - y_n) + \lim_{n\to\infty}(y_n - z_n) = 0 + 0 = 0.\end{aligned}$$

It follows that $\{x_n\} \sim \{z_n\}$. □

Note: the previous proof also shows that $\sim$ is an equivalence relation on the larger set $\mathcal{S}$. But we will not use this fact.

An equivalence class $[\{x_n\}]$ in $\mathcal{R}$ for the equivalence relation $\sim$ consists of all Cauchy sequences $\{y_n\}$ such that $\{x_n - y_n\}$ converges to 0. We can think of this as the set of all sequences which converge to "the same thing" as $\{x_n\}$. We define the set of *real numbers* $\mathbb{R}$ to be the set of equivalence classes in $\mathcal{R}$ for the relation $\sim$. We can view the rational numbers $\mathbb{Q}$ as a subset of $\mathbb{R}$ by associating to each rational number c the constant sequence $x_n = c$ for all n. This map is one-to-one (why?) but far from being onto.

Given two real numbers x and y, we can add, subtract, multiply and divide them (except by 0), as well as compare them, and all of the usual properties that we expect to hold will in fact hold. We will not try and prove all of this from scratch but simply say a few words about how these operations are defined. For example, suppose that $[\{x_n\}]$ and $[\{y_n\}]$ are two equivalence classes. We define $[\{x_n\}]+[\{y_n\}] = [\{x_n+y_n\}]$ and $[\{x_n\}][\{y_n\}] = [\{x_n y_n\}]$. In other words, given the classes $[\{x_n\}]$ and $[\{y_n\}]$, in order to define their sum we pick any sequence $\{x_n\}$ whose equivalence class is $[\{x_n\}]$, say, and likewise for $[\{y_n\}]$, and then add the sequences to get a new sequence $\{x_n + y_n\}$. The last proposition implies that this new sequence is again a Cauchy sequence and so its equivalence class defines an element of $\mathbb{R}$. We must then check that the equivalence class you get doesn't depend on the specific choice of $\{x_n\}$ and $\{y_n\}$. In other words, if $\{x_n\} \sim \{x'_n\}$, then $\{x_n + y_n\} \sim \{x'_n + y_n\}$. It then follows that, if $\{y_n\} \sim \{y'_n\}$, then $\{x_n + y_n\} \sim \{x'_n + y_n\} \sim \{x'_n + y'_n\}$. So it is enough to show a result of the following type:

Lemma D.23. *Let $\{x_n\}$ and $\{y_n\}$ be two Cauchy sequences.*

1. *If $\{x_n\} \sim \{x'_n\}$, then $\{x_n + y_n\} \sim \{x'_n + y_n\}$.*
2. *If $\{x_n\} \sim \{x'_n\}$, then $\{x_n y_n\} \sim \{x'_n y_n\}$.*
3. *If $\{x_n\} \sim \{x'_n\}$ and $\{x_n\}$ does not converge to 0, then $\{x'_n\}$ does not converge to 0 either and $\{1/x_n\} \sim \{1/x'_n\}$, with the understanding that these sequences are defined for all n sufficiently large.*

Proof. (1) If $\lim_{n\to\infty}(x_n - x'_n) = 0$, then

$$\lim_{n\to\infty}\left[x_n + y_n - (x'_n + y_n)\right] = \lim_{n\to\infty}(x_n - x'_n) = 0.$$

(2) We need to show that $x_n y_n - x'_n y_n$ converges to 0, i.e. that for every $\epsilon > 0$, there exists an N such that, for all $n \geq N$, $|x_n y_n - x'_n y_n| < \epsilon$. We have $|x_n y_n - x'_n y_n| = |y_n||x_n - x'_n|$. Since $\{y_n\}$ is a Cauchy sequence, there is a $C > 0$ such that $|y_n| \leq C$ for all n. As $x_n - x'_n$ converges to 0, given $\epsilon > 0$, there is an N such that for all $n \geq N$, $|x_n - x'_n| < \epsilon/C$. Thus

$$|x_n y_n - x'_n y_n| = |y_n||x_n - x'_n| \leq C(\epsilon/C) = \epsilon.$$

(3) Use the fact that there exist positive rational numbers D and D' such that, for all n sufficiently large, $|x_n| \geq D$ and $|x'_n| \geq D'$. Thus, as in the limits proof,

$$\left|\frac{1}{x_n} - \frac{1}{x'_n}\right| = \frac{|x'_n - x_n|}{|x_n x'_n|} \leq \frac{|x'_n - x_n|}{DD'}.$$

Given $\epsilon > 0$, the argument then follows as in (2), by finding an N such that, if $n \geq N$, then $|x'_n - x_n| < DD'\epsilon$. □

Likewise we can compare two real numbers. It is enough to define the set P of positive real numbers and show that it has the usual properties:

1. For every real number r, exactly one of the following holds: $r \in P$, $-r \in P$, or $r = 0$;

2. the sum and product of two positive numbers is positive;

3. If $-r, -s \in P$, then $-(r + s) \in P$ and $rs \in P$.

Then we can define $r > 0$ by saying that $r \in P$, and $r > s$ by saying that $r - s \in P$.

To define the set P, we say that a real number $[\{x_n\}]$ is positive, i.e. is in P, if, choosing a sequence $\{x_n\}$ in the equivalence class $[\{x_n\}]$, there is a positive rational number D and a natural number N such that, for all $n \geq N$, $x_n \geq D$. It is easy to see that, if $\{x_n\} \sim \{x'_n\}$ and $\{x_n\}$ is positive, then $\{x'_n\}$ is positive as well. To see this, suppose that D is as in the definition and that $x_n \sim x'_n$. Thus there is an N_0 such that, if $n \geq N_0$, then $|x_n - x'_n| < D/2$. Thus $x'_n \geq D - D/2 > 0$. Moreover, every real number r satisfies: either r is positive, or $-r$ is positive, or $r = 0$, by (3) of Proposition D.20. Absolute values are defined in the usual way. In

fact, if $\{x_n\}$ is a Cauchy sequence defining the real number r, then it is an exercise that $\{|x_n|\}$ is also a Cauchy sequence, and in fact $|r| = [\{|x_n|\}]$. With these definitions, the real numbers have all of the familiar properties: you can add, subtract, multiply, and divide real numbers (except by 0), addition and multiplication are associative and commutative, multiplication distributes over addition, and the rational numbers 0 and 1 (viewed as real numbers by using constant sequences) are the additive and multiplicative identities, respectively. The sum and product of two positive numbers is positive, and given a real number r, either r is positive, $-r$ is positive, or $r = 0$.

We have the following:

Lemma D.24. *Suppose that $\{x_n\}$ is a Cauchy sequence and that $\{x_n\} \geq 0$ for all n sufficiently large. Then the corresponding real number $[\{x_n\}] \geq 0$.*

Proof. If $\{x_n\}$ converges to 0 this is clear. Otherwise there is a $D > 0$ and an $N \in \mathbb{N}$ such that either $x_n \geq D$ for all $n \geq N$ or $x_n \leq -D < 0$ for all $n \geq N$. As we have assumed that $x_n \geq 0$ for all large n, we must have $x_n \geq D > 0$ for all $n \geq N$. By definition then $[\{x_n\}]$ is either 0 or positive, i.e. $[\{x_n\}] \geq 0$. □

Note that, even if $x_n > 0$ for all n, it is still possible that $[\{x_n\}] = 0$.

Corollary D.25. *Suppose that $D \in \mathbb{Q}$ and that $\{x_n\}$ is a Cauchy sequence and that $\{x_n\} \geq D$ for all n sufficiently large. Then the corresponding real number $[\{x_n\}] \geq D$.*

Proof. Apply the above to the Cauchy sequence $\{x_n - D\}$. □

Corollary D.26. *Suppose that $r \in \mathbb{R}$ is positive. Then there exists a positive rational number D such that $r \geq D > 0$.*

Proof. Write $r = [\{x_n\}]$; we know that there exists a $D > 0$ such that $\{x_n\} \geq D$ for all n sufficiently large. Thus, by the previous corollary, $r \geq D$. □

There is also the following, which is another way of saying that real numbers can be approximated by rational numbers:

Lemma D.27. *For every real number r and $\epsilon > 0$, there exists a rational number c such that $|c - r| < \epsilon$. In fact, if $r = [\{x_n\}]$ is the equivalence class of the Cauchy sequence $\{x_n\}$, where the $x_n \in \mathbb{Q}$, then we can take $c = x_k$ for k sufficiently large.*

Proof. With $r = [\{x_n\}]$ as in the statement of the lemma, there exists $N \in \mathbb{N}$ such that, for all $n, k \geq N$, $|x_n - x_k| < \epsilon/2$. Choose one such rational number x_k. Setting $c = x_k$, we obtain $|c - x_n| < \epsilon/2$ for all $n \geq N$ and therefore $\epsilon - |c - x_n| \geq \epsilon/2 > 0$. It follows that $\{\epsilon - |c - x_n|\}$, which is a Cauchy sequence, defines the positive real number $\epsilon - |c - r|$, i.e. $\epsilon - |c - r| > 0$. Hence $|c - r| < \epsilon$. □

Another way that the real numbers resemble the rational numbers is that $\mathbb{R}$ is Archimedean:

Proposition D.28. *Let $r, A \in \mathbb{R}$ be positive. Then there exists an $N \in \mathbb{N}$ such that, for all $n \geq N$, $nr > A$.*

Proof. To say that $nr > A$ is equivalent to saying that $n > A/r$, since r is positive. We know by the previous lemma that there exists a $B \in \mathbb{Q}$ with $|A/r - B| < 1$, and hence $A/r < B + 1$. But since $\mathbb{Q}$ is Archimedean, we cn find an N such that, if $n \geq N$, then $n \geq B + 1 > A/r$. Thus $nr > A$, as desired. □

We can now define sequences of real numbers, limits, subsequences, and so on. In particular, every sequence of rational numbers is also a sequence of real numbers after identifying a rational number a with the constant sequence $x_n = a$ for all n.

Suppose that $\{x_n\}$ is a Cauchy sequence of rational numbers. Thus it defines a real number $r = [\{x_n\}]$. We claim that the sequence of rational numbers $\{x_n\}$, viewed as a sequence of real numbers, actually converges to the real number r:

Lemma D.29. *Let $\{x_n\}$ be a Cauchy sequence of rational numbers and let $r = [\{x_n\}]$. Then $\{x_n\}$ converges to r. In other words, for every $\epsilon > 0$, there is an N such that, for all $n \geq N$, $|x_n - r| < \epsilon$.*

Proof. There exists $N \in \mathbb{N}$ such that, for all $n, m \geq N$, $|x_n - x_m| < \epsilon/2$. As in the proof of Lemma D.27, it follows that, if $n \geq N$, then $|x_n - r| \leq \epsilon/2 < \epsilon$. □

We can define Cauchy sequence of **real** numbers. All of the results we proved about Cauchy sequences of rational numbers hold true for Cauchy sequences of real numbers; the proofs are the same. It is then natural to ask if every Cauchy sequence of real numbers converges; otherwise, we would have to repeat the process we used to construct the real numbers and perhaps this would go on for ever. Fortunately, we do not need to do so: every Cauchy sequence of real numbers converges to a real number.

Theorem D.30 (Completeness of the real numbers)**.** *Let $\{r_n\}$ be a Cauchy sequence of real numbers. Then $\{r_n\}$ converges to a real number.*

Proof. For each n, we can choose a rational number x_n so that $|x_n - r_n| < 1/n$. Thus we obtain a sequence $\{x_n\}$ of rational numbers. Clearly the sequence $\{x_n - r_n\}$ of real numbers converges to 0. Applying part (3) of Proposition D.19, we see that $\{x_n\}$ is a Cauchy sequence of rational numbers. Thus it defines a real number $r = [\{x_n\}]$.

We must still show that r_n converges to r. Suppose that we are given $\epsilon > 0$. We know by the previous lemma that $\{x_n\}$ converges to r. Choose N such that $|x_n - r| < \epsilon/2$ if $n \geq N$. Choose N also so large that $1/N < \epsilon/2$ and hence $1/n < \epsilon/2$ for all $n \geq N$. It follows that

$$\begin{aligned} |r_n - r| = |r_n - x_n + x_n - r| &\leq |r_n - x_n| + |x_n - r| \\ &< \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

Thus r_n converges to r. □

Corollary D.31. *A sequence of real numbers is convergent $\iff$ it is a Cauchy sequence.* □

The fact that every Cauchy sequence of real numbers converges to a real number is sometimes described by saying that $\mathbb{R}$ is *complete*. In other words, we do not have to enlarge $\mathbb{R}$ any further.

We have now defined the set of real numbers $\mathbb{R}$. The real numbers contain the rational numbers $\mathbb{Q}$ (technically we actually have identified a rational number q with the equivalence class of the constant sequence all of whose terms are q) and we can add, subtract, multiply and divide real numbers as well as compare them. Likewise we can define sequences of real numbers and limits of sequences. All of the above can be done for rational numbers as well, so in these ways the real numbers do not look any different from the rational numbers.

The essential difference between the rational numbers and the real numbers is the following: the real numbers are *complete*, in other words every Cauchy sequence of real numbers converges to a real number. Thus for example $\mathbb{R}$ contains all infinite decimal expansions (and in fact every element of $\mathbb{R}$ can be so expressed). Here we shall use the completeness of $\mathbb{R}$ to describe two other fundamental properties of $\mathbb{R}$. Actually all of these properties are equivalent, in the sense that any one of them logically implies the other two (and all three hold for $\mathbb{R}$).

To define these properties we begin with some notation and a definition. We shall use the standard notation for intervals: given $a, b \in \mathbb{R}$, define

$$(a, b) = \{x \in \mathbb{R} : a < x < b\}$$
$$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}.$$

Of course $(a, b) = \emptyset$ if $a \geq b$ and $[a, b] = \emptyset$ if $a > b$. We refer to (a, b) as an *open interval* and to $[a, b]$ as a *closed interval*. We can also define "half open intervals $(a, b]$ and $[a, b)$. Finally we set

$$(-\infty, a) = \{x \in \mathbb{R} : x < a\}.$$

We can also define $(-\infty, a]$, (a, ∞) and $[a, \infty)$. However we do not try to give a meaning to the symbols ∞ and $-\infty$ or to define $[-\infty, a)$, say. The sets $(-\infty, a)$ and (a, ∞) will also be called open (unbounded) intervals, whereas $(-\infty, a]$ and $[a, \infty)$ are closed (unbounded) intervals.

Now let us define what it means for a subset of $\mathbb{R}$ to be bounded.

Definition D.32. Let A be a subset of $\mathbb{R}$. The set A is *bounded above* if there exists $C \in \mathbb{R}$ such that $x \leq C$ for all $x \in A$. The number C is a *least upper bound* for A if it is the smallest number with this property, in other words if for every $C' < C$ there exists an $x \in A$ with $C' < x$. Note that we do **not** require that $C \in A$. (In this sense, least upper bounds are different from well-ordering properties.) Similarly, the set A is *bounded below* if there exists a real number D such that $x \geq D$ for all $x \in A$. Greatest lower bounds are defined similarly. Finally the set A is *bounded* if it is bounded above and below.

For example, $\emptyset$ is bounded above and below by any choice of real number C. Thus $\emptyset$ does not have a least upper bound. A closed interval $[a, b]$ is bounded above by b and below by a. Likewise an open interval (a, b) is bounded above by b and below by a. Of course, (a, b) is also bounded above by $b + 1$ or $b + (1/2)$ or $b + 27$; here b is the least upper bound. The set $(-\infty, a)$ is bounded above, with a as the least upper bound, but is not bounded below. Every set which is bounded above by C and below by D is by definition a subset of the interval $[D, C]$. Finally notice that if a set A has a least upper bound C_0, then C_0 is smaller than every other bound for A:

Lemma D.33. *Let $A \subseteq \mathbb{R}$, and let C_0 be a least upper bound for A. If C is any other upper bound for A, then $C \geq C_0$. In particular, C_0 is the unique least upper bound for A.*

Proof. Given another upper bound C for A, we suppose that $C < C_0$ and will derive a contradiction. By the definition of a least upper bound, if $C < C_0$ there is an $x \in A$ with $x > C$. Hence C cannot be an upper bound for A.

It follows that $C_0 \leq C$ for every upper bound C for A. Now suppose that C is also a least upper bound for A. Then $C_0 \leq C$ and also by symmetry $C \leq C_0$. Thus $C = C_0$. □

We shall use the notation $\sup A$ for the least upper bound of A, if it exists (if it exists it is unique by the above lemma). Here sup is read "soup" and stands for *supremum*, Latin for greatest. Likewise a greatest lower bound for A, if it exists, is denoted $\inf A$, where inf stands for *infimum* (smallest in Latin). Sometimes one uses the notation $\operatorname{lub} A$ for $\sup A$ and $\operatorname{glb} A$ (read "glub") for $\inf A$.

With this said, here is the first basic theorem about real numbers:

Theorem D.34 (Existence of least upper bounds)**.** *Let A be a nonempty subset of $\mathbb{R}$ which is bounded above. Then there is a least upper bound for A, necessarily unique.*

Proof. We begin with the following lemma:

Lemma D.35. *Let A be a nonempty set which is bounded above. Then, for every real number $\epsilon > 0$, there exists an element $x \in A$ (depending on ϵ) such that, for every element a of A, $a < x + \epsilon$. In other words, $x \in A$, but every element of A is smaller than $x + \epsilon$.*

Proof. Suppose that the lemma is false. Then for every $x \in A$, there exists an element $a \in A$ with $a \geq x + \epsilon$. Choose $x_0 \in A$, which is possible since A is nonempty. Then there is an $x_1 \in A$ with $x_1 \geq x_0 + \epsilon$. Repeating the above with x_1, there is an $x_2 \in A$ with $x_2 \geq x_1 + \epsilon = x_0 + 2\epsilon$. By induction we find an element $x_n \in A$ with $x_n \geq x_0 + n\epsilon$ for every $n \in \mathbb{N}$. But on the other hand, A is bounded above, by some number C, say. Then $x_0 + n\epsilon \leq x_n \leq C$, so that $n\epsilon \leq C - x_0$ for every $n \in \mathbb{N}$. This contradicts the Archimedean property of $\mathbb{R}$. Thus we must eventually find an x_n such that, for every $a \in A$, $a < x_n + \epsilon$. □

Returning to the proof of the theorem, define a sequence a_n of elements of A as follows. By applying the above claim to the set A and to $\epsilon = 1$, choose a_1 to be an element of A such that every element of A is $< a_1 + 1$. Now apply the above claim to $\epsilon = 1/2$ and the set

$$A_1 = \{x \in A : x \geq a_1\} \subseteq A;$$

here A_1 is nonempty because $a_1 \in A_1$ and A_1 is bounded above because it is a subset of A. We conclude that there there is an $a_2 \in A$ with $a_2 \geq a_1$ and such that, for every $x \in A_1$, $x < a_2 + (1/2)$. Of course, the same holds true for $x \in A$, since if $x \notin A_1$, then $x < a_1 \leq a_2 < a_2 + 1/2$. Continuing in this way by induction, we find a sequence $a_1, a_2, \ldots, a_n, \ldots$ with $a_i \in A$, $a_1 \leq a_2 \leq a_3 \leq \ldots$, and such that, for every $n \in \mathbb{N}$ and every $x \in A$, we have $x < a_n + (1/n)$. In particular, applying this to a_m with $m \geq n$, we see that $0 \leq a_m - a_n < 1/n$. It follows that $\{a_n\}$ is a Cauchy sequence: given $\epsilon > 0$, choose N so that $1/N \leq \epsilon$. Then for $n, m \geq N$, where we assume say that $m \geq n$, we have $|a_m - a_n| < 1/n \leq 1/N \leq \epsilon$. Thus since $\{a_n\}$ is a Cauchy sequence, it converges to some $C_0 \in \mathbb{R}$. We claim that C_0 is a least upper bound for A.

To begin with, notice that $a_n \leq C_0$ for every n. Thus for every $x \in A$ and $n \in \mathbb{N}$, $x < a_n + (1/n) \leq C_0 + (1/n)$. It follows that $x \leq \lim_{n\to\infty}(C_0 + (1/n)) = C_0$, so that C_0 is an upper bound for A. On the other hand, let C be a real number such that $C < C_0$. We claim that C is not an upper bound for A. Since $C_0 - C$ is positive, there is a natural number n with $1/n < C_0 - C$ (just choose $n > 1/(C - C_0)$, by the Archimedean property). Now for every $m \geq n$, $a_m < a_n + (1/n)$, so that $C_0 = \lim_{m\to\infty} a_m \leq a_n + (1/n)$ for every n. Thus $a_n \geq C_0 - (1/n) > C_0 - (C_0 - C) = C$. Since $a_n \in A$ and $a_n > C$, it follows that C cannot be an upper bound for A. This concludes the proof of the theorem. □

Of course, a similar argument shows that every set of real numbers which is bounded below has a greatest lower bound. This also follows from the above theorem directly, by noting that, if D_0 is a greatest lower bound for $A \subset \mathbb{R}$, then $-D_0$ is a least upper bound for the set $-A = \{-a : a \in A\}$, and conversely.

As a corollary, we have the following useful criterion for convergence of a sequence:

Corollary D.36. *Let $\{x_n\}$ be an increasing sequence of real numbers, i.e. for all n, $x_n \leq x_{n+1}$. Then $\{x_n\}$ converges $\iff$ it is bounded above, i.e. there exists a $C \in \mathbb{R}$ such that, for all n, $x_n \leq C$. In this case, $\lim_{n\to\infty} x_n = L$, where L is the least upper bound of the set $\{x_n : n \in \mathbb{N}\}$.*

Proof. As we have seen, if $\{x_n\}$ converges then it is bounded and hence it is bounded above (note that an increasing sequence is trivially bounded below by x_1). Conversely, suppose that $\{x_n\}$ is bounded above. As $\{x_n : n \in \mathbb{N}\}$ is clearly nonempty and bounded above, it has a least upper bound L. We shall show that $\lim_{n\to\infty} x_n = L$. Choose $\epsilon > 0$. Then $L - \epsilon$ is not an upper bound

for $\{x_n : n \in \mathbb{N}\}$. Thus there exists an N such that $L - \epsilon < x_N \leq L$ and hence $0 \leq L - x_N < \epsilon$. Since $\{x_n\}$ is increasing, for all $n \geq N$, $x_N \leq x_n \leq L$, and hence

$$0 \leq L - x_n \leq L - x_N < \epsilon.$$

Thus $|L - x_n| < \epsilon$ for all $n \geq N$, so that $\lim_{n\to\infty} x_n = L$. □

Corollary D.37. *Let $\{x_n\}$ be a sequence of non-negative real numbers i.e. $x_n \geq 0$ for all n. Let s_n be the n^{th} partial sum of the sequence $\{x_n\}$: $s_n = \sum_{i=1}^{n} x_i$. Then the sequence $\{s_n\}$ converges $\iff$ the partial sums s_n are bounded, i.e. there exists a $C \in \mathbb{R}$ such that, for all n, $s_n \leq C$.*

Proof. Since the $x_n \geq 0$, the partial sums s_n are increasing in the sense of the previous corollary: $s_{n+1} = s_n + x_n \geq s_n$. Thus, the result is immediate from the previous corollary. □

Here is the second main property of the real numbers which we shall use. It says that a sequence $\{x_n\}$ such that the terms x_n are bounded has to have a convergent subsequence, in other words there must be some real number r such that infinitely many of the x_n are getting closer and closer to r.

Theorem D.38 (Bolzano-Weierstrass)**.** *Let $\{x_n\}$ be a bounded sequence of real numbers. Then $\{x_n\}$ has a convergent subsequence.*

Typical examples are the sequence $(-1)^n$ or the sequence of rational numbers between 0 and 1 described previously.

Proof. To say that $\{x_n\}$ is bounded says that there are real numbers D and C such that $x_n \in [D, C]$ for every $n \in \mathbb{N}$. Let $\ell = C - D$ denote the length of this interval. Now divide the interval into two halves:

$$[D, C] = \left[D, \frac{D+C}{2}\right] \cup \left[\frac{D+C}{2}, C\right].$$

Each of these new intervals has length $(C - D)/2 = \ell/2$. Because the sequence contains infinitely many terms, at least one of these halves, say $\left[D, \frac{D+C}{2}\right]$, contains infinitely many terms of the sequence. In particular, we can choose n_1 such that $x_{n_1} \in \left[D, \frac{D+C}{2}\right]$. Now take the interval $\left[D, \frac{D+C}{2}\right]$ and divide it again into two half intervals, each of length $\frac{1}{2}(\ell/2) = \ell/4$. Again at least one of the new half intervals must contain infinitely many terms of the sequence. Choose an $n_2 > n_1$ such that x_{n_2} lies in one of the new half intervals with infinitely many terms, and repeat this process.

At stage k we have found a subinterval of $[D, C]$ of length $\ell/2^k$, containing infinitely many of the terms of the sequence, and we choose an x_{n_k} in this interval. Moreover all future choices of x_{n_j} for $j \geq k$ are also going to lie in this interval of length $\ell/2^k$. Thus, given $\epsilon > 0$, if we choose N so that $\ell/2^N < \epsilon$, then for all $k, j \geq N$, x_{n_k} and x_{n_j} both lie in an interval of length $\ell/2^N$ and so $|x_{n_k} - x_{n_j}| \leq \ell/2^N < \epsilon$. Thus $\{x_{n_k}\}$ is a Cauchy sequence and so converges. We have therefore found a convergent subsequence of $\{x_n\}$, as required. This concludes the proof of the Bolzano-Weierstrass theorem. $\square$

Chapter 7

Basic topology of $\mathbb{R}^n$

7.1 Norms

In $\mathbb{R}$, we have the basic measure of the size of a real number, the absolute value, and the distance between two real numbers x and y, i.e. $|x-y|$. Using the absolute value, we define sequences, convergent sequences, Cauchy sequences, bounded sets, and can then state the basic facts about real numbers:

1. Completeness (every Cauchy sequence of real numbers converges).

2. Least upper bound property: Every nonempty subset of $\mathbb{R}$ which is bounded above has a least upper bound. (In terms of sequences, this is essentially the statement that every increasing sequence of real numbers which is bounded above converges.)

3. The Bolzano-Weierstrass theorem: every bounded sequence of real numbers has a convergent subsequence.

How might we go about generalizing some of the above to $\mathbb{R}^n$? In $\mathbb{R}^n$, as a generalization of absolute value, we have the distance of a vector $\mathbf{x}$ from the origin, namely $\|\mathbf{x}\|$, and hence the distance $\|\mathbf{x}-\mathbf{y}\|$ between two vectors. The function $\|\cdot\|$ is a function $\mathbb{R}^n \to \mathbb{R}$ with the following properties:

1. (Positivity) For all $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.

2. (Homogeneity) For all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, $\|t\mathbf{x}\| = |t|\|\mathbf{x}\|$.

3. (Triangle inequality) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

In general, any function N from $\mathbb{R}^n$ to $\mathbb{R}$, or more generally from a vector space V to $\mathbb{R}$, satisfying the above properties will be called a *norm*. A norm N is a way to measure the size of a vector. From N we can define a "distance" between two vectors $\mathbf{x}, \mathbf{y}$ by taking $N(\mathbf{x} - \mathbf{y})$. Such a distance is *translation invariant* in the sense that for any fixed $\mathbf{p} \in \mathbb{R}^n$, the distance from $\mathbf{x} + \mathbf{p}$ to $\mathbf{y} + \mathbf{p}$ is the same as the distance from $\mathbf{x}$ to $\mathbf{y}$. Likewise (2), the homogeneity property says that if we scale the vectors $\mathbf{x}, \mathbf{y}$ by a real number t, then the distance changes by $|t|$. Taking $t = -1$ we see that the distance from $\mathbf{x}$ to $\mathbf{y}$ is the same as the distance from $\mathbf{y}$ to $\mathbf{x}$. Finally (3), the triangle inequality, says that for all $\mathbf{x}, \mathbf{y}, \mathbf{z}$, the distance from $\mathbf{x}$ to $\mathbf{z}$ is less than the sum of the distance from $\mathbf{x}$ to $\mathbf{y}$ plus the distance from $\mathbf{y}$ to $\mathbf{z}$, since

$$\begin{aligned} N(\mathbf{x} - \mathbf{z}) &= N(\mathbf{x} - \mathbf{y} + \mathbf{y} - \mathbf{z}) \\ &\leq N(\mathbf{x} - \mathbf{y}) + N(\mathbf{y} - \mathbf{z}). \end{aligned}$$

We will refer to the function $\|\cdot\|$ as the *usual* or *standard* norm on $\mathbb{R}^n$. It is easy to see that every norm on $\mathbb{R}$ is a positive multiple of the usual one, i.e. of absolute value. But for $n > 1$, there are a great number of different norms on $\mathbb{R}^n$ which are not multiples of the standard norm or of each other. Here are some important examples of norms:

Proposition 7.1. *The following are all norms on $\mathbb{R}^n$:*

1. $\|\mathbf{x}\|_\infty = \max\{|x_i| : 1 \leq i \leq n\}$, *the largest absolute value of the components of* $\mathbf{x}$.

2. $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$, *the sum of the absolute values of the components of* $\mathbf{x}$.

3. $\sqrt{Q(\mathbf{x})} = \sqrt{B(\mathbf{x}, \mathbf{x})}$, *where B is a positive definite bilinear form. For example, if λ_i is a positive real number, $1 \leq i \leq n$, then $\sqrt{\sum_{i=1}^n \lambda_i x_i^2}$ is a norm.*

Proof. The proof that $\|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_1$ are norms is a straightforward application of the standard properties of the absolute value in $\mathbb{R}$. The proof that the square root of $Q(\mathbf{x}) = B(\mathbf{x}, \mathbf{x})$ is a norm if B is a positive definite bilinear form consists in checking that the proof of the Cauchy-Schwarz inequality for the usual inner product and the triangle inequality for the standard norm work in the context of a more general positive definite bilinear form. □

In many ways, the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are also very natural ways to measure the size of a vector, or the distance between two vectors. For example, in New York, where we have to travel vertically or horizontally using the numbered streets and avenues, the distance between two points is naturally given by the norm $\|\cdot\|_1$. The norm $\|\cdot\|_\infty$ is also technically very useful.

There are the following relations between the norms $\|\mathbf{x}\|, \|\mathbf{x}\|_\infty, \|\mathbf{x}\|_1$: first, since $|x_i| \leq \|\mathbf{x}\|_\infty$, and $\|\mathbf{x}\|_\infty = |x_i|$ for at least one i,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \qquad \text{and} \qquad \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty.$$

Likewise

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\| \qquad \text{and} \qquad \|\mathbf{x}\| \leq \sqrt{n}\|\mathbf{x}\|_\infty.$$

Using the above, we can compare $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|$. This would give

$$\|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty \leq n\|\mathbf{x}\| \qquad \text{and} \qquad \|\mathbf{x}\| \leq \sqrt{n}\|\mathbf{x}\|_\infty \leq \sqrt{n}\|\mathbf{x}\|_1.$$

In fact, there are better inequalities

$$\|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\| \qquad \text{and} \qquad \|\mathbf{x}\| \leq \|\mathbf{x}\|_1.$$

These are left as exercises.

In general, if we can compare two different norms using some universal constant as above we shall call them equivalent:

Definition 7.2. Two norms N_1 and N_2 are *equivalent* if there exist positive real numbers C_1 and C_2 such that, for all $\mathbf{x} \in \mathbb{R}^n$,

$$N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x}) \qquad \text{and} \qquad N_2(\mathbf{x}) \leq C_2 N_1(\mathbf{x}).$$

Thus the norms $\|\cdot\|_1, \|\cdot\|_\infty, \|\cdot\|$ are all equivalent.

For example, if $N(\mathbf{x}) = \sqrt{\sum_{i=1}^n \lambda_i x_i^2}$, where the λ_i are positive real numbers, and we let C_1 be the minimum value of the $\sqrt{\lambda_i}$ and C_2 be the maximum value, then clearly

$$C_1\|\mathbf{x}\| \leq N(\mathbf{x}) \leq C_2\|\mathbf{x}\|.$$

Thus $N(\mathbf{x})$ is equivalent to the standard norm.

The following is straightforward and is left as an exercise:

Proposition 7.3. *The relation $\sim$ on norms defined by $N_1 \sim N_2 \iff N_1$ and N_2 are equivalent is an equivalence relation.* □

It is a fact which we shall prove in the next chapter that all norms on $\mathbb{R}^n$ are equivalent. This fails for infinite-dimensional spaces. For example, let V be the vector space $C^0([a,b])$ of continuous functions on the closed interval $[a,b]$. Then there are three different but very natural norms on V:

1. $\|f\|_\infty = \sup\{f(x) : x \in [a,b]\}$ (which exists by the extreme value theorem);
2. $\|f\|_1 = \int_a^b |f(x)|\,dx$;
3. $\|f\|_2 = \left(\int_a^b (f(x))^2\,dx\right)^{1/2}$.

One can show that no two of the three norms above are equivalent.

Definition 7.4. Given any norm N, the *ball of radius r about* $\mathbf{v} \in \mathbb{R}^n$ (with respect to N) as follows :

$$B_r(\mathbf{v}) = \{\mathbf{x} \in \mathbb{R}^n : N(\mathbf{x}-\mathbf{v}) < r\}.$$

It is the set of all points $\mathbf{x}$ of distance less that r (using N to define distance) from $\mathbf{v}$. We sometimes call $B_r(\mathbf{v})$ the *open ball of radius r about* $\mathbf{v}$, for reasons which we explain below. Thus if $n = 1$ a ball of radius r about $a \in \mathbb{R}$ (for the usual distance) is an open interval $(a-r, a+r)$ centered at a. We can also define the *closed ball of radius r about* $\mathbf{v}$ with respect to N:

$$\overline{B_r(\mathbf{v})} = \{\mathbf{x} \in \mathbb{R}^n : N(\mathbf{x}-\mathbf{v}) \le r\},$$

and the *sphere of radius r* with respect to N:

$$S_r(\mathbf{v}) = \{\mathbf{x} \in \mathbb{R}^n : N(\mathbf{x}-\mathbf{v}) = r\}.$$

If $\mathbf{v}$ is the origin we shall just write B_r, $\overline{B_r}$ and S_r. Also, since N will usually be clear from the context we shall omit any reference to it, and if we have not specified a norm in advance then it is always understood that we mean the standard norm. For $r = 1$ we shall call B_1 and S_1 the *unit ball* and *unit sphere* about the origin.

Example 7.5. (1) For the standard norm, S^1 is the unit circle $\{(x_1, x_2) : x_1^2 + x_2^2 = 1\}$, S^2 is the unit sphere $\{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 = 1\}$ in $\mathbb{R}^3$, and S^0 is the two points ± 1 in $\mathbb{R}$. We write $S^{n-1} = S_1 \subset \mathbb{R}^n$ (again, for the standard norm). It is the set of all points in $\mathbb{R}^n$ of distance 1 from the origin, and is called the $(n-1)$-*sphere* (of radius one).

(2) For $n = 2$, the closed unit ball for the norm $\|\mathbf{x}\|_\infty$ is the square with the four vertices $(\pm 1, \pm 1)$, and hence the square parallel to the x- and y-axes, centered at the origin, with side length equal to 2. The picture in higher dimensions is similar. If instead we consider the the closed unit ball for the norm $\|\mathbf{x}\|_1$, again in $\mathbb{R}^2$, then it is the square with the four vertices $(\pm 1, 0)$ and $(0, \pm 1)$, i.e. the square formed by the lines $x_1 + x_2 = \pm 1$ and $x_1 - x_2 = \pm 1$.

One important property of balls (for any norm) is the following: if $\mathbf{x}, \mathbf{y} \in B_r(\mathbf{v})$, then the line segment joining $\mathbf{x}$ to $\mathbf{y}$ lies in $B_r(\mathbf{v})$. In general, given two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the line joining $\mathbf{x}$ to $\mathbf{y}$ is given in parametric form by

$$\{t\mathbf{x} + (1-t)\mathbf{y} : t \in \mathbb{R}\}.$$

Note that $t = 0$ corresponds to $\mathbf{y}$ and that $t = 1$ corresponds to $\mathbf{x}$. So for $0 \leq t \leq 1$ the corresponding points of the line run between $\mathbf{y}$ and $\mathbf{x}$. (This should be intuitively clear, but it isn't so easy to say in precise terms what we mean. For example, what does "between" mean here?)

Definition 7.6. Let $X \subseteq \mathbb{R}^n$. Then X is *convex* if, for every $\mathbf{x}, \mathbf{y} \in X$, the line segment between $\mathbf{x}$ and $\mathbf{y}$ is contained in X.

For example, $\mathbb{R}^n$ is convex. The subset $\{\mathbf{0}\}$ is convex, as is every subset consisting of a single point. A line is convex. Every vector subspace of $\mathbb{R}^n$ is convex, as is every affine subspace. The empty set is (unfortunately) also convex.

Proposition 7.7. *If N is a norm on $\mathbb{R}^n$, then for every $r \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$, the ball $B_r(\mathbf{v})$ of radius r about $\mathbf{v}$ with respect to N is convex.*

Proof. Suppose that $\mathbf{x}, \mathbf{y} \in B_r(\mathbf{v})$. Thus $N(\mathbf{x} - \mathbf{v}) < r$ and $N(\mathbf{y} - \mathbf{v}) < r$. Given $t \in \mathbb{R}$ with $0 \leq t \leq 1$, since $t\mathbf{v} + (1-t)\mathbf{v} = \mathbf{v}$,

$$\begin{aligned} N(t\mathbf{x} + (1-t)\mathbf{y} - \mathbf{v}) &= N(t\mathbf{x} + (1-t)\mathbf{y} - (t\mathbf{v} + (1-t)\mathbf{v})) \\ &= N(t(\mathbf{x} - \mathbf{v}) + (1-t)(\mathbf{y} - \mathbf{v})) \\ &\leq N(t(\mathbf{x} - \mathbf{v})) + N((1-t)(\mathbf{y} - \mathbf{v})) \\ &= |t|N(\mathbf{x} - \mathbf{v}) + |1-t|N(\mathbf{y} - \mathbf{v}) \\ &< tr + (1-t)r = r. \end{aligned}$$

□

In terms of balls, we can restate the meaning of equivalent norms as follows:

Lemma 7.8. *Suppose that N_1 and N_2 are two norms on $\mathbb{R}^n$ and let C_1 be a positive real number. Then the following are equivalent:*

1. *For all $\mathbf{x} \in \mathbb{R}^n$ $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$.*

2. *For all $\mathbf{x}$ such that $N_2(\mathbf{x}) \leq 1$ (i.e. for $\mathbf{x}$ in the closed unit ball for N_2), $N_1(\mathbf{x}) \leq C_1$.*

3. *For all $\mathbf{x}$ such that $N_2(\mathbf{x}) = 1$ (i.e. for $\mathbf{x}$ in the unit sphere for N_2), $N_1(\mathbf{x}) \leq C_1$.*

Proof. (1) $\implies$ (2): If $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$, then for $N_2(\mathbf{x}) \leq 1$ we have $N_1(\mathbf{x}) \leq C_1$. Thus if $\mathbf{x}$ lies in the closed unit ball for N_2, then $N_1(\mathbf{x}) \leq C_1$.

(2) $\implies$ (3): This is clear since the closed unit sphere is a subset of the closed unit ball.

(3) $\implies$ (1): Suppose that $N_1(\mathbf{x}) \leq C_1$ for all $\mathbf{x}$ such that $N_2(\mathbf{x}) = 1$, and let $\mathbf{x}$ be an arbitrary vector in $\mathbb{R}^n$. If $\mathbf{x} = 0$, then $N_1(\mathbf{x}) = N_2(\mathbf{x}) = 0$, so that we certainly have $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$. If $\mathbf{x} \neq 0$, then let $t = 1/N_2(\mathbf{x})$, which is defined since $N_2(\mathbf{x}) \neq 0$. Then

$$N_2(t\mathbf{x}) = |t| N_2(\mathbf{x}) = N_2(\mathbf{x})/N_2(\mathbf{x}) = 1.$$

Thus by assumption

$$N_1(t\mathbf{x}) = |t| N_1(\mathbf{x}) \leq C_1.$$

Since $|t| = 1/N_2(\mathbf{x})$, this says that $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$. □

The proof of the following is very similar:

Lemma 7.9. *Suppose that N_1 and N_2 are two norms on $\mathbb{R}^n$ and let C_1 be a positive real number. Then the following are equivalent:*

1. *For all $\mathbf{x} \in \mathbb{R}^n$, $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$.*

2. *For every $r > 0$, the open ball of radius r/C_1 about the origin for N_2 is contained the ball of radius r about the origin for N_1.*

3. *There exists some $r > 0$ such that the open ball of radius r/C_1 about the origin for N_2 is contained the ball of radius r about the origin for N_1.*

Proof. Clearly (1) $\implies$ (2) $\implies$ (3). To see that (3) $\implies$ (1), note that (3) is equivalent to: if $C_1 N_2(\mathbf{x}) < r$, then $N_1(\mathbf{x}) < r$. By the equivalence of (2) and (1) in Lemma 7.8, it suffices to show that, if $N_2(\mathbf{x}) \leq 1$, then $N_1(\mathbf{x}) \leq C_1$. Let ϵ be such that $0 < \epsilon < r$. Then $0 < (r-\epsilon)/C_1 < r/C_1$. Thus, if $N_2(\mathbf{x}) \leq 1$, then

$$N_2\left(\left(\frac{r-\epsilon}{C_1}\right)\mathbf{x}\right) = \left(\frac{r-\epsilon}{C_1}\right) N_2(\mathbf{x}) \leq \left(\frac{r-\epsilon}{C_1}\right) < r/C_1.$$

Hence, by the assumption of (3),

$$N_1\left(\left(\frac{r-\epsilon}{C_1}\right)\mathbf{x}\right) = \left(\frac{r-\epsilon}{C_1}\right) N_1(\mathbf{x}) < r.$$

Thus, for all $\epsilon > 0$, if $\epsilon < r$, then

$$N_1(\mathbf{x}) < C_1\left(\frac{r}{r-\epsilon}\right).$$

Since this is true for all $\epsilon > 0$, we see that $N_1(\mathbf{x}) \leq C_1$. □

Corollary 7.10. *Let N_1 and N_2 be two norms on $\mathbb{R}^n$. Then N_1 and N_2 are equivalent $\iff$ there exist positive real numbers r_1 and ϵ_1 such that the open ball of radius ϵ_1 for N_2 is contained in the open ball of radius r_1 for N_1 and there exist positive real numbers r_2 and ϵ_2 such that the open ball of radius ϵ_2 for N_1 is contained in the open ball of radius r_2 for N_2.*

Proof. By Lemma 7.9, if there exists a positive real number C_1 such that $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$, then for every $r_1 > 0$, the open ball of radius $\epsilon_1 = r_1/C_1$ about the origin for N_2 is contained the ball of radius r_1 about the origin for N_1. Conversely suppose that there exist positive real numbers r_1 and ϵ_1 such that the open ball of radius ϵ_1 for N_2 is contained in the open ball of radius r_1 for N_1. Set $C_1 = r_1/\epsilon_1$ and hence $\epsilon_1 = r_1/C_1$. By Lemma 7.9 again, $N_1(\mathbf{x}) \leq C_1 N_2(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. A similar argument handles the analogous statement in case there exists a positive real number C_2 such that $N_2(\mathbf{x}) \leq C_2 N_1(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. □

An easy argument using the translation invariance of distance then shows:

Corollary 7.11. *Suppose that N_1 and N_2 are equivalent norms on $\mathbb{R}^n$. Then, for all $\mathbf{p} \in \mathbb{R}^n$ and all $r > 0$, there exists an $\epsilon > 0$ such that the open ball of radius ϵ about $\mathbf{p}$ for N_2 is contained the ball of radius r about $\mathbf{p}$ for N_1.* □

7.2 Sequences; open and closed sets

Just as in $\mathbb{R}$, sequences will play an important role:

Definition 7.12. A *sequence* in $\mathbb{R}^n$ is a collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots$ indexed by the positive integers. Technically a sequence is a function from $\mathbb{N}$ to $\mathbb{R}^n$, whose value at k is denoted $\mathbf{v}_k$ (or by whatever letter you choose). If we write $\mathbf{v}_k$ in terms of its components, then $\mathbf{v}_k = (v_{1,k}, \ldots, v_{n,k})$. Here each $v_{i,k}$ is a sequence in its own right (we fix i and let k, the sequence variable, range over $\mathbb{N}$). Thus a sequence in $\mathbb{R}^n$ is the same thing as an ordered collection of n sequences in $\mathbb{R}$. We say that the sequence $\mathbf{v}_k$ *converges* to $\mathbf{L} \in \mathbb{R}^n$ (with respect to the norm N) if, for all $\epsilon > 0$, there exists a $K \in \mathbb{N}$ such that $N(\mathbf{L} - \mathbf{v}_k) < \epsilon$ if $k \geq K$. As usual we write $\lim_{k\to\infty} \mathbf{v}_k = \mathbf{L}$ to mean that the sequence $\{\mathbf{v}_k\}$ converges to $\mathbf{L}$ (with respect to the norm N, which is supposed to be understood from the discussion), and say that the sequence $\mathbf{v}_k$ *converges* if it converges to some $\mathbf{L}$. In this case $\mathbf{L}$ is necessarily unique, by the same argument we used for the one variable case.

An easy argument shows the following:

Lemma 7.13. *Suppose that N_1 and N_2 are two equivalent norms. Then a sequence $\{\mathbf{v}_k\}$ converges with respect to N_1 if and only if it converges with respect to N_2, and the limits are the same.* □

Thus, once we know that all norms on $\mathbb{R}^n$ are equivalent, we can drop the qualifier "with respect to N" in the definition.

The next result says that a sequence of vectors in $\mathbb{R}^n$ converges to a limit vector $\iff$ each on the n sequences of components converge to a limit in $\mathbb{R}$.

Proposition 7.14. *The sequence $\{\mathbf{v}_k\} = \{(v_{1,k}, \ldots, v_{n,k})\}$ converges to $\mathbf{L} = (L_1, \ldots, L_n)$ for any of the equivalent norms $\|\cdot\|$, $\|\cdot\|_\infty$, $\|\cdot\|_1$, if and only if each of the n sequences $v_{i,k}$ converges to L_i.*

Proof. It suffices to work with the norm $\|\cdot\|_\infty$. First suppose that, for $\mathbf{v}_k = (v_{1,k}, \ldots, v_{n,k})$, the sequence $\{\mathbf{v}_k\}$ converges to $\mathbf{L} = (L_1, \ldots, L_n)$, in the norm $\|\cdot\|_\infty$. Thus, for all $\epsilon > 0$, there exists a K such that, for all $k \geq K$, $\|\mathbf{v}_k - \mathbf{L}\|_\infty = \max\{|v_{i,k} - L_i|\} < \epsilon$. Since $|v_{j,k} - L_j| \leq \max\{|v_{i,k} - L_i|\}$, this says that for all $j, 1 \leq j \leq n$ and for $k \geq K$, $|v_{j,k} - L_j| < \epsilon$. Thus by definition $\{v_{j,k}\}$ converges to L_j for every j.

Conversely, suppose that $\{v_{j,k}\}$ converges to L_j for every j. Given $\epsilon > 0$, for each j we can choose a K_j such that, if $k \geq K_j$, then $|v_{j,k} - L_j| < \epsilon$.

Now if $K = \max\{K_j : j = 1, \ldots, n\}$, then for $k \geq K$ we have $|v_{j,k} - L_j| < \epsilon$ for every j. Thus $\max\{|v_{j,k} - L_j| : j = 1, \ldots, n\} = \|\mathbf{v}_k - \mathbf{L}\|_\infty < \epsilon$. So by definition $\{\mathbf{v}_k\}$ converges to $\mathbf{L}$. □

Just as in the one variable case, we have:

Lemma 7.15. *1. If $\{\mathbf{v}_k\}$ and $\{\mathbf{w}_k\}$ are two sequences in $\mathbb{R}^n$ such that $\{\mathbf{v}_k\}$ converges to $\mathbf{L}$ and $\{\mathbf{w}_k\}$ converges to $\mathbf{M}$ then $\{\mathbf{v}_k + \mathbf{w}_k\}$ converges to $\mathbf{L} + \mathbf{M}$.*

2. If $\{\mathbf{v}_k\}$ is a sequence in $\mathbb{R}^n$ and $\{t_k\}$ is a sequence in $\mathbb{R}$ such that $\{\mathbf{v}_k\}$ converges to $\mathbf{L}$ and $\{t_k\}$ converges to t then $\{t_k\mathbf{v}_k\}$ converges to $t\mathbf{L}$. □

Lemma 7.15 can be proved directly by imitating the one variable case. For a norm equivalent to the standard norm, it also follows directly from the one-variable case by invoking Proposition 7.14.

Cauchy sequences in $\mathbb{R}^n$ are defined just as they are for $\mathbb{R}$. An argument very similar to the above gives:

Proposition 7.16. *The sequence $\{\mathbf{v}_k\} = \{(v_{1,k}, \ldots, v_{n,k})\}$ is a Cauchy sequence for any of the equivalent norms $\|\cdot\|$, $\|\cdot\|_\infty$, $\|\cdot\|_1$, if and only if each of the n sequences $\{v_{i,k}\}$ is a Cauchy sequence. Hence, every Cauchy sequence in $\mathbb{R}^n$ (for any of the equivalent norms $\|\cdot\|$, $\|\cdot\|_\infty$, $\|\cdot\|_1$) converges.* □

When we look at real valued functions of one variable, we usually don't require that they be defined on all real numbers, but only on certain subsets. The reasonable subsets of $\mathbb{R}$ are usually the intervals. They come in three kinds (here it is understood that $a, b \in \mathbb{R}$ with $a < b$):

1. the *open* intervals $(a, b), (a, \infty), (-\infty, b), (-\infty, \infty)$;
2. the *closed* intervals $[a, b], [a, \infty), (-\infty, b], (-\infty, \infty)$;
3. the *half-open* intervals $(a, b], [a, b)$.

The natural generalizations of open and closed intervals are open and closed balls in $\mathbb{R}^n$. However, because of the many possibilities for the geometry of a subset of $\mathbb{R}^n$, it is better to have a more inclusive definition:

Definition 7.17. Let N be a norm on $\mathbb{R}^n$. A subset X of $\mathbb{R}^n$ is *open* (with respect to N) if, for all $\mathbf{v} \in X$, there exists an $r > 0$ (depending on $\mathbf{v}$ and X) such that X contains a ball of radius r about $\mathbf{v}$, i.e. such that $B_r(\mathbf{v}) \subseteq X$.

The meaning of openness is that X is open exactly when, for all $\mathbf{v} \in X$, X contains all points "close enough" to $\mathbf{v}$. For example, $\mathbb{R}^n$ is open. The empty set is (vacuously) open. An open interval in $\mathbb{R}$ is open; more generally we have:

Proposition 7.18. *For all $\mathbf{v} \in \mathbb{R}^n$ and $r \in \mathbb{R}$, the ball $B_r(\mathbf{v})$ with respect to N is an open set with respect to N.*

Proof. Suppose that $\mathbf{w} \in B_r(\mathbf{v})$, so that $N(\mathbf{w} - \mathbf{v}) = t < r$. We claim that $B_{r-t}(\mathbf{w}) \subseteq B_r(\mathbf{v})$. This follows from the triangle inequality: if $\mathbf{x} \in B_{r-t}(\mathbf{w})$, then

$$\begin{aligned} N(\mathbf{x} - \mathbf{v}) &= N(\mathbf{x} - \mathbf{w} + \mathbf{w} - \mathbf{v}) \\ &\leq N(\mathbf{x} - \mathbf{w}) + N(\mathbf{w} - \mathbf{v}) < r - t + t = r. \end{aligned}$$

□

Next we show that the definition of open set does not depend on the choice of norm:

Lemma 7.19. *If N_1 and N_2 are equivalent norms, then a set X is open with respect to N_1 if and only if it is open with respect to N_2.*

Proof. This is immediate from Corollary 7.11. □

Remark 7.20. (1) From now on we omit the qualifier "with respect to N," taking N as clear from the context. As we shall show that all norms on $\mathbb{R}^n$ are equivalent, it is reasonable that we don't in fact ever need to specify N.

(2) It is easy to see from Lemma 7.9 that the converse to Lemma 7.19 also holds: If N_1 and N_2 define the same collection of open sets, then N_1 and N_2 are equivalent.

Proposition 7.21. *The union of two open sets in $\mathbb{R}^n$ is open; in fact, the union of possibly infinitely many subsets of $\mathbb{R}^n$ is open. The intersection of two open subsets of $\mathbb{R}^n$ is open, and in fact the intersection of finitely many open subsets of $\mathbb{R}^n$ is open.*

Proof. If $X_i, i \in I$ is a collection of open subsets of $\mathbb{R}^n$, then

$$\bigcup_{i \in I} X_i = \{\mathbf{x} \in \mathbb{R}^n : \text{for some } i \in I,\ \mathbf{x} \in X_i\}.$$

Thus $X_i \subseteq \bigcup_{i \in I} X_i$ for all i. Now given $\mathbf{x} \in \bigcup_{i \in I} X_i$, there exists an i such that $\mathbf{x} \in X_i$. Since X_i is open, there is an $r > 0$ such that $B_r(\mathbf{x}) \subseteq X_i$. Thus $B_r(\mathbf{x}) \subseteq X_i \subseteq \bigcup_{i \in I} X_i$. By definition therefore $\bigcup_{i \in I} X_i$ is open.

Now suppose that X_1 and X_2 are two open sets and that $\mathbf{x} \in X_1 \cap X_2$. Thus $\mathbf{x} \in X_1$, and by definition there is a ball $B_{r_1}(\mathbf{x})$ contained in X_1. Likewise $\mathbf{x} \in X_2$, and there is a ball $B_{r_2}(\mathbf{x})$ contained in X_2. Take $r = \min\{r_1, r_2\}$. Then $B_r(\mathbf{x}) = B_{r_1}(\mathbf{x}) \cap B_{r_2}(\mathbf{x}) \subseteq X_1 \cap X_2$. By definition therefore $X_1 \cap X_2$ is open. The case of finitely many open sets follows by induction. □

The above argument does not work for the intersection of infinitely many open sets and in fact the intersection of infinitely many open sets need not be open (exercise).

As a consequence of Proposition 7.21, a union of open intervals in $\mathbb{R}$ is open. Now the union of two non-disjoint open intervals, say $(a_1, b_1) \cup (a_2, b_2)$ with $a_2 < b_1$, is again an open interval (a_1, b_2). So the above is really only interesting for a union of disjoint intervals. One can show: a subset X of $\mathbb{R}$ is open $\iff$ X is a disjoint union of (possibly infinitely many) open intervals. For example, $\mathbb{R} - \mathbb{Z} = \bigcup_{n \in \mathbb{Z}} (n, n+1)$ is a disjoint union of open intervals. Another example is $\bigcup_{n \in \mathbb{N}} \left(\frac{1}{n+1}, \frac{1}{n} \right)$.

However, for $n > 1$, there is no straightforward description of open sets, which can come in a wide variety of shapes and sizes.

Definition 7.22. A subset X of $\mathbb{R}^n$ is *closed* with respect to N if $\mathbb{R}^n - X$ is open with respect to N. (Here if X is a subset of a set Z, then by definition $Z - X = \{z \in Z : z \notin X\}$.) As for open sets, we will drop the phrase "with respect to N" since it will be clear from the context, and eventually unnecessary.

The sets $\mathbb{R}^n$ and $\emptyset$ are closed since $\mathbb{R}^n = \mathbb{R}^n - \emptyset$ and $\emptyset = \mathbb{R}^n - \mathbb{R}^n$. A subset of $\mathbb{R}^n$ may be open, closed, or neither. Also, a subset of $\mathbb{R}^n$ may be both open and closed; in fact, as we shall see later, the only two such subsets are $\mathbb{R}^n$ itself and $\emptyset$. For example, a closed interval $[a, b]$ in $\mathbb{R}$ is closed, since $\mathbb{R} - [a, b] = (-\infty, a) \cap (b, \infty)$ is open. Likewise a closed ball and a sphere are closed:

Proposition 7.23. *For all $\mathbf{v} \in \mathbb{R}^n$ and $r \in \mathbb{R}$, the sets $\overline{B_r(\mathbf{v})}$ and $S_r(\mathbf{v})$ are closed.*

Proof. We must show first that

$$\mathbb{R}^n - \overline{B_r(\mathbf{v})} = \{\mathbf{x} \in \mathbb{R}^n : N(\mathbf{x} - \mathbf{v}) > r\}$$

is open. To do so, we can just imitate the argument of the proof that $B_r(\mathbf{v})$ was open to show that, if $\mathbf{x} \in \mathbb{R}^n - \overline{B_r(\mathbf{v})}$ and $N(\mathbf{x}-\mathbf{v}) = t > r$, then $B_{t-r}(\mathbf{x}) \subseteq \mathbb{R}^n - \overline{B_r(\mathbf{v})}$—proof as exercise. Then

$$\mathbb{R}^n - S_r(\mathbf{v}) = (\mathbb{R}^n - \overline{B_r(\mathbf{v})}) \cup B_r(\mathbf{v}).$$

As we have seen, this is the union of two open sets and is therefore open. □

By logic, (de Morgan's laws)

$$\mathbb{R}^n - \bigcup_i X_i = \bigcap_i (\mathbb{R}^n - X_i) \qquad \text{and} \qquad \mathbb{R}^n - (X_1 \cap X_2) = (\mathbb{R}^n - X_1) \cup (\mathbb{R}^n - X_2).$$

Thus the intersection of arbitrarily many closed sets is closed and the union of finitely many closed sets is closed.

The subset $\mathbb{Z}$ of $\mathbb{R}$ is closed since its complement is open. The subset $\{1, 1/2, \dots\} \cup \{0\} = \{1/n : n \in \mathbb{N}\} \cup \{0\}$ is closed since its complement is $\bigcup_{n\in\mathbb{N}} \left(\frac{1}{n+1}, \frac{1}{n}\right) \cup (-\infty, 0) \cup (1, \infty)$. But $\{1/n : n \in \mathbb{N}\}$ is not closed, since $0 \in \mathbb{R} - \{1/n : n \in \mathbb{N}\}$ but there is no open interval about 0 which is contained in $\mathbb{R} - \{1/n : n \in \mathbb{N}\}$, or in other words does not intersect $\{1/n : n \in \mathbb{N}\}$. Hence $\mathbb{R} - \{1/n : n \in \mathbb{N}\}$ is not open and so $\{1/n : n \in \mathbb{N}\}$ is not closed.

More generally, we have the following characterization of closed sets in terms of sequences:

Proposition 7.24. *A subset X of $\mathbb{R}^n$ is closed if and only if, for every sequence $\{\mathbf{v}_k\}$ such that $\mathbf{v}_k \in X$ for all k, if $\lim_{k\to\infty} \mathbf{v}_k$ exists, then $\lim_{k\to\infty} \mathbf{v}_k \in X$.*

Proof. First suppose that X is closed and that $\{\mathbf{v}_k\}$ is a convergent sequence such that $\mathbf{v}_k \in X$ for all k. Let $\lim_{k\to\infty} \mathbf{v}_k = \mathbf{L}$. Thus, for all $\epsilon > 0$, there exists a K such that, if $k \geq K$, then $N(\mathbf{v}_k - \mathbf{L}) < \epsilon$. We suppose that $\mathbf{L} \notin X$ and shall reach a contradiction. If $\mathbf{L} \notin X$, then $\mathbf{L} \in \mathbb{R}^n - X$, which is open. Thus there is an open ball $B_r(\mathbf{L})$ contained in $\mathbb{R}^n - X$. This says that if $N(\mathbf{x} - \mathbf{L}) < r$, then $\mathbf{x} \notin X$. In other words, if $\mathbf{x} \in X$, then $N(\mathbf{x} - \mathbf{L}) \geq r$. Applying this to $\mathbf{x} = \mathbf{v}_k$, it follows that $N(\mathbf{v}_k - \mathbf{L}) \geq r$ for every k. But then $\mathbf{v}_k$ cannot converge to $\mathbf{L}$: we violate the definition with $\epsilon = r$. This is a contradiction. So $\mathbf{L} \in X$.

Conversely suppose that, for for every sequence $\{\mathbf{v}_k\}$ such that $\mathbf{v}_k \in X$ for all k, if $\lim_{k\to\infty} \mathbf{v}_k$ exists, then $\lim_{k\to\infty} \mathbf{v}_k \in X$. Let $\mathbf{w} \in \mathbb{R}^n - X$. We must show that there exists an open ball of radius r about $\mathbf{w}$ contained

in $\mathbb{R}^n - X$ for some $r > 0$. Suppose not. Taking for example $r = 1/k$ for a positive integer k, this says that there must exist a $\mathbf{x} \in X$ such that $N(\mathbf{x}-\mathbf{w}) < 1/k$. For each k choose $\mathbf{v}_k$ to be such an element $\mathbf{x} \in X$. Thus for every positive integer k, we have chosen $\mathbf{v}_k \in X$ such that $N(\mathbf{v}_k - \mathbf{w}) < 1/k$. Then the sequence $\mathbf{v}_k$ converges to $\mathbf{w}$. In this case by assumption $\mathbf{w} \in X$, contrary to the choice of $\mathbf{w}$. So it must be that $\mathbb{R}^n - X$ is open, or in other words X is closed. □

Warning: a subset of $\mathbb{R}^n$ is open if and only if its complement is closed. But in general a subset of $\mathbb{R}^n$ is neither open nor closed. For example, $(3, 5) \cup \{0\}$ is not an open or closed subset of $\mathbb{R}$. A half open interval is neither open nor closed. The set of all rational numbers is not an open or closed subset of $\mathbb{R}$. The set $\{(x_1, 0) : 0 < x_1 < 1\}$ is not an open or closed subset of $\mathbb{R}^2$.

7.3 Continuous functions in one variable

We turn now to the definition of continuous functions. We begin with ordinary, real-valued functions $f\colon \mathbb{R} \to \mathbb{R}$. As usual in calculus, we can identify the function with its graph and define (roughly) a continuous function to be one whose graph you can draw "without lifting pencil from paper." We begin with the definition of a limit:

Definition 7.25. Let $f\colon I \to \mathbb{R}$ be a function, where I is an interval in $\mathbb{R}$, and let $a \in I$ be an interior point (i.e. not an endpoint). Then

$$\lim_{x \to a} f(x) = L$$

if, given $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $x \in I$, $|x - a| < \delta$ and $x \neq a$ (in other words, $0 < |x - a| < \delta$). Left and right hand limits $\lim_{x \to a^-} f(x)$ and $\lim_{x \to a^+} f(x)$ are similarly defined, where for $\lim_{x \to a^-} f(x)$ we just consider those $x \in I$ with $x < a$ and for $\lim_{x \to a^+} f(x)$ we just consider those $x \in I$ with $x > a$.

The understanding here, as with sequences, is that the interesting choices of ϵ are small positive numbers. In this case, the definition means the following: given a measure of how close we want the distance from $f(x)$ to L to be (the number ϵ), as long as x is sufficiently close to a (distance no more that δ), then indeed $f(x)$ will be within distance ϵ of L. If δ works for a given ϵ, then any positive real number smaller than δ also works.

Note that, in the above definition, we do not allow $x = a$. Thus, it does not matter if $f(a) = L$, or even if $f(x)$ is defined at $x = a$, and we can think of the limit as really defined for a function $f\colon I - \{a\} \to \mathbb{R}$, where a is an interior point of the interval I. Similarly, left or right hand limits are defined for $f\colon (c, a) \to \mathbb{R}$ or $f\colon (a, c) \to \mathbb{R}$.

Next we define continuous functions.

Definition 7.26. Let $f\colon I \to \mathbb{R}$ be a function, where I is an interval, and let a be an interior point of I. We say that f is *continuous at* a if $\lim_{x\to a} f(x)$ exists and is equal to $f(a)$. Finally f is *continuous* or *continuous on* I if it is continuous at a for every $a \in I$. Continuity for functions defined on a closed or half open interval I is defined similarly, using left or right hand limits at the endpoints.

It is easy to see the following:

Lemma 7.27. *The function $f\colon I \to \mathbb{R}$ is continuous at a if and only if, given $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(a)| < \epsilon$ for all $x \in I$ such that $|x - a| < \delta$.*

The general meaning of the continuity of f at a is as follows: the value of f at x is close to $f(a)$ as long as x is close to a. In other words, the constant function $f(a)$ is a reasonable approximation to $f(x)$, at least for x close to a. Here, close, reasonable, etc. are given more precise meaning by the numbers ϵ and δ.

Here are some of the basic examples of continuous functions:

1. A constant function $f(x) = c$ is continuous. Here $|f(x)-f(a)| = 0 < \epsilon$ for every positive ϵ, and any choice of δ will do.

2. The identity function $f(x) = x$ is continuous. Here $|f(x) - f(a)| = |x - a|$. Given $\epsilon > 0$, choose $\delta = \epsilon$. Thus if $|x - a| < \delta$, then $|f(x) - f(a)| = |x - a| < \delta = \epsilon$.

3. The absolute value function $f(x) = |x|$ is continuous. This follows from the inequality $||x| - |a|| \leq |x - a|$ and taking $\delta = \epsilon$ as in the previous example.

Some familiar properties of continuous functions are given below:

Proposition 7.28. *Let $f, g\colon I \to \mathbb{R}$ be two functions, and let $a \in I$. Suppose that $\lim_{x\to a} f(x) = L$ and that $\lim_{x\to a} g(x) = M$. Then:*

1. $\lim_{x\to a}(f(x)+g(x)) = L+M$. *Thus if f and g are continuous at a, so is $f+g$.*

2. $\lim_{x\to a}(f(x)g(x)) = LM$. *Thus if f and g are continuous at a, so is fg.*

3. *If $M \neq 0$, then $\lim_{x\to a}(f(x)/g(x)) = L/M$, in the sense that, if $x \in I$ is in some open interval containing a, then $g(x) \neq 0$ and the quotient is therefore defined, possibly in a smaller interval, and has the limit L/M. Thus if f and g are continuous at a, then f/g is defined in a subinterval of I containing a and it is continuous at a.*

The proofs are minor variations on the proofs of the corresponding statements for sequences.

Corollary 7.29. 1. *Let $p(x) = \sum_{i=0}^{n} a_i x^i$ be a polynomial. Then the function $p\colon \mathbb{R} \to \mathbb{R}$ is continuous.*

2. *Let $r(x) = p(x)/q(x)$ be a rational function, i.e. a quotient of two polynomials $p(x)$, $q(x)$, where $q(x)$ is not identically zero, and let I be an interval such that $q(x) \neq 0$ for all $x \in I$. Then the function $r(x)\colon I \to \mathbb{R}$ is continuous.* □

One additional operation we can do with functions which we could not really do with sequences is compose them. We then have:

Proposition 7.30. *Let $f\colon I \to \mathbb{R}$ be a function and let $a \in I$. Let g be a function defined in an interval containing the image of f. Suppose that f is continuous at a and that g is continuous at $f(a)$. Then $g \circ f$ is continuous at a.*

Proof. Given $\epsilon > 0$, there exists a $\delta_1 > 0$ such that, if $|y - f(a)| < \delta_1$, then $|g(y) - g(f(a))| < \epsilon$. Applying the definition of continuity to f at a and using δ_1 in place of ϵ, there exists a δ such that, if $|x - a| < \delta$, then $|f(x) - f(a)| < \delta_1$. Thus, if $|x - a| < \delta$, then $|g(f(x)) - g(f(a))| < \epsilon$. It follows that $g \circ f$ is continuous at a. □

A very similar argument, which is left as an exercise, shows the following:

Proposition 7.31. *Let $f\colon I \to \mathbb{R}$ be a function which is continuous at $a \in I$. Suppose that $\{x_n\}$ is a sequence, with $x_n \in I$ for all n, such that $\lim_{n\to\infty} x_n = a$. Then $\lim_{n\to\infty} f(x_n) = f(a)$.* □

We come now to the two fundamental theorems concerning continuous functions.

Theorem 7.32 (Intermediate value theorem)**.** *Let I be an interval and let $f\colon I \to \mathbb{R}$ be a continuous function. Suppose $a, b \in I$ with $a \leq b$ and that $f(a) \leq f(b)$. Let c be a real number such that $f(a) \leq c \leq f(b)$. Then there exists an x with $a \leq x \leq b$ such that $f(x) = c$.*

Pictorially, this says that the horizontal line $y = c$ must cross the graph of $f(x)$.

Of course, there is a similar statement if instead we assume that $f(a) \geq f(b)$ and that $f(a) \geq c \geq f(b)$.

Theorem 7.33 (Extreme value theorem)**.** *Let $I = [a, b]$ be a closed, bounded, and nonempty interval and let $f\colon I \to \mathbb{R}$ be a continuous function. Then the image $f(I)$ is bounded. Moreover, there is an $x_0 \in I$ such that $f(x_0)$ is the greatest lower bound of $f(I)$, in other words $f(x) \geq f(x_0)$ for all $x \in I$, and similarly there is an $x_1 \in I$ such that $f(x_1)$ is the least upper bound of $f(I)$, in other words $f(x) \leq f(x_1)$ for all $x \in I$.*

We can summarize the extreme value theorem by saying that f has a maximum and a minimum value on I.

Notice that both theorems assert that some number exists, but do not in general tell us how to find this number—this question can usually only be answered explicitly, if it can be answered at all, via calculus. Also, the x in the intermediate value theorem and the x_0, x_1 in the extreme value theorem need not in general be unique.

Let us prove the extreme value theorem–we will discuss the intermediate value theorem later. The main ingredients will be the Bolzano-Weierstrass theorem and the existence of least upper bounds. Suppose that $I = [a, b]$ is closed and nonempty. The proof of the extreme value theorem proceeds in two steps:

Step I. We begin by showing that f is bounded on I, i.e. that the image $f(I)$ is bounded. Otherwise, for each $n \in \mathbb{N}$ there exists an $x_n \in I$ such that $|f(x_n)| \geq n$. Since $x_n \in [a, b]$, the sequence $\{x_n\}$ is bounded. Thus, by the Bolzano-Weierstrass theorem, there is a convergent subsequence $\{x_{n_k}\}$. Let $L = \lim_{k\to\infty} x_{n_k}$. Since $[a, b]$ is closed, it contains all limits of sequences whose terms lie in I, and so $L \in I$. Now $\{f(x_{n_k})\}$ converges to $f(L)$ and is therefore bounded. But by construction $|f(x_{n_k})| \geq n_k \geq k$, and thus the sequence $\{f(x_{n_k})\}$ is unbounded. This is a contradiction. It follows that f is bounded.

Step II. Now let us show that f attains its maximum value. Since $f(I)$ is bounded, and by hypothesis it is nonempty, there exists a least upper

bound M for $f(I)$. We shall show that $M = f(t)$ for some $t \in I$. It then follows that $f(x) \leq f(t) = M$ for all $x \in I$, so that M is in fact the maximum value of f. To see this, use the fact that, for all $n \in \mathbb{N}$, $M - 1/n$ is not an upper bound for $f(I)$. Thus, for all n, there exists $y_n \in f(I)$ such that $M - 1/n < y_n \leq M$. Clearly the sequence $\{y_n\}$ converges to M. By definition there exists an $x_n \in I$ such that $y_n = f(x_n)$. Again by applying the Bolzano-Weierstrass theorem, there is a convergent subsequence $\{x_{n_k}\}$ of x_n. Suppose that $\lim_{k\to\infty} x_{n_k} = t$. As above, $t \in I$. Moreover $\lim_{k\to\infty} f(x_{n_k}) = f(t)$. On the other hand, $f(x_{n_k}) = y_{n_k}$ and so the sequence $\{f(x_{n_k})\}$ is a subsequence of the convergent sequence $\{y_n\}$. But every subsequence of a convergent sequence converges, and to the same limit. Thus

$$f(t) = \lim_{k\to\infty} f(x_{n_k}) = \lim_{k\to\infty} y_{n_k} = \lim_{n\to\infty} y_n = M.$$

It follows that $M = f(t)$ as claimed.

The proof that f also attains its greatest lower bound is similar. □

7.4 Functions of several variables

We turn now to functions of several variables. As before, it will be simplest to begin first with functions defined on all of $\mathbb{R}^n$. We will also look at functions just defined on a subset X of $\mathbb{R}^n$, which will usually be open or (less often) closed. We can consider real-valued functions $f\colon \mathbb{R}^n \to \mathbb{R}$, or more generally vector-valued functions $F\colon \mathbb{R}^n \to \mathbb{R}^m$. Such a function F can always be written as $F = (f_1, \ldots, f_m)$, where each $f_i\colon \mathbb{R}^n \to \mathbb{R}$ is a real-valued function. In this way, many questions concerning vector-valued functions can be reduced to questions concerning real-valued functions, of several variables. However, there is not usually any way to reduce questions about functions of several variables to functions of just one variable.

In general, it is hard, if not impossible, to visualize functions of several variables. If $f\colon \mathbb{R}^2 \to \mathbb{R}$ is a real-valued function of two variables, then its graph is a subset of $\mathbb{R}^3$. If we visualize $\mathbb{R}^3$ in the usual way, by letting the xy-plane be horizontal, then we can think of the z-coordinate as height above the xy-plane (positive or negative), and imagine $f(x, y)$ as giving the height of the point on the graph above (x, y). Thus the graph will cover $\mathbb{R}^2$, and intuitively this says that the graph is a two-dimensional object–a surface. Another way to visualize the graph is by drawing the family of *level sets* or *level curves* $\{(x, y); f(x, y) = c\}$ for various values of c. For a fixed c, the set $\{(x, y); f(x, y) = c\}$ consists of all the points on the graph

at height (or level) c, and is the set of (x, y)-coordinates of the intersection of the graph with the horizontal plane $z = c$. We can think of the level sets as giving a contour map of the graph. In particular, when level sets bunch close together (for evenly spaced choices of c), it means that the graph is steep.

Of course, we can define graphs and level sets in general but are usually not able to draw them. If $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is a function, which we also write as $(f_1, \ldots, f_m)$, then its graph lies in $\mathbb{R}^{n+m}$. Given $\mathbf{c} = (c_1, \ldots, c_m) \in \mathbb{R}^m$, the level set is the preimage

$$F^{-1}(\mathbf{c}) = \{\mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) = \mathbf{c})\}.$$

This is of course just the set of $\mathbf{x}$ such that the system of equations $f_1(\mathbf{x}) = c_1, \ldots, f_m(\mathbf{x}) = c_m$ is satisfied. If F is linear and $\mathbf{c} = \mathbf{0}$, then the level set $F^{-1}(\mathbf{0})$ is just the kernel of F, which is a vector subspace V of $\mathbb{R}^n$, and more generally $F^{-1}(\mathbf{c})$ is either empty or an affine subspace of $\mathbb{R}^n$ parallel to V (in other words, it is of the form $\mathbf{x}_0 + V$, where $\mathbf{x}_0$ is a particular solution to the equation $F(\mathbf{x}) = \mathbf{c}$).) Just as in linear algebra, for a general function F, we are interested in the level sets $F^{-1}(\mathbf{c})$ both as solutions to systems of equations and as geometric objects in their own right. Note that either the level sets $F^{-1}(\mathbf{c})$ and $F^{-1}(\mathbf{c}')$ are disjoint or $\mathbf{c} = \mathbf{c}'$ and thus $F^{-1}(\mathbf{c}) = F^{-1}(\mathbf{c}')$. This is simply a reflection of the fact that F is a function: if $\mathbf{x} \in F^{-1}(\mathbf{c}) \cap F^{-1}(\mathbf{c}')$, then $F(\mathbf{x}) = \mathbf{c} = \mathbf{c}'$. Of course, the only time we can draw level sets of a real-valued function $f\colon \mathbb{R}^n \to \mathbb{R}$, say, is when $n = 2$, which we have discussed above, or $n = 3$, in which case level sets are often called *level surfaces*.

One final type of function which we can visualize is at the opposite extreme from a real-valued function $f\colon \mathbb{R}^n \to \mathbb{R}$, namely a function $\mathbf{r}\colon I \to \mathbb{R}^m$, where I is an interval, which is a vector-valued function of a single variable. Writing $\mathbf{r}(t) = (x_1(t), \ldots, x_m(t))$, we say that $\mathbf{r}$ is a *path* or *parametrized curve* if all of the $x_i(t)$ are continuous. In this case, we often think of the parameter t as time and imagine g as tracing out a curve as a function of t. If $I = [a, b]$ is a closed interval, then the point $\mathbf{r}(a)$ is the *starting point* of the path and the point $\mathbf{r}(b)$ is the *endpoint*.

Keeping the above in mind, let's give some examples of graphs and level sets for functions of two variables.

$f(x, y) = ax + by$. The graph is $\{(z, y, z) \in \mathbb{R}^3 : z = ax + by\}$ and is thus a plane. It intersects the xz-plane in the line $z = ax$ of slope a through the origin, and the yz-plane in the line $z = by$. The level sets are the parallel lines $ax + by = c$ and they are uniformly spaced for uniform spacing of the

c, in other words, the distance between two level sets corresponding to two different levels c and c' only depends on the distance from c to c'.

$f(x, y) = x^3$. In the xz-plane, the graph is just the graph of the function $z = x^3$. In fact the same is true in every plane $y = \text{const.}$ parallel to the xz-plane, and so we can imagine the graph is given by sliding the graph of $z = x^3$ along the y-axis. The level sets are $x^3 = c$, or $x = c^{1/3}$. Thus they are all vertical lines, but they bunch together for increasing c.

$f(x, y) = x^2 + y^2$. In the xz-plane, the graph is the graph of the function $z = x^2$ and is thus a parabola. Since z only depends on $x^2 + y^2 = \|(x, y)\|^2$, the value of the function is unchanged along any circle whose center is the origin. Thus the graph has rotational symmetry, and is obtained by rotating a parabola in the xz-plane about the z-axis. The level curves at height $c > 0$ are all circles about the origin, reflecting the rotational symmetry, and of radius $\sqrt{c}$–thus they bunch up for $c > 1$, reflecting the steepness. For $c = 0$ the "level curve" is just the origin, and for $c < 0$ it is empty.

$f(x, y) = xy$. Here there is no obvious symmetry or way to simplify the picture down to a function of one variable, and the intersection of the graph with the xz-plane and the yz-plane both give the unilluminating graph $z = 0$. On the other hand, if we look at the graph above the lines through the origin $y = tx$, we get the parabola $z = tx^2 = \frac{1}{t}y^2$, if $t \neq 0$. For $t > 0$, as t increases from 0 to 1, i.e. as the line rotates from the x-axis, these become a moving family of progressively steeper parabolas. As we keep rotating to a vertical line, the family of parabolas becomes more shallow again, and we get the y-axis on the yz-plane. Continuing to rotate means that $t < 0$, so that the parabolas open downward. In terms of level curves, the level sets $xy = c$ are a hyperbola (in two pieces), in the first and third quadrant for $c > 0$ and the second and fourth quadrant for $c < 0$, and are a union of the x and y-axes for $c = 0$. It is a good idea to try to visualize this graph, to see that it is both increasing at the origin, and to see why it is called a saddle.

As we saw above, there are no very systematic ways to try and understand the graph of a function $z = f(x, y)$, even ones with very simple formulas such as xy. General suggestions are: look for symmetry, see if the behavior reduces somehow to a function of one variable, draw the intersections with the xz-plane and the yz-plane, look at the behavior of the function over certain lines or other simple curves (such as circles) in the xy-plane, try to find some representative level curves. Beyond that, each function f must be approached individually.

Next we turn to the definition of a continuous function. Let $F\colon X \to \mathbb{R}^m$ be a function, where X is a subset of $\mathbb{R}^n$ for some n. The most interesting

examples will be $X = \mathbb{R}^n$ or an open subset of $\mathbb{R}^n$. A function F from X to $\mathbb{R}^m$ is given by m functions from X to $\mathbb{R}$: $F(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$. We can carry over the usual ϵ, δ definition of continuity. For simplicity, we will just write this out for the usual norm $\|\cdot\|$ on $\mathbb{R}^n$ and $\mathbb{R}^m$. In general, to be on the safe side, we would have to choose **two** norms, one for $\mathbb{R}^n$ and another for $\mathbb{R}^m$. Of course, since all norms are equivalent it won't matter. With that said, we can essentially copy the definitions in one variable.

Definition 7.34. Let $F\colon \mathbb{R}^n \to \mathbb{R}^m$ be a function. Then $\lim_{\mathbf{x}\to\mathbf{a}} F(\mathbf{x}) = \mathbf{L}$ if, for all $\epsilon > 0$, there exists $\delta > 0$ such that $\|F(\mathbf{x}) - \mathbf{L}\| < \epsilon$ for all $\mathbf{x}$ such that $0 < \|\mathbf{x} - \mathbf{a}\| < \delta$.

In case F is not defined on all of $\mathbb{R}^n$, we shall just consider the case where X is an **open** subset of $\mathbb{R}^n$, $\mathbf{a} \in X$, and the domain of F contains $X - \{\mathbf{a}\}$. The definition is the same as above except that we require in addition that $\mathbf{x} \in X$.

The function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is *continuous at* $\mathbf{a}$ if $\lim_{\mathbf{x}\to\mathbf{a}} F(\mathbf{x}) = F(\mathbf{a})$, or equivalently if, for every $\epsilon > 0$, there exists $\delta > 0$ such that $\|F(\mathbf{x}) - F(\mathbf{a})\| < \epsilon$ if $\|\mathbf{x} - \mathbf{a}\| < \delta$. The function F is *continuous* if it is continuous at every point.

If F is just defined on an (arbitrary) subset X of $\mathbb{R}^n$, we say that F is *continuous at* $\mathbf{a} \in X$ if, for every $\epsilon > 0$, there exists $\delta > 0$ such that $\|F(\mathbf{x}) - F(\mathbf{a})\| < \epsilon$ if $\|\mathbf{x} - \mathbf{a}\| < \delta$ and $\mathbf{x} \in X$, and that F is *continuous* if it is continuous at every point of X. In case $n = 1$ and I is a closed or half open interval, it is easy to see that this reduces to the usual definition in terms of left and right hand limits.

Here are some easy examples of continuous functions:

1. Constant functions;
2. For $n = m$, the function $F(\mathbf{x}) = \mathbf{x}$ (i.e. $F = \mathrm{Id}$);
3. The coordinate functions $F(\mathbf{x}) = x_i$ (to make $|x_i - a_i| < \epsilon$, use the $\|\cdot\|_\infty$ norm and the fact that $|x_i - a_i| \le \|\mathbf{x} - \mathbf{a}\|_\infty$. So to make $|x_i - a_i| < \epsilon$, it suffices to take $\|\mathbf{x} - \mathbf{a}\|_\infty < \epsilon$);
4. The function $f(\mathbf{x}) = \|\mathbf{x}\|$ (use the corollary of the triangle inequality which says: $|\|\mathbf{x}\| - \|\mathbf{a}\|| \le \|\mathbf{x} - \mathbf{a}\|$).

Before we state the next result, let us review the following terminology. Let $F\colon X \to Y$ be a function from a set X to a set Y. The *image* of F is the set

$$\{y \in Y : y = F(x) \text{ for some } x \in X\}.$$

Call this set $F(X)$. Similarly, if A is a subset of X we let

$$F(A) = \{y \in Y : y = F(x) \text{ for some } x \in A\}.$$

If B is a subset of Y, we define the *preimage* or *inverse image* $F^{-1}(B)$ as follows:

$$F^{-1}(B) = \{x \in X : F(x) \in B\}.$$

Thus we always have $F(F^{-1}(B)) \subseteq B$ and $A \subseteq F^{-1}(F(A)$ but it may not be true that $F(F^{-1}(B)) = B$ or that $F^{-1}(F(A)) = A$. Of course, if F is onto then $F(F^{-1}(B)) = B$, and if F is one-to-one then $F^{-1}(F(A)) = A$. In general, it is easy to check that $F^{-1}(A \cup B) = F^{-1}(A) \cup F^{-1}(B)$ and that $F^{-1}(A \cap B) = F^{-1}(A) \cap F^{-1}(B)$.

We have the following set of equivalent characterizations of continuous functions:

Proposition 7.35. *Let $F\colon \mathbb{R}^n \to \mathbb{R}^n$ be a function. Then the following are equivalent:*

1. *F is continuous.*
2. *For all open subsets X of $\mathbb{R}^m$, $F^{-1}(X)$ is an open subset of $\mathbb{R}^n$.*
3. *For all closed subsets X of $\mathbb{R}^m$, $F^{-1}(X)$ is a closed subset of $\mathbb{R}^n$.*
4. *For all sequences $\{\mathbf{v}_k\}$ in $\mathbb{R}^n$, if $\{\mathbf{v}_k\}$ converges in $\mathbb{R}^n$ to a limit $\mathbf{w}$, then the sequence $\{F(\mathbf{v}_k)\}$ converges in $\mathbb{R}^m$ to $F(\mathbf{w})$.*

Proof. (1) $\implies$ (2): Say that F is continuous and that X is an open subset of $\mathbb{R}^m$. Let $\mathbf{v} \in F^{-1}(X)$, so that $\mathbf{x} = F(\mathbf{v}) \in X$. There exists an $\epsilon > 0$ such that $B_\epsilon(\mathbf{x}) \subseteq X$. For this ϵ, there is a $\delta > 0$ such that $\|\mathbf{w} - \mathbf{v}\| < \delta$ implies that

$$\|F(\mathbf{w}) - F(\mathbf{v})\| = \|F(\mathbf{w}) - \mathbf{x}\| < \epsilon.$$

This says that if $\mathbf{w} \in B_\delta(\mathbf{v})$, then $F(\mathbf{w}) \in B_\epsilon(\mathbf{x}) \subseteq X$. In other words, for every $\mathbf{v} \in F^{-1}(X)$, $F^{-1}(X)$ contains some ball $B_\delta(\mathbf{v})$ around $\mathbf{v}$. Thus $F^{-1}(X)$ is open.

(2) $\implies$ (1): Suppose that $\mathbf{v} \in \mathbb{R}^n$. Given $\epsilon > 0$, the ball $B_\epsilon(F(\mathbf{v}))$ is open in $\mathbb{R}^m$. Thus $F^{-1}(B_\epsilon(F(\mathbf{v})))$ is open in $\mathbb{R}^n$. Since $\mathbf{v} \in F^{-1}(B_\epsilon(F(\mathbf{v})))$ (as $F(\mathbf{v}) \in B_\epsilon(F(\mathbf{v}))$), there exists a $\delta > 0$ such that $B_\delta(\mathbf{v}) \subseteq F^{-1}(B_\epsilon(F(\mathbf{v})))$. This exactly says that if $\|\mathbf{x} - \mathbf{v}\| < \delta$, then $\|F(\mathbf{x}) - F(\mathbf{v})\| < \epsilon$, so that F is continuous at $\mathbf{v}$.

(2) $\iff$ (3): This follows from $F^{-1}(\mathbb{R}^m - X) = \mathbb{R}^n - F^{-1}(X)$.

(1) $\implies$ (4): Suppose that $\{\mathbf{v}_k\}$ converges to $\mathbf{w}$. We wish to prove that $\{F(\mathbf{v}_k)\}$ converges to $F(\mathbf{w})$. Given $\epsilon > 0$, there exists $\delta > 0$ such that $\|\mathbf{x}-\mathbf{w}\| < \delta$ implies that $\|F(\mathbf{x})-F(\mathbf{v})\| < \epsilon$. Given δ, there exists a K such that $k \geq K$ implies that $\|\mathbf{v}_k - \mathbf{w}\| < \delta$. So $\|F(\mathbf{v}_k) - F(\mathbf{w})\| < \epsilon$ as long as $k \geq K$ and hence $\{F(\mathbf{v}_k)\}$ converges to $F(\mathbf{w})$.

(4) $\implies$ (1): It suffices to prove the contrapositive. If F is not continuous at some $\mathbf{w}$, then there exists an $\epsilon > 0$ such that, for every $\delta > 0$, there exists a $\mathbf{x}$ with $\|\mathbf{x}-\mathbf{w}\| < \delta$ but $\|F(\mathbf{x})-F(\mathbf{w})\| \geq \epsilon$. Thus, taking $\delta = 1/k$, we can find a $\mathbf{v}_k$ with $\|\mathbf{v}_k - \mathbf{w}\| < 1/k$ but $\|F(\mathbf{v}_k) - F(\mathbf{w})\| \geq \epsilon$ for all k. The first inequality says that $\mathbf{v}_k$ converges to $\mathbf{w}$ and the second says that $F(\mathbf{v}_k)$ does not converge to $F(\mathbf{w})$. □

There are similar results for functions defined on a subset X of $\mathbb{R}^n$, not necessarily open nor closed. The proof of the following is an easy modification of the above:

Proposition 7.36. *Let $F\colon X \to \mathbb{R}^n$ be a function. Then the following are equivalent:*

1. *F is continuous.*
2. *For all open subsets U of $\mathbb{R}^m$, $F^{-1}(U)$ is the intersection of an open subset A of $\mathbb{R}^n$ with X.*
3. *For all closed subsets V of $\mathbb{R}^m$, $F^{-1}(V)$ is the intersection of a closed subset B of $\mathbb{R}^n$ with X.*
4. *For all sequences $\{\mathbf{v}_k\}$ in X, if $\{\mathbf{v}_k\}$ converges in $\mathbb{R}^n$ to a limit $\mathbf{w} \in X$, then the sequence $\{F(\mathbf{v}_k)\}$ converges in $\mathbb{R}^m$ to $F(\mathbf{w})$.*

Note that, if X above is open, then Condition (2) just says that, for all open subsets U of $\mathbb{R}^m$, $F^{-1}(U)$ is open. On the other hand, if X is closed, then Condition (3) just says that, for all closed subsets V of $\mathbb{R}^m$, $F^{-1}(V)$ is closed.

We also have the following, which follows easily from (4) and Lemma 7.14, or by a direct argument:

Proposition 7.37. *Let $F\colon X \to \mathbb{R}^m$ be a function, with*

$$F(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})).$$

Then F is continuous if and only if $f_i(\mathbf{x})$ is continuous for all i. □

Continuous functions of several variables have properties similar to continuous real-valued functions of one variable:

Proposition 7.38. *1. If $F\colon X \to \mathbb{R}^m$ and $G\colon X \to \mathbb{R}^m$ are continuous, so is $F + G$.*

2. If $F\colon X \to \mathbb{R}^m$ and $g\colon X \to \mathbb{R}$ are continuous, so is gF (the function from X to $\mathbb{R}^m$ obtained by scalar multiplication of g with the vector valued function F).

3. If $F\colon X \to \mathbb{R}^m$ and $G\colon X \to \mathbb{R}^m$ are continuous, so is $\langle F, G\rangle$ (the function from X to $\mathbb{R}^m$ obtained by taking the inner product of the vector valued functions F and G).

4. If $g\colon X \to \mathbb{R}$ is continuous and $g(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in X$, then $1/g$ is continuous.

5. If $F\colon X \to \mathbb{R}^m$ is continuous, if G is defined on $F(X)$ and $G\colon F(X) \to \mathbb{R}^k$ is continuous, then $G \circ F\colon X \to \mathbb{R}^k$ is continuous.

The proofs are similar to those in the one-variable case.

Using the above, we can construct lots of continuous functions and use them to define lots of open sets. For example, the i coordinate functions $p_i(\mathbf{x}) = x_i$ are continuous. So all linear functions $f\colon \mathbb{R}^n \to \mathbb{R}$ are continuous, since $f(x_1, \ldots, x_n) = \sum_{i=1}^n a_i x_i$ is a sum of coordinate functions times constants. Similarly a linear function $F\colon R^n \to \mathbb{R}^m$ is continuous. Likewise all polynomials in the x_i are continuous, also all rational functions except where the denominator is zero. Taking compositions, we see that every expression of the form $f(F(x_1, \ldots, x_n))$ is continuous provided that f and F are continuous. For example,

$$e^{\frac{x_1 x_2 - 3\sin x_2}{1+x_1^2+x_2^2}}$$

defines a continuous function. Once we have a large stock of continuous functions, we can show that many sets are open by writing them as $f^{-1}(X)$ where X is an open subset of $\mathbb{R}$ and $f\colon \mathbb{R}^n \to \mathbb{R}$ is continuous, and similarly for $X \subseteq \mathbb{R}^m$. For example, given $\mathbf{v} \in \mathbb{R}^n$, let $f\colon \mathbb{R}^n \to \mathbb{R}$ be the function $f(\mathbf{x}) = \|\mathbf{v} - \mathbf{x}\|$. Then $B_r(\mathbf{v}) = f^{-1}(-\infty, r)$, $\bar{B}_r(\mathbf{v}) = f^{-1}(-\infty, r]$, and $S_r(\mathbf{v}) = f^{-1}(r)$. We see (again) that the first set is open and the other two are closed. Likewise

$$\{(x_1, x_2, x_3) : 5x_1 + 2x_2 - x_3 = 12\}.$$

is $A^{-1}(12)$, where $A\colon \mathbb{R}^3 \to \mathbb{R}$ is defined by $A(\mathbf{x}) = 5x_1 + 2x_2 - x_3$. Finally

$$\{(x_1, x_2, x_3) : x_3 \geq x_1^2 + x_2^2\}$$

is equal to $f^{-1}([0, \infty))$, where $f(\mathbf{x}) = x_3 - x_1^2 + x_2^2$, and is therefore closed. More generally, if $f_1, \ldots, f_m\colon \mathbb{R}^n \to \mathbb{R}$ are continuous, then a set defined via the f_i and non-strict inequalities such as

$$\{\mathbf{x} \in \mathbb{R}^n : f_1(\mathbf{x}) \leq c_1, \ldots, f_m(\mathbf{x}) \leq c_m\}$$

is closed, whereas the set

$$\{\mathbf{x} \in \mathbb{R}^n : f_1(\mathbf{x}) < c_1, \ldots, f_m(\mathbf{x}) < c_m\}$$

is open.

7.5 Compact sets

Our final goal in this chapter will be to generalize the intermediate value theorem and the extreme value theorem to continuous functions in several variables. We begin with the extreme value theorem. Going over the proof of the extreme value theorem given above, we see that the main point, aside from the existence of least upper bounds and the fact that the image under a continuous function of a convergent sequence is convergent, is the Bolzano-Weierstrass theorem. We thus make this into a definition:

Definition 7.39. A subset X of $\mathbb{R}^n$ is *compact* if, for every sequence $\{\mathbf{v}_k\}$ with $\mathbf{v}_k \in X$ for all k, there exists a subsequence $\{\mathbf{v}_{k_i}\}$ which converges to a limit $\mathbf{w} \in X$. (**Note:** that we require not only that the limit exist, but that it lie in X as well.)

Theorem 7.40. *If X is a nonempty compact subset of $\mathbb{R}^n$, then every continuous function $f\colon X \to \mathbb{R}$ satisfies the extreme value theorem: there exists an $\mathbf{x} \in X$ such that*

$$f(\mathbf{x}) = \sup\{f(\mathbf{y}) : \mathbf{y} \in X\}$$

and similarly for the minimum value. □

The proof follows the proof of the extreme value theorem given above.

Of course, the theorem above is useless if we don't know how to find compact sets. Clearly the empty set is compact, also a set consisting of one element, or finitely many. On the other hand, $\mathbb{R}^n$ is not compact, since for example the sequence $\{(n, 0, \ldots, 0)\}$ does not have a convergent

subsequence. Likewise the open ball B_1 of radius 1 about the origin is not compact: while the sequence $\{(1-1/n, 0, \ldots, 0)\}$ converges, its limit is not in B_1. The following gives a more explicit characterization of compact sets:

Theorem 7.41. *A subset X of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. (Here, by bounded, we mean that there is a $C > 0$ such that $\|\mathbf{x}\| \leq C$ for all $\mathbf{x} \in X$.)*

Proof. First we claim that, if X is compact, then it is closed and bounded. To see that X is closed, let $\{\mathbf{v}_k\}$ be a sequence in X converging to some $\mathbf{w} \in \mathbb{R}^n$. We must show that $\mathbf{w} \in X$. But $\{\mathbf{v}_k\}$ has a subsequence converging to some limit $\mathbf{w}' \in X$, by the definition of compactness. On the other hand, since $\{\mathbf{v}_k\}$ is convergent, every subsequence is convergent to the same limit. Hence $\mathbf{w}' = \mathbf{w}$ and in particular $\mathbf{w} \in X$. Thus X is closed.

Next we claim that X is bounded. If not, for each positive integer k there exists a $\mathbf{v}_k$ such that $\|\mathbf{v}_k\| \geq k$. But then the sequence $\{\mathbf{v}_k\}$ has no convergent subsequence: if $\{\mathbf{v}_{k_i}\}$ is a subsequence which converges to some limit $\mathbf{w} \in \mathbb{R}^n$, then there is an integer i_0 such that, for $i \geq i_0$, $\|\mathbf{w} - \mathbf{v}_{k_i}\| < 1$, say, and so, since

$$\|\mathbf{v}_{k_i}\| - \|\mathbf{w}\| \leq |\|\mathbf{v}_{k_i}\| - \|\mathbf{w}\|| \leq \|\mathbf{v}_{k_i} - \mathbf{w}\| = \|\mathbf{w} - \mathbf{v}_{k_i}\| < 1,$$

for all $i \geq i_0$ we have $\|\mathbf{v}_{k_i}\| < \|\mathbf{w}\| + 1$. But also $\|\mathbf{v}_{k_i}\| \geq k_i \geq i$ for all i, which is impossible since the positive integers are not bounded. (We essentially showed above that a convergent sequence is bounded.)

Conversely, suppose that X is closed and bounded. We must show that X is compact. The idea of the proof is to reduce the question to the one-variable Bolzano-Weierstrass theorem. First, to say that X is bounded in some norm implies that X is bounded in any equivalent norm. So we can assume that X is bounded in the $\|\cdot\|_\infty$ norm, in other words that, for all $\mathbf{x} = (x_1, \ldots, x_n) \in X$, $|x_i| \leq C$. Thus X is a subset of the product $[-C, C] \times \cdots \times [-C, C] = [-C, C]^n$ of the closed intervals $[-C, C]$. Let $\{\mathbf{v}_k\}$ be a sequence such that $\mathbf{v}_k \in X$ for every k. Writing $\mathbf{v}_k = (v_{1,k}, \ldots, v_{n,k})$, this says in particular that $|v_{i,k}| \leq C$ for every i, so that the sequences $\{v_{i,k}\}$ are bounded for every i. Begin with $i = 1$, and apply the Bolzano-Weierstrass theorem to the sequence $\{v_{1,k}\}$ to find a convergent subsequence $\{v_{1,k_j}\}$. After relabeling, we can replace the original sequence $\{\mathbf{v}_k\}$ by the subsequence, so that we can asume that $\{v_{1,k}\}$ converges. Now consider the sequence $\{v_{2,k}\}$. Since it is still bounded, it has a convergent subsequence $\{v_{2,k_j}\}$, again by the Bolzano-Weierstrass theorem. Since $\{v_{1,k}\}$ converges, the same is true for the subsequence $\{v_{1,k_j}\}$ converges. So after passing to a

further subsequence, we can ssume that the sequence $\{v_{2,k}\}$ also converges. (Note that we have implicitly used the fact that a subsequence of a subsequence is a subsequence.) Continuing in this way, we can find a subsequence of the original sequence $\{\mathbf{v}_k\}$, say $\{\mathbf{v}_{k_j}\}$, such that $\{v_{i,k_j}\}$ converges for every i. Since every component subsequence converges, so does $\{\mathbf{v}_{k_j}\}$, to some limit $\mathbf{L} \in \mathbb{R}^n$. Since X is closed, $\mathbf{L} \in X$. Hence X is compact. □

Corollary 7.42. *Every continuous function on a closed bounded set has a maximum and a minimum value.* □

Let us give one application: Using the above, we can (finally!) show that all norms in $\mathbb{R}^n$ are equivalent. It is enough to show:

Proposition 7.43. *Every norm on $\mathbb{R}^n$ is equivalent to $\|\cdot\|_1$, and hence every two norms on $\mathbb{R}^n$ are equivalent.*

Proof. Let N be a norm on $\mathbb{R}^n$, and let $C_1 = \max\{N(\mathbf{e}_1), \ldots, N(\mathbf{e}_n)\}$. Then

$$N(\mathbf{x}) = N\left(\sum_{i=1}^{n} x_i \mathbf{e}_i\right) \leq \sum_i N(x_i \mathbf{e}_i) = \sum_i |x_i| N(\mathbf{e}_i) \leq C_1 \sum_i |x_i| = C_1 \|\mathbf{x}\|_1.$$

Conversely, we want to show that there is a constant C_2 such that $\|\mathbf{x}\|_1 \leq C_2 N(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. First, we claim that $N\colon \mathbb{R}^n \to \mathbb{R}$ is continuous with respect to the norm $\|\cdot\|_1$. Let $\mathbf{v} \in \mathbb{R}^n$ and let $\epsilon > 0$ be given, Using the above argument, we have $N(\mathbf{w}-\mathbf{v}) \leq C_1 \|\mathbf{w}-\mathbf{v}\|_1$. Thus, given $\mathbf{v} \in \mathbb{R}^n$ and $\epsilon > 0$, if we set $\delta = \epsilon/C_1$, then for $\|\mathbf{w}-\mathbf{v}\|_1 < \delta$,

$$N(\mathbf{w}-\mathbf{v}) \leq C_1 \|\mathbf{w}-\mathbf{v}\|_1 < C_1 \delta = C_1 \frac{\epsilon}{C_1} = \epsilon.$$

On the other hand we always have the inequality $|N(\mathbf{w}) - N(\mathbf{v})| \leq N(\mathbf{w}-\mathbf{v})$, which as we have seen follows from the triangle inequality, and so

$$|N(\mathbf{w}) - N(\mathbf{v})| < \epsilon.$$

So, for all $\epsilon > 0$, taking $\delta = \epsilon/C_1$, if $\|\mathbf{w}-\mathbf{v}\|_1 < \delta$, then $|N(\mathbf{w}) - N(\mathbf{v})| < \epsilon$, i.e. N is continuous with respect to $\|\cdot\|_1$. In particular consider the function N on the compact set $S_1(0)$, the unit sphere for $\|\cdot\|_1$. Since $S_1(0)$ is closed and bounded, it is compact, and so N has a minimum value D on $S_1(0)$. Also $D > 0$. This says that, for all $\mathbf{x}$ such that $\|\mathbf{x}\|_1 = 1$, $N(\mathbf{x}) \geq D$, or in other words

$$1 = \|\mathbf{x}\|_1 \leq \frac{1}{D} N(\mathbf{x}).$$

Let $C_2 = 1/D$. We claim that $\|\mathbf{x}\|_1 \leq C_2 N(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. This obviously true if $\mathbf{x} = 0$. Otherwise let $t = \|\mathbf{x}\|_1$ and let $\mathbf{y} = \frac{1}{t}\mathbf{x}$. Then $\|\mathbf{y}\|_1 = 1$ and

$$N(\mathbf{y}) = \frac{1}{t} N(\mathbf{x}) \geq D.$$

Thus $t = \|\mathbf{x}\|_1 \leq \frac{1}{D} N(\mathbf{x}) = C_2 N(\mathbf{x})$ for all $\mathbf{x} \neq 0$, and so for all $\mathbf{x}$.

We have thus found constants $C_1, C_2 > 0$ such that, for all $\mathbf{x} \in \mathbb{R}^n$, $N(\mathbf{x}) \leq C_1 \|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_1 \leq C_2 N(\mathbf{x})$. It follows that N and $\|\cdot\|_1$ are equivalent. □

Lastly, we want to give another and very useful characterization of compact sets.

Definition 7.44. If A is a subset of a set X, $U_i, i \in I$ is a collection of subsets of X, and $A \subseteq \bigcup_{i \in I} U_i$, then we call the sets $U_i, i \in I$ a *cover* of A. We will mainly be concerned with the case where $X = \mathbb{R}^n$ and the U_i are open subsets of $\mathbb{R}^n$, in which case we call the sets $U_i, i \in I$ an *open cover* of A. If J is a subset of I such that $A \subseteq \bigcup_{i \in J} U_i$, then we call the sets $U_i, i \in J$ a *subcover* of $U_i, i \in I$ (and an open subcover if the U_i are open sets).

For example, the collection of open intervals $\{(n-1, n+1) : n \in \mathbb{Z}\}$ is an open cover of $\mathbb{R}$, as is the collection $\{(-n, n) : n \in \mathbb{N}\}$. The collection $\{(1/n, 1 - 1/n) : n \in \mathbb{N}\}$ is an open cover of the open interval $(0, 1)$. In neither of these cases can we find a finite subcover, i.e. a finite collection of elements of the cover which again form a cover. The situation is very different if we work with closed (bounded) intervals. More generally, we make the following definition:

Definition 7.45. A subset X of $\mathbb{R}^n$ has the *Heine-Borel property* if every open cover of X has a finite subcover.

Theorem 7.46. *A subset X of $\mathbb{R}^n$ has the Heine-Borel property if and only X is compact.*

Proof. We have seen that a subset X of $\mathbb{R}^n$ is compact $\iff$ X is closed and bounded. We will therefore prove that a subset X satisfying the Heine-Borel property is closed and bounded, and that a compact set X has the Heine-Borel property.

Suppose that X has the Heine-Borel property. First we claim that X is bounded: since $X \subseteq \bigcup_{k \in \mathbb{N}} B_k(\mathbf{0})$, and each ball $B_k(\mathbf{0})$ is open, the collection $\{B_k(\mathbf{0}) : k \in \mathbb{N}\}$ is an open cover of X and hence has a finite subcover

$\{B_{k_1}(\mathbf{0}), \ldots, B_{k_\ell}(\mathbf{0})\}$. Setting $C = \sup\{k_1, \ldots, k_\ell\}$, this says that $X \subseteq \bigcup_{i=1}^{\ell} B_{k_i}(\mathbf{0}) = B_C(\mathbf{0})$, and hence X is bounded.

To see that X is closed, or equivalently that $\mathbb{R}^n - X$ is open, suppose that $\mathbf{y} \notin X$. We must show that there is an open ball $B_r(\mathbf{y}) \subseteq \mathbb{R}^n - X$. For all $\mathbf{x} \in X$, $\|\mathbf{x}-\mathbf{y}\| > 0$, and hence there exists a $k \in \mathbb{N}$ such that $\|\mathbf{x}-\mathbf{y}\| > 1/k$. Thus, $\mathbf{x} \in U_k = \mathbb{R}^n - \overline{B_{1/k}(\mathbf{y})}$ for some k, so that $\{U_k : k \in \mathbb{N}\}$ is an open cover of X. It follows that there is a finite subcover $\{U_{k_1}, \ldots, U_{k_\ell}\}$. Taking $M = \sup\{k_1, \ldots, k_\ell\}$, we see that, for all $\mathbf{x} \in X$, $\|\mathbf{x} - \mathbf{y}\| > 1/M$. Hence, if $\mathbf{p} \in B_{1/M}(\mathbf{y})$, then $\|\mathbf{y} - \mathbf{p}\| < 1/M$, and so $\mathbf{p} \notin X$. It follows that $B_{1/M}(\mathbf{y}) \subseteq \mathbb{R}^n - X$, so that $\mathbb{R}^n - X$ is open and hence X is closed.

Conversely, suppose that X is compact. We must show that X has the Heine-Borel property. Actually, we shall something a little weaker: suppose that $\{U_k : k \in \mathbb{N}\}$ is an open cover of X. Then there exists a finite subcover $\{U_{k_1}, \ldots, U_{k_\ell}\}$. (The difference between this statement and the general Heine-Borel property is that we assume that the cover is indexed by the natural numbers, or equivalently is countably infinite. Using different ideas, one can prove the general statement. Alternatively, one can show that, if $\{U_i : i \in I\}$ is a collection of open sets in $\mathbb{R}^n$, then there exists a countable subset $J \subseteq I$ such that $\bigcup_{i\in J} U_i = \bigcup_{i\in I} U_i$. Thus one can reduce to the case of a countable open cover.)

So suppose that $X \subseteq \bigcup_{k\in\mathbb{N}} U_k$, where each U_k is open. Suppose that no finite subset of the U_k is a cover; we will find a contradiction. For each k, let $V_k = \bigcup_{i=1}^{k} U_i$. Then V_k is an open set, $V_k \subseteq V_{k+1}$, and, by assumption, for all k, X is not contained in V_k. Thus, for all $k \in \mathbb{N}$, there exists a $\mathbf{v}_k \in X$ with $\mathbf{v}_k \notin V_k$. Since X is compact, the sequence $\{\mathbf{v}_k\}$ has a convergent subsequence $\{\mathbf{v}_{k_\ell}\}$ whose limit $\mathbf{L}$ is in X. Hence $\mathbf{L} \in U_N$ for some N, since the U_k are an open cover of X, and therefore $\mathbf{L} \in V_N$. Since V_N is open, there is a ball $B_\epsilon(\mathbf{L}) \subseteq V_N$. Now, if ℓ is large enough, $\|\mathbf{v}_{k_\ell} - \mathbf{L}\| < \epsilon$, since $\lim_{\ell\to\infty} \mathbf{v}_{k_\ell} = \mathbf{L}$, and so $\mathbf{v}_{k_\ell} \in V_N$ as long as ℓ is large enough. But, if $\ell \geq N$, $k_\ell \geq N$, and by construction $\mathbf{v}_{k_\ell} \notin V_{k_\ell}$ and hence $\mathbf{v}_{k_\ell} \notin V_N$ since $V_N \subseteq V_{k_\ell}$. This is a contradiction. Hence there must exist a finite subcover of the U_k. □

7.6 Connected sets

Now we turn to the analogue of the intermediate value theorem. The idea, as in the definition of compact sets, will be to find some abstract definition of a class of sets which will satisfy the intermediate value theorem, and then give a method for describing when a given set satisfies the definition.

However, the definitions will be a little more involved:

Definition 7.47. Let X be an open subset of $\mathbb{R}^n$. Then X is *connected* if there do not exist open sets U and V of $\mathbb{R}^n$ such that $X = U \cup V$, both U and V are nonempty, but $U \cap V = \emptyset$. An arbitrary subset X of $\mathbb{R}^n$ is *connected* if there do not exist open sets U and V of $\mathbb{R}^n$ such that $X \subseteq U \cup V$, both $X \cap U$ and $X \cap V$ are nonempty, but $X \cap U \cap V = \emptyset$.

Thus, a connected open set X is not the union of two open sets U and V, such that U and V are nonempty and disjoint. Conversely, any union of two open disjoint nonempty sets is *disconnected*, i.e. not connected. If $X = \{\mathbf{p}, \mathbf{q}\}$, where $\mathbf{p}$ and $\mathbf{q}$ are two distinct points of $\mathbb{R}^n$, then X is not connected. In fact, if $r = \|\mathbf{p} - \mathbf{q}\|$, then the open balls $U = B_{r/2}(\mathbf{p})$ and $V == B_{r/2}(\mathbf{q})$ are disjoint, by the triangle inequality, and clearly $X \subseteq U \cup V$.

For another example, suppose that $X \subseteq \mathbb{R}$ and that there exist two points $a, b \in X$ with $a < b$, say. If there exists a c with $a < c < b$ but $c \notin X$, then take $U = (-\infty, c)$ and $V = (c, \infty)$. Clearly, since $c \notin X$, $X \subseteq U \cap V = \mathbb{R} - \{c\}$. Also, $a \in X \cap U$ and $b \in X \cap V$, so both $X \cap U$ and $X \cap V$ are nonempty. But $X \cap U \cap V \subseteq U \cap V = (-\infty, c) \cap (c, \infty) = \emptyset$. Thus X is not connected.

A deeper fact is:

Theorem 7.48. *An interval is a connected subset of* $\mathbb{R}$.

Proof. Let I be an interval (open. closed, half-open, bounded or unbounded). We must show that I is connected. The basic fact we will use about intervals is: if $x_1, x_2 \in I$ with, say, $x_1 < x_2$, then for all t, if $x_1 \leq t \leq x_2$, then $t \in I$.

If I is not connected, we can write $I \subseteq U \cup V$ where U and V are open with $I \cap U \cap V = \emptyset$, and $I \cap U \neq \emptyset$, $I \cap V \neq \emptyset$. Choose $x_1 \in I \cap U$ and $x_2 \in I \cap V$. We may assume that $x_1 < x_2$ by symmetry (note $x_1 \neq x_2$ for otherwise $x_1 = x_2 \in I \cap U \cap V = \emptyset$). Let

$$Z = \{x \in I \cap U : x < x_2\}.$$

Then Z is nonempty since $x_1 \in Z$ and Z is bounded above (by x_2). Let y be the least upper bound of Z. Since $x_1 \leq y \leq x_2$, $y \in I$. We will show that y cannot be in either $I \cap U$ or $I \cap V$, a contradiction to the assumption that $I \subseteq U \cup V$. Thus it follows that I is connected.

Suppose first that $y \in I \cap U$. In particular $y \neq x_2$, and so $y < x_2$ since x_2 is an upper bound for Z. Since U is open, there exists an $r > 0$ such that $(y - r, y + r) \subseteq U$, and we may assume that r is so small that $y + r < x_2$. But then $[y, y + r) \subseteq I \cap U$, so that for example $y + r/2 \in I \cap U$ and $y + r/2 < x_2$.

This contradicts the fact that y was a upper bound for the set of elements of $I \cap U$ less than x_2.

Thus $y \in V$ and so $y \neq x_1$. It follows as above that there exists an $r > 0$ such that $(y - r, y + r) \subseteq B$, with $y - r > x_1$, and that $(y - r, y] \subseteq I \cap B$. But then $y - r$ is another upper bound for Z, contradicting the choice of y as the least upper bound. So $y \notin V$. This contradicts the assumption that $I \subseteq U \cup V$. □

In fact, it is not hard to show that a subset of $\mathbb{R}$ is connected if and only if it is an interval (open, closed, half open, bounded or unbounded), but the proof involves another application of the least upper bound property and will not be given here.

To apply the above to continuous functions, we have:

Theorem 7.49. *If X is a connected subset of $\mathbb{R}^n$ and $F\colon X \to \mathbb{R}^m$ is continuous, then $F(X)$ is connected.*

Proof. Suppose that X is connected, that $F\colon X \to \mathbb{R}^m$ is continuous, and that $F(X)$ is not connected. Then there exist open subsets $A, B \subseteq \mathbb{R}^m$ such that $F(X) \subseteq A \cup B$, $F(X) \cap A \neq \emptyset$, $F(X) \cap B \neq \emptyset$, and $F(X) \cap A \cap B = \emptyset$. Now since F is continuous, $F^{-1}(A) = X \cap U$ and $F^{-1}(B) = X \cap V$, where U and V are open subsets of $\mathbb{R}^n$. Moreover $X \subseteq F^{-1}(A) \cup F^{-1}(B) \subseteq U \cup V$, $X \cap U = F^{-1}(A) \neq \emptyset$, $X \cap V = F^{-1}(B) \neq \emptyset$, and $X \cap U \cap V \subseteq F^{-1}(A) \cap F^{-1}(B) = \emptyset$. (For example, if $\mathbf{x} \in X \cap U \cap V$, then $\mathbf{x} \in X \cap U = F^{-1}(A)$ and so by definition $F(\mathbf{x}) \in F(X) \cap A$. Likewise $F(\mathbf{x}) \in F(X) \cap B$, and so $F(\mathbf{x}) \in F(X) \cap A \cap B = \emptyset$, which is impossible.) It follows that X is not connected, a contradiction. □

Corollary 7.50. *If X is a connected subset of $\mathbb{R}^n$ and $f\colon X \to \mathbb{R}$ is continuous, then the conclusions of the intermediate value theorem hold for X. In other words, for all $\mathbf{x}_1, \mathbf{x}_2 \in X$ such that $f(\mathbf{x}_1) < f(\mathbf{x}_2)$ and for all c with $f(\mathbf{x}_1) < c < f(\mathbf{x}_2)$, there exists an $\mathbf{x} \in X$ such that $f(\mathbf{x}) = c$.*

Proof. By the above theorem, $f(X)$ is connected. By the example after the definition of connectedness, this implies that, since $f(X)$ contains the points $f(\mathbf{x}_1)$, $f(\mathbf{x}_2)$ with $f(\mathbf{x}_1) < f(\mathbf{x}_2)$, it must contain all c such that $f(\mathbf{x}_1) < c < f(\mathbf{x}_2)$, for otherwise it would not be connected. Thus the intermediate value theorem hold for X. □

To be able to use the above results, we need to be able to find lots of connected sets. Recall that we have defined a *path* in $X \subseteq \mathbb{R}^n$ (or a *parametrized curve*) is a continuous function $\mathbf{r}$ from a closed interval $[a, b]$

to $\mathbb{R}^n$ such that $\mathbf{r}([a,b]) \subseteq X$, with starting point $\mathbf{r}(a)$ and endpoint $\mathbf{r}(b)$. For example, if $\mathbf{p} \in X$, then the function $\mathbf{r}\colon [a,b] \to X$ defined by $\mathbf{r}(t) = \mathbf{p}$ for all $t \in [a,b]$ is a path in X, the *constant path*.

Definition 7.51. If $\mathbf{r}\colon [a,b] \to \mathbb{R}^n$ is a path and $\varphi\colon [\alpha,\beta] \to [a,b]$ is a bijection such that φ and φ^{-1} are continuous, then $\mathbf{r} \circ \varphi\colon [\alpha,\beta] \to \mathbb{R}^n$ is another path. We say that $\mathbf{r} \circ \varphi$ is *obtained from* $\mathbf{r}$ *via reparametrization*. It is easy to see that reparametrization is an equivalence relation on the set of paths. Using a reparametrization, we often normalize the domain of $\mathbf{r}$ to be $[0,1]$. Note that a path is *oriented*: we have a starting point and an endpoint and hence a direction. If $\varphi\colon [\alpha,\beta] \to [a,b]$ satisfies $\varphi(\alpha) = a$ and $\varphi(\beta) = b$, then we say that φ *preserves the orientation*, whereas if $\varphi(\alpha) = b$ and $\varphi(\beta) = a$, then we say that φ *reverses the orientation*. We can always find some reparametrization which reverses the orientation, for example by taking $\varphi(t) = -t$ (in which case $[\alpha,\beta] = [-b,-a]$), or $\varphi(t) = a + b - t$ (in which case $[\alpha,\beta] = [a,b]$ again).

If $\mathbf{r}_1\colon [a_1,b_1] \to \mathbb{R}^n$ and $\mathbf{r}_2\colon [a_2,b_2] \to \mathbb{R}^n$ are two paths, and $\mathbf{r}(b_1) = \mathbf{r}_2(a_2)$, in other words the starting point of $\mathbf{r}_2$ is the endpoint of $\mathbf{r_1}$, then we can make a new path as follows: choose some point $c > b_1$ and some continuous bijection $\varphi\colon [b_1,c] \to [a_2,b_2]$ such that $\varphi(a_2) = b_1$. For example, we could take $\varphi(t)$ to be of the form $At + B$ for some constants A and B. Then we can define $\mathbf{r}\colon [a_1,c] \to \mathbb{R}^n$ by:

$$\mathbf{r}(t) = \begin{cases} \mathbf{r}_1(t), & \text{if } a_1 \le t \le b_1; \\ \mathbf{r}_2 \circ \varphi(t), & \text{if } b_1 \le t \le c. \end{cases}$$

The main point is to check that $\mathbf{r}$ is continuous, and it is enough to check at the point b_1. Here $\mathbf{r}$ is really only defined up to reparametrization. We will sometimes denote the path $\mathbf{r}$ by $\mathbf{r}_1 + \mathbf{r}_2$, although this notation could be confused with the path which is a vector sum of two paths both defined on the same interval.

Definition 7.52. A set $X \subseteq \mathbb{R}^n$ is *path connected* if, for every $\mathbf{x}$ and $\mathbf{y} \in X$, there exists a path $\mathbf{r}\colon [a,b] \to X$ with $\mathbf{r}(a) = \mathbf{x}$ and $\mathbf{r}(b) = \mathbf{y}$.

For example, $\mathbb{R}^n$ itself is path connected. A ball (with respect to any norm) is path connected, as is every convex set. The graph of a continuous function $f\colon \mathbb{R}^n \to \mathbb{R}$ is a path connected subset of $\mathbb{R}^{n+1}$. More generally, suppose that X is a path connected subset of $\mathbb{R}^n$ and that $F\colon X \to \mathbb{R}^m$ is continuous. Then the graph of F is a path connected subset of $\mathbb{R}^{n+m}$

(exercise). For example, the top half of the unit sphere in $\mathbb{R}^n$ (in the $\|\cdot\|$ norm) is path connected: it is the graph of the continuous function

$$x_n = \sqrt{1 - x_1^2 - \cdots - x_{n-1}^2},$$

defined on the closed unit ball in $\mathbb{R}^{n-1}$. In fact, the whole sphere is connected. More generally:

Lemma 7.53. *If X_1 and X_2 are two path connected subsets of $\mathbb{R}^n$ and $X_1 \cap X_2 \neq \emptyset$, then $X_1 \cup X_2$ is path connected.*

Proof. We must show that any two points $\mathbf{x}$ and $\mathbf{y} \in X_1 \cup X_2$ are connected by a path in $X_1 \cup X_2$. If $\mathbf{x}$ and $\mathbf{y} \in X_1$, then we can connect them by a path lying in X_1; likewise if $\mathbf{x}$ and $\mathbf{y} \in X_2$. So the only problem is when one of the points, say $\mathbf{x}$, is in X_1 and the other, say $\mathbf{y}$, is in X_2. Choose a point $\mathbf{p} \in X_1 \cap X_2$. Then there is a path $\mathbf{r}_1 \colon [a_1, b_1] \to X_1$ such that $\mathbf{r}_1(a_1) = \mathbf{x}$ and $\mathbf{r}_1(b_1) = \mathbf{p}$, and similarly there is a path $\mathbf{r}_2 \colon [a_2, b_2] \to X_2$ such that $\mathbf{r}_2(a_2) = \mathbf{p}$ and $\mathbf{r}_2(b_2) = \mathbf{y}$. By the procedure described in Definition 7.51, there is a real number $c > b_1$ and a path $\mathbf{r} = \mathbf{r}_1 + \mathbf{r}_2 \colon [a_1, c] \to X_1 \cup X_2$, such that $\mathbf{r}(a_1) = \mathbf{x}$ and $\mathbf{r}(c) = \mathbf{y}$. Hence $X_1 \cup X_2$ is path connected. □

Theorem 7.54. *A path connected set X in $\mathbb{R}^n$ is connected.*

Proof. Let X be path connected, and suppose that X is not connected. Let A and B be open subsets of $\mathbb{R}^n$ such that $X \subseteq A \cup B$ with $X \cap A \cap B = \emptyset$, and $X \cap A \neq \emptyset$ and $X \cap B \neq \emptyset$. Choose $\mathbf{v} \in X \cap A$, $\mathbf{w} \in X \cap B$, and let $\mathbf{r} \colon [a, b] \to X$ be a path with $\mathbf{r}(a) = \mathbf{v}$ and $\mathbf{r}(b) = \mathbf{w}$. It follows that $\mathbf{r}([a, b]) \subseteq X \subseteq A \cup B$ with $\mathbf{r}([a, b]) \cap A \cap B \subseteq X \cap A \cap B = \emptyset$, and $\mathbf{r}([a, b]) \cap A \neq \emptyset$ and $\mathbf{r}([a, b]) \cap B \neq \emptyset$ since $\mathbf{v} \in \mathbf{r}([a, b]) \cap A$ and $\mathbf{w} \in \mathbf{r}([a, b]) \cap B$. But then $\mathbf{r}([a, b])$ is not connected. This contradicts the previous theorems, since we know that $[a, b]$ is connected and that $\mathbf{r}$ is continuous. Thus X must be connected. □

Corollary 7.55. *The only subsets of $\mathbb{R}^n$ which are both open and closed are $\mathbb{R}^n$ and $\emptyset$.*

Proof. Since $\mathbb{R}^n$ is path connected, it is connected. Now suppose that U is a subset of $\mathbb{R}^n$ which is both open and closed. If $U \neq \emptyset$ and $U \neq \mathbb{R}^n$, then both U and $V = \mathbb{R}^n - U$ are open and nonempty. Clearly $U \cap V = \emptyset$ and $U \cup V = \mathbb{R}^n$. But this contradicts the fact that $\mathbb{R}^n$ is connected. □

Note however that there exist connected subsets of $\mathbb{R}^n$ for $n \geq 2$ which are not path connected (although they look somewhat complicated). However, for **open** subsets of $\mathbb{R}^n$, we have the following:

Lemma 7.56. *If X is an open connected subset of $\mathbb{R}^n$, then X is path connected.*

Proof. Trivially, the empty set is both connected and path connected. So we can assume that $X \neq \emptyset$. Choose a point $\mathbf{p} \in X$, and define U to be the subset of X consisting of all $\mathbf{x} \in X$ such that there exists a path $\mathbf{r}\colon [a,b] \to X$ with $\mathbf{r}(a) = \mathbf{p}$ and $\mathbf{r}(b) = \mathbf{x}$. Note that $\mathbf{p} \in U$ since the constant path joins $\mathbf{p}$ to itself. We claim that U is an open subset of X, and that $X - U = V$ is also open. Since $X = U \cup V$ and $U \cap V = \emptyset$, the only possibility, since X is connected, is that $V = \emptyset$ and $U = X$. In other words, for every $\mathbf{x} \in X$, there exists a path $\mathbf{r}\colon [a,b] \to X$ with $\mathbf{r}(a) = \mathbf{p}$ and $\mathbf{r}(b) = \mathbf{x}$. Now if $\mathbf{x}$ and $\mathbf{y}$ are any two points of X, there are paths $\mathbf{r}_1\colon [a,b] \to X$ with $\mathbf{r}_1(a) = \mathbf{p}$ and $\mathbf{r}_1(b) = \mathbf{x}$ and $\mathbf{r}_2\colon [\alpha,\beta] \to X$ with $\mathbf{r}_2(\alpha) = \mathbf{p}$ and $\mathbf{r}_2(\beta) = \mathbf{y}$. The path $(-\mathbf{r}_1) + \mathbf{r}_2$ defined in Definition 7.51 is then a path from $\mathbf{x}$ to $\mathbf{y}$, so that X is path connected.

To see that U and $V = X - U$ are open, suppose first that $\mathbf{x} \in U$. Since X is open, there exists a ball B centered at $\mathbf{x}$ such that $B \subseteq X$. We will show that $B \subseteq U$, showing that U is open. By the definition of U, there exists a path $\mathbf{r}\colon [a,b] \to X$ with $\mathbf{r}(a) = \mathbf{p}$ and $\mathbf{r}(b) = \mathbf{x}$. Now, if $\mathbf{y}$ is any point of B, since B is convex and hence path connected, there is a path (for example a line segment) $\mathbf{r}_1\colon [\alpha,\beta] \to B$ with $\mathbf{r}_1(\alpha) = \mathbf{x}$ and $\mathbf{r}_1(\beta) = \mathbf{y}$. Thus the path $\mathbf{r} + \mathbf{r}_1$ defined in Definition 7.51 is a path from $\mathbf{p}$ to $\mathbf{x}$, so that by definition $\mathbf{x} \in U$. Hence $B \subseteq U$ and U is open.

A very similar argument shows that $V = X - U$ is open: if $\mathbf{y} \in V$ and B is any ball centered at $\mathbf{y}$ and contained in X, then $B \subseteq V$ as well, because if there were a point $\mathbf{z}$ of B in U, then we could connect $\mathbf{p}$ to $\mathbf{z}$ by a path in X, and then we could connect $\mathbf{z}$ to $\mathbf{y}$ by a path in the convex set B, so that there would be a path in X connecting $\mathbf{p}$ to $\mathbf{y}$, implying that $\mathbf{y} \in U$, a contradiction. Thus V is also open, and so, since $X = U \cup V$ with $U \cap V = \emptyset$ and X is connected, $V = \emptyset$ and $X = U$ is path connected. □

Chapter 8

Derivatives

8.1 A review of calculus in one variable

There are two different ways of looking at the derivative $f'(a)$ of a function $f(x)$ in one variable. The first and perhaps more intuitive way is to think of the limit $\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = f'(a)$ as giving the slope of the graph $y = f(x)$ at the point $(a, f(a))$, or (equivalently) the "instantaneous rate of change" of the function $f(x)$ at $x = a$. The second way to understand the derivative is via the tangent line approximation. The idea here is to find the line $y = mx + b$ which best approximates $f(x)$ at $x = a$, and the main problem is to give a quantitative meaning to the phrase "best approximates." Clearly we should assume that the line passes through the point $(a, f(a))$, so that it is reasonable to write the equation for the line as $y = f(a) + m(x - a)$. To see what property m should satisfy so as to give the value $f'(a)$, we try to work backwards from the definition of the derivative and rewrite this definition as follows: we can replace $\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = f'(a)$ by any one of the equivalent statements

$$\lim_{x \to a} \left(\frac{f(x) - f(a)}{x - a} - f'(a) \right) = 0;$$

$$\lim_{x \to a} \left(\frac{f(x) - f(a) - f'(a)(x - a)}{x - a} \right) = 0;$$

$$\lim_{x \to a} \left| \frac{f(x) - f(a) - f'(a)(x - a)}{x - a} \right| = 0;$$

$$\lim_{x \to a} \frac{|f(x) - f(a) - f'(a)(x - a)|}{|x - a|} = 0.$$

This gives a quantitative measure of how well the tangent line approximates the function: if we let $e(x)$ be the error $|f(x) - f(a) - f'(a)(x-a)|$, and we set

$$h(x) = \frac{|f(x) - f(a) - f'(a)(x-a)|}{|x-a|},$$

then by definition

$$e(x) = |f(x) - f(a) - f'(a)(x-a)| = h(x)|x-a|$$

is of the form $h(x)|x-a|$ where $\lim_{x\to a} h(x) = 0$. This says that the error tends to 0 **faster** than $|x-a|$—of course, we will often want to know how much faster. Let us make the following definition:

Definition 8.1. A function $f(x)$ is $o(g(x))$ at a if there exists a function $h(x)$ with $\lim_{x\to a} h(x) = 0$ such that, for all x close to a (i.e. in some open interval containing a), $|f(x)| = h(x)|g(x)|$. For example, to say that $f(x)$ is $o(1)$ says that $\lim_{x\to a} f(x) = 0$. The function $|x-a|^r$ is $o(|x-a|)$ if and only if $r > 1$. The above discussion says that $|f(x) - f(a) - f'(a)(x-a)|$ is $o(|x-a|)$. In general, to say that $f(x)$ is $o(g(x))$ means roughly that f is of a smaller order of magnitude (decreases faster) than g.

Likewise, we define f to be $O(g(x))$ if there exists a constant $C > 0$ such that, for all x close to a, $|f(x)| \leq C|g(x)|$. For example, $|x-a|^r$ is $O(|x-a|)$ if and only if $r \geq 1$. If $f(x)$ is $o(g(x))$, then it is $O(g(x))$. If $f(x)$ is $O(|x-a|^r)$, then it is $o(|x-a|^s)$ for every $s < r$. To say that f to be $O(g(x))$ is to say roughly that f is at worst the same order of magnitude as g (decreases at least as fast).

We now go back and restate the definition of the derivative in this language:

Proposition 8.2. *Let $f(x)\colon I \to \mathbb{R}$ be a function, where I is an open interval, and suppose that $a \in I$. There exists a real number m such that $f(x) - f(a) - m(x-a)$ is $o(|x-a|)$ $\iff$ $\lim_{x\to a} \dfrac{f(x)-f(a)}{x-a} = f'(a)$ exists, and in this case $m = f'(a)$.*

Proof. As we have seen above, the statement that $f(x) - f(a) - m(x-a)$ is $o(|x-a|)$ is equivalent to the statement that

$$\lim_{x\to a} \left| \frac{f(x)-f(a)}{x-a} - m \right| = 0,$$

which in turn is equivalent to the statement that the derivative exists and equals m. □

As usual, we define a function $f\colon I \to \mathbb{R}$ to be *differentiable at* $a \in I$ if the above limit exists and simply to be *differentiable* if it is differentiable at all points of I. In this case, we get a new function $f'(x)$, the *derivative* of $f(x)$.

Example 8.3. 1) The constant function is everywhere differentiable and its derivative is identically 0.

2) The linear function $mx+b$ is everywhere differentiable and its derivative is constant, equal to m.

Note the following property of a differentiable function:

Lemma 8.4. *If f is differentiable at a, it is continuous at a. Moreover, if we define a function $F(x)$ by*

$$F(x) = \begin{cases} \dfrac{f(x)-f(a)}{x-a}, & \text{if } x \neq a; \\ f'(a), & \text{if } x = a, \end{cases}$$

then $F(x)$ is continuous at a.

Proof. To see the first statement, note that

$$\lim_{x\to a}(f(x)-f(a)) = \lim_{x\to a} f'(a)(x-a) = 0.$$

Thus $\lim_{x\to a} f(x) = f(a)$, and so by definition $f(x)$ is continuous at a. The second statement also follows easily from the definitions and basic properties of continuity. □

We then have the usual rules for derivatives:

Proposition 8.5. *Let $f(x)$ and $g(x)$ be functions defined on an interval containing a and suppose that f and g are differentiable at a. Then*

(i) (Sum rule) $f+g$ is differentiable at a, and its derivative at a is

$$(f+g)'(a) = f'(a) + g'(a);$$

(ii) (Product rule) fg is differentiable at a, and its derivative at a is

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a);$$

(iii) (Quotient rule) If $g(a) \neq 0$, thenf/g is differentiable at a, and its derivative at a is

$$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g(a)^2};$$

(iv) (Chain rule) If f is differentiable at a and g is differentiable at $b = f(a)$, then $g \circ f$ is differentiable at a and

$$(g \circ f)'(a) = g'(f(a)) \cdot f'(a).$$

Proof. (i) is easy and is left as an exercise. For (ii), we calculate the difference quotient for fg. As usual the basic trick is to add and subtract something. We have

$$\begin{aligned}\frac{(fg)(x) - (fg)(a)}{x-a} &= \frac{f(x)g(x) - f(a)g(x) + f(a)g(x) - f(a)g(a)}{x-a} \\ &= \frac{f(x)-f(a)}{x-a}g(x) + f(a)\frac{g(x)-g(a)}{x-a}.\end{aligned}$$

Taking the limit as $x \to a$ and noting that $g(x)$ is continuous at a since it is differentiable at a, we see that the limit of the above expression is indeed $f'(a)g(a) + f(a)g'(a)$.

For (iii), let us first calculate $(1/g)'$; the general formula will then follow by applying the product rule to the function $f/g = f \cdot (1/g)$. By definition we need to calculate the limit as $x \to a$ of

$$\begin{aligned}\frac{\frac{1}{g(x)} - \frac{1}{g(a)}}{x-a} &= \frac{\frac{g(a)-g(x)}{g(x)g(a)}}{x-a} \\ &= \frac{1}{g(x)g(a)}\frac{g(a)-g(x)}{x-a}.\end{aligned}$$

Since g is continuous and $g(a) \neq 0$, the first term goes to $1/g(a)^2$, and by definition the second goes to $-g'(a)$. Thus we have the formula

$$\left(\frac{1}{g}\right)'(a) = -\frac{g'(a)}{g(a)^2}.$$

Applying the product formula to f/g then gives (iii).

Finally we must prove the Chain Rule. If we go ahead and write down the difference quotient, we obtain

$$\frac{g(f(x)) - g(f(a))}{x-a}.$$

We try to use the standard trick of multiplying by 1 in the following form

$$\frac{g(f(x)) - g(f(a))}{x - a} = \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} \cdot \frac{f(x) - f(a)}{x - a}. \qquad (*)$$

This is allowed unless $f(x) = f(a) = b$. Thus first assume that for all x sufficiently close to a and $\neq a$, $f(x) \neq b$. Then the right hand side of the expression above makes sense. If instead it is possible that $f(x) = b$ for a sequence of x converging to a, we proceed as follows: define $G(y)$ by the formula

$$G(y) = \begin{cases} \dfrac{g(y) - g(b)}{y - b}, & \text{if } y \neq b; \\ g'(b), & \text{if } y = b. \end{cases} \qquad (**)$$

As we have seen in Lemma 8.4, G is continuous at b. We claim that in any case

$$\frac{g(f(x)) - g(f(a))}{x - a} = G(f(x)) \cdot \frac{f(x) - f(a)}{x - a}.$$

This follows from our trick above of multiplying and dividing by $f(x) - f(a)$, if $f(x) \neq f(a)$. But if $f(x) = f(a)$, then both $g(f(x)) - g(f(a))$ and $f(x) - f(a)$ are 0, so $(**)$ is still true.

Now take the limit as $x \to a$ of both sides of $(*)$. Since $G(y)$ is continuous at b, and $\lim_{x \to a} f(x) = f(a)$, it follows that $\lim_{x \to a} G(f(x)) = G(f(a)) = G(b) = g'(b) = g'(f(a))$. Moreover the limit of the term $\dfrac{f(x) - f(a)}{x - a}$ as $x \to a$ is $f'(a)$. So the limit of the product is $g'(f(a)) \cdot f'(a)$ as desired. □

We also record the familiar result about extreme values:

Definition 8.6. Let $f(x)$ be a function defined on an open interval I. A point a is a *local maximum* of f if there is an open interval $J \subset I$ with $a \in J$ such that, for all $x \in J$, $f(x) \leq f(a)$. Local minima are similarly defined. Note that a maximum value of f on I (if it exists) is necessarily a local maximum, but the converse need not hold.

Proposition 8.7. *Let $f(x)$ be differentiable at a, and suppose that a is a local maximum or minimum for f. Then $f'(a) = 0$.*

Proof. Suppose that a is a local maximum. Then for all x close to a, $f(x) - f(a) \leq 0$. Now for $x < a$, $x - a < 0$ and so

$$\frac{f(x) - f(a)}{x - a} \geq 0.$$

Thus

$$\lim_{x \to a^-} \frac{f(x) - f(a)}{x - a} \geq 0.$$

Similar reasoning shows that

$$\lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} \leq 0.$$

Since the limit exists, both the upper and lower limits must be equal, and so the limit must be zero. The case of a local minimum is similar. □

In general it is hard to use the definition of a derivative directly to say something about a function. The Mean Value Theorem allows us to go from "local" information about a function f, namely some statement about its derivative at every point a (which is defined by the values of f near a), to get "global" information about the function f (e.g. f is constant or increasing). Before stating and proving the Mean Value Theorem, we begin with the special case known as Rolle's theorem:

Theorem 8.8 (Rolle's theorem)**.** *Let g be continuous on $[a, b]$ and differentiable on (a, b), and suppose that $g(a) = g(b)$. Then there exists a c with $a < c < b$ such that $g'(c) = 0$.*

Proof. If g is constant there is nothing to prove, since g' is identically zero. Otherwise g has either a maximum or a minimum at some point $c \neq a, b$, by the extreme value theorem, as g is continuous on $[a, b]$. Then by Proposition 8.7 $g'(c) = 0$. □

Theorem 8.9 (Mean Value Theorem)**.** *Let $f(x)$ be continuous on the interval $[a, b]$ and differentiable on the interval (a, b). Then there exists a c with $a < c < b$ such that*

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

Proof. The idea is to replace the function $f(x)$ by a function to which the hypotheses of Rolle's theorem apply. In other words we need to replace f by some new function F where F is again continuous on the interval $[a, b]$ and differentiable on the interval (a, b), and further F satisfies $F(a) = F(b)$. The trick is to subtract off the equation for the secant line to the graph of f passing through $(a, f(a))$ and $(b, f(b))$. This is a new function which has

the value $f(a)$ at a and $f(b)$ at b, and so the difference will be zero at a and b and will thus satisfy the assumptions of Rolle's theorem. Explicitly we let

$$F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

Since the equation for the secant line is a linear function, it is everywhere differentiable, and so $F(x)$ is continuous on $[a, b]$ and differentiable on (a, b) and $F(a) = F(b) = 0$. Applying Rolle's theorem to F, there is a c with $a < c < b$ such that $F'(c) = 0$. But

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

Thus to say that $F'(c) = 0$ is exactly to say that

$$f'(c) = \frac{f(b) - f(a)}{b - a},$$

as desired. □

Corollary 8.10. *Let $f(x)$ be continuous on the interval $[a, b]$ and differentiable on the interval (a, b), and suppose that $f'(x) = 0$ for all $x \in (a, b)$. Then f is constant.*

Proof. Let us show that $f(t) = f(a)$ for every $t \in [a, b]$. This is of course obvious if $t = a$. Otherwise we consider the restriction of f to $[a, t]$, where it still satisfies the hypotheses of the Mean Value Theorem. Thus

$$\frac{f(t) - f(a)}{t - a} = f'(d) = 0,$$

where d is some point in $(a, t) \subseteq (a, b)$. So $f(t) - f(a) = 0$ and hence $f(t) = f(a)$. □

Corollary 8.11. *Let f and g be continuous on the interval $[a, b]$ and differentiable on the interval (a, b), and suppose that $f' = g'$. Then there is a constant C such that $f = g + C$.*

Proof. Apply the above corollary to the function $f - g$. □

Corollary 8.12. *Let $f(x)$ be differentiable on the open interval I and suppose that $f'(x) > 0$ for all $x \in I$. Then f is strictly increasing in I. Likewise if $f'(x) < 0$ for all $x \in I$, then f is strictly decreasing in I.*

Proof. We shall just consider the case where $f'(x) > 0$ for all $x \in I$; the other case is similar. Given $a < b \in I$, apply the Mean Value Theorem to the interval $[a, b] \subset I$. It follows that there is some $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c)(b - a) > 0.$$

Thus $f(b) > f(a)$ whenever $b > a$, $a, b \in I$. So f is strictly increasing. □

It is easy to see that if f is increasing in I and differentiable, then $f'(x) \geq 0$ for all $x \in I$. For if $f(x_1) \leq f(x_2)$ whenever $x_1 < x_2$, then consider

$$\frac{f(x) - f(a)}{x - a}.$$

If $x < a$, then the numerator is ≤ 0 and the denominator is < 0, so the quotient is ≥ 0. Likewise if $x > a$, then the numerator is ≥ 0 and the denominator is > 0, so that once again the quotient is ≥ 0. Taking the limit as $x \to a$, we see that $f'(a) \geq 0$. However it is possible to have $f'(x) = 0$ at some point x for a strictly increasing function f. For example the function $f(x) = x^3$ is strictly increasing on $\mathbb{R}$ but its derivative at 0 is 0.

We remark that, however obvious the Mean Value Theorem or any of the corollaries might appear, they all depend in some deep sense on the completeness of the real numbers. For example we can define in an identical way the concept that a function defined on the rational numbers $\mathbb{Q}$ or on an interval (a, b) of rational numbers (i.e. on a set of the form $\{x \in \mathbb{Q} : a < x < b\}$ is differentiable or continuous. On the other hand the Mean Value Theorem and all of the corollaries above do not in general remain true for such functions. For example the function defined on $[0, 4] \cap \mathbb{Q}$ defined by

$$f(x) = \begin{cases} 1, & \text{if } 0 \leq x < \sqrt{2}; \\ 2, & \text{if } \sqrt{2} < x \leq 4, \end{cases}$$

has derivative zero but is not constant. Likewise the function defined by

$$f(x) = \begin{cases} x, & \text{if } 0 \leq x < \sqrt{2}; \\ x - 4, & \text{if } \sqrt{2} < x \leq 4, \end{cases}$$

has a positive derivative but is not increasing.

8.2 Derivatives in several variables

We now try to define derivatives of functions $F \colon \mathbb{R}^n \to \mathbb{R}^m$, or more generally functions F defined on an open subset X of $\mathbb{R}^n$. Such functions can be

separated into two extremes: first there are functions $f\colon \mathbb{R}^n \to \mathbb{R}$ which are real valued functions of several variables, and then there are functions $\mathbf{r}\colon \mathbb{R} \to \mathbb{R}^m$ which are vector valued functions of a single real variable. Thus $\mathbf{r}(t) = (x_1(t), \ldots, x_m(t))$. The most general function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is just the vector-valued function $F = (f_1, \ldots, f_m)$ where each f_i is a real valued function of n variables. It turns out that there is no problem in defining the derivative of a vector valued function $\mathbf{r}$ of a single variable (if it exists): we let

$$\mathbf{r}'(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left(\mathbf{r}(t + \Delta t) - \mathbf{r}(t)\right).$$

Geometrically this expression measures the difference vector (up to scalar multiples) $\mathbf{r}(t + \Delta t) - \mathbf{r}(t)$, and can be thought of as the instantaneous rate of change of the position vector $\mathbf{r}(t)$. In the limit we can think of $\mathbf{r}'(t)$ as giving the tangent vector to the curve. If it exists, it is equal to

$$\mathbf{r}'(t) = \frac{d}{dt}\mathbf{r}(t) = (x_1'(t), \ldots, x_m'(t)).$$

A path $\mathbf{r}\colon [a, b] \to \mathbb{R}^n$ such that $\mathbf{r}'(t) = (x_1'(t), \ldots, x_m'(t))$ exists for all $t \in (a, b)$ will be called a *differentiable path*. Usually we will also make the requirement that the functions $x_i'(t)$ extend to continuous functions on $[a, b]$. It is then natural to only allow reparametrizations of $[a, b]$ of the form $\varphi\colon [\alpha, \beta] \to [a, b]$ such that $\varphi'(t)$ is a continuous function on (α, β) which extends to a continuous function on $[\alpha, \beta]$. We will refer to $\mathbf{r}'(t)$ as the *tangent vector* to the (parametrized) curve $\mathbf{r}(t)$; sometimes we call $\mathbf{r}'(t)$ the *velocity vector* and write it as $\mathbf{v}(t)$. Under our assumptions above, it is a continuous function of t. As we shall see, the magnitude $\|\mathbf{v}(t)\|$ of the velocity vector is also important and is called the *speed* of $\mathbf{r}(t)$.

There are standard formulas for the derivative $\mathbf{r}'$ under the usual operations, which follow immediately from the rules for derivatives of one variable:

Proposition 8.13. *1. If $\mathbf{r}_1\colon (a, b) \to \mathbb{R}^n$ and $\mathbf{r}_2\colon (a, b) \to \mathbb{R}^n$ are two differentiable paths, then so is $\mathbf{r}_1 + \mathbf{r}_2$, and $(\mathbf{r}_1 + \mathbf{r}_2)' = \mathbf{r}_1' + \mathbf{r}_2'$. (Here by the sum we mean the usual pointwise vector sum of the functions, not the sum of paths as defined earlier.)*

2. If $\mathbf{r}$ is a differentiable path $(a, b) \to \mathbb{R}^n$ and $g\colon (a, b) \to \mathbb{R}$ is a differentiable function, then $g\mathbf{r}$ is a differentiable path and $(g\mathbf{r}) = g'\mathbf{r} + g\mathbf{r}'$.

3. If $\mathbf{r}_1$ and $\mathbf{r}_2$ are two differentiable paths $(a, b) \to \mathbb{R}^n$, then $\langle \mathbf{r}_1, \mathbf{r}_2 \rangle$ is a differentiable (real-valued) function and

$$\frac{d}{dt}\langle \mathbf{r}_1, \mathbf{r}_2 \rangle = \langle \mathbf{r}_1', \mathbf{r}_2 \rangle + \langle \mathbf{r}_1, \mathbf{r}_2' \rangle. \quad \square$$

Corollary 8.14. *If* $\mathbf{r}$ *is a differentiable path* $(a,b) \to \mathbb{R}^n$*, then the real-valued function* $\|\mathbf{r}\|^2$ *is differentiable and that*

$$\frac{d}{dt}\|\mathbf{r}\|^2 = 2\langle \mathbf{r}, \mathbf{r}' \rangle.$$

Hence, if $\mathbf{r}$ *is a differentiable path* $(a,b) \to \mathbb{R}^n$ *such that* $\|\mathbf{r}\|^2$ *is constant, i.e. such that the image of* $\mathbf{r}$ *is contained in some sphere centered at the origin, then the tangent vector* $\mathbf{r}'(t)$ *is always orthogonal to* $\mathbf{r}(t)$*.* □

The proofs are left as exercises.

In this context, the chain rule reads as follows:

Lemma 8.15. *If* $\mathbf{r}$ *is a differentiable path* $(a,b) \to \mathbb{R}^n$ *and* $\varphi\colon (\alpha,\beta) \to (a,b)$ *is differentiable, then* $\mathbf{r} \circ \varphi$ *is differentiable and*

$$(\mathbf{r} \circ \varphi)' = \varphi' \cdot (\mathbf{r}' \circ \varphi). \quad \square$$

We can write this in differential notation, if $t = \varphi(u)$, as

$$\frac{d\mathbf{r}}{du} = \frac{dt}{du} \cdot \frac{d\mathbf{r}}{dt},$$

where we have switched the usual order to keep the scalar multiplication on the left. In particular, we will usually apply this formula when φ is a reparametrization with φ' never 0. It then follows by the intermediate value theorem that either φ' is always positive and φ is strictly increasing or φ' is always negative and φ is strictly decreasing. In the first case, φ is orientation preserving and the tangent vector $\dfrac{d\mathbf{r}}{du}$ is a positive scalar multiple of $\dfrac{d\mathbf{r}}{dt}$; in the second case φ is orientation reversing and the tangent vector $\dfrac{d\mathbf{r}}{du}$ is a negative scalar multiple of $\dfrac{d\mathbf{r}}{dt}$. In both cases the tangent line, i.e. the line spanned by the tangent vector, is unchanged, but the speed, i.e. $\left\|\dfrac{d\mathbf{r}}{dt}\right\|$, is changed by the factor $\left|\dfrac{dt}{du}\right|$.

We turn now to the case of a single real valued function of several variables. If we look at a real valued function $f(x_1,\ldots,x_n)$, one problem with defining a derivative is that there are too many possibilities! We could try to analyze the change of f in each variable separately, by considering the *partial derivative* $\dfrac{\partial f}{\partial x_i}$ with respect to x_i. By definition, if it exists,

$$\frac{\partial}{\partial x_i} f(a_1,\ldots,a_n) = \lim_{\Delta t \to 0} \frac{f(a_1,\ldots,a_i+\Delta t,\ldots,a_n) - f(a_1,\ldots,a_n)}{\Delta t}.$$

In other words, it is the usual derivative $\left.\frac{d}{dt} f(a_1, \ldots, t, \ldots, a_n)\right|_{t=a_i}$ of the function $f(a_1, \ldots, t, \ldots, a_n)$ obtained by holding all of the variables fixed except for x_i, thus obtaining a function of a single variable. Geometrically we may think of the number $\partial f/\partial x_i$ as the slope of the part of the graph of f (which lives in $\mathbb{R}^{n+1}$) which lies above a line parallel to the x_i-axis, at the point $x_i = a_i$. Defined this way, $\partial f/\partial x_i(\mathbf{a})$ is a number (if it exists). In the usual way, we can also define a function $\partial f/\partial x_i$ of $\mathbf{x}$ by taking the partial derivative with respect to x_i at every point $\mathbf{x}$ where the partial derivative is defined. To make this computation, we just take d/dx_i of f, treating f as a function of x_i alone and all the other variables as constants. For example:

$$\frac{\partial}{\partial x_1}(x_1^2 e^{x_1 x_2}) = 2x_1 e^{x_1 x_2} + x_1^2 x_2 e^{x_1 x_2};$$
$$\frac{\partial}{\partial x_2}(x_1^2 e^{x_1 x_2}) = x_1^3 e^{x_1 x_2}.$$

There are several ways we could generalize the definition of partial derivative, all of which might be interesting. For example, why restrict to lines parallel to one of the coordinate axes? In general, given a point $\mathbf{a} \in \mathbb{R}^n$, every line through $\mathbf{a}$ may be described parametrically by

$$\{\mathbf{a} + t\mathbf{v} : t \in \mathbb{R}\}.$$

Here $\mathbf{v}$ is a nonzero vector in $\mathbb{R}^n$, unique up to a nonzero scalar, we which can think of as the direction of the line. To make $\mathbf{v}$ unique, we can assume that $\|\mathbf{v}\| = 1$, in which case $\mathbf{v}$ is called a *unit direction*. It is then unique up to sign, and the two choices of sign correspond to the choices of "direction" (or more properly "orientation") on the line. For an arbitrary vector $\mathbf{v}$, not necessarily a unit vector (or even nonzero) we define the *directional derivative of* f at $\mathbf{a}$ in the direction $\mathbf{v}$ (if it exists) by the formula

$$\left.\frac{d}{dt}\right|_{t=0} f(\mathbf{a} + t\mathbf{v}).$$

If $\mathbf{v} = \mathbf{e}_i$ then the directional derivative is just the partial derivative. More generally still, we could take any differentiable curve $\mathbf{r}(t)$ from $\mathbb{R}$ to $\mathbb{R}^n$ with $\mathbf{r}(0) = \mathbf{a}$. Consider $f \circ \mathbf{r}(t)$, which is a function of one variable, and take its derivative at 0. In fact, it will turn out that this only depends on $\mathbf{r}'(0)$, and can be expressed in terms of the partial derivatives of f, at least for reasonable functions f. In particular there is a formula for the directional

derivative of f in the direction $\mathbf{v}$ in terms of $\mathbf{v}$ and the partial derivatives of f.

However, instead of proceeding in this way, we shall use a different approach, one which does not depend on coordinates. The main reason to avoid coordinates if possible is that there is no natural choice of coordinates on $\mathbb{R}^n$, and it is always best to have a definition which does not depend on a choice. Also, our definition will have the advantage of fitting together all of the partial derivatives at once! To make this definition, note that the limit of the difference quotient

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

cannot make sense when we replace x and a by vectors. So we must find another way to generalize the defintion of derivative. The main idea is to use the tangent line approximation instead:

Definition 8.16. Let $F\colon X \to \mathbb{R}^m$ be a function, where X is an open subset of $\mathbb{R}^n$. We say that F is *differentiable* at $\mathbf{a} \in X$, if there exists a linear map $A\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{\mathbf{x} \to \mathbf{a}} \frac{\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x} - \mathbf{a})\|}{\|\mathbf{x} - \mathbf{a}\|} = 0.$$

Making the change of variable $\mathbf{h} = \mathbf{x} - \mathbf{a}$, we can rewrite this as

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\|F(\mathbf{a} + \mathbf{h}) - F(\mathbf{a}) - A(\mathbf{h})\|}{\|\mathbf{h}\|} = 0.$$

Another way to say this is to say that the difference $F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x} - \mathbf{a})$ satisfies:

$$\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x} - \mathbf{a})\| = H(\mathbf{x})\|\mathbf{x} - \mathbf{a}\|,$$

where $\lim_{\mathbf{x} \to \mathbf{a}} H(\mathbf{x}) = 0$. By analogy with the one variable case, we also say that the difference $F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x} - \mathbf{a})$ is $o(\|\mathbf{x} - \mathbf{a}\|)$. Note that $H(\mathbf{x})$ is necessarily equal to

$$\frac{\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x} - \mathbf{a})\|}{\|\mathbf{x} - \mathbf{a}\|}$$

and is not defined at $\mathbf{a}$, which is irrelevant as far as the limit is concerned. However there is a unique extension of H to a continuous function defined at $\mathbf{a}$, by setting $H(\mathbf{a}) = 0$.

Thus we can think of the (affine) linear function $A(\mathbf{x}-\mathbf{a})+F(\mathbf{a})$ as the "best" linear approximation to F. As we shall see in a minute, the linear map A is unique. We will call the linear function A the *derivative* of F at $\mathbf{a}$ (if it exists). We use the notation $DF_{\mathbf{a}}$ for the derivative of F at $\mathbf{a}$.

One additional remark is that, while we have given the definition of the derivative in terms of the standard norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, we are free to use any equivalent norms in checking that a function is differentiable (and hence any norms).

Proposition 8.17. *1. If A_1 and A_2 are two linear maps $\mathbb{R}^n \to \mathbb{R}^m$ such that $F(\mathbf{x})-F(\mathbf{a})-A_i(\mathbf{x}-\mathbf{a})$ is $o(\|\mathbf{x}-\mathbf{a}\|)$ for $i=1,2$, then $A_1=A_2$. Thus the derivative, if it exists, is unique.*

2. If F is differentiable at $\mathbf{a}$, it is continuous at $\mathbf{a}$.

3. A constant function is differentiable at $\mathbf{a}$ and its derivative is zero.

4. A linear function A is differentiable at $\mathbf{a}$ and its derivative $DA_{\mathbf{a}}=A$.

5. If F, G are two functions $X \to \mathbb{R}^m$ and F, G are differentiable at $\mathbf{a}$, then so is $F+G$, and $D(F+G)_{\mathbf{a}} = DF_{\mathbf{a}} + DG_{\mathbf{a}}$.

6. The function $F=(f_1,\dots,f_m)$ is differentiable at $\mathbf{a}$ if and only each function $f_i\colon X \to \mathbb{R}$ is differentiable at $\mathbf{a}$. Moreover, the derivative $DF_{\mathbf{a}} = ((Df_1)_{\mathbf{a}},\dots,(Df_m)_{\mathbf{a}})$ in the obvious sense.

Proof. 1. Suppose that A_1 and A_2 are two possible derivatives for F at $\mathbf{a}$. Then

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|A_1(\mathbf{h})-A_2(\mathbf{h})\|}{\|\mathbf{h}\|} =$$

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|A_1(\mathbf{h})-(F(\mathbf{a}+\mathbf{h})-F(\mathbf{a}))+(F(\mathbf{a}+\mathbf{h})-F(\mathbf{a}))-(A_2(\mathbf{h})\|}{\|\mathbf{h}\|}$$

$$\le \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|F(\mathbf{a}+\mathbf{h})-F(\mathbf{a})-A_1(\mathbf{h})}{\|\mathbf{h}\|} + \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|F(\mathbf{a}+\mathbf{h})-F(\mathbf{a})-A_2(\mathbf{h})}{\|\mathbf{h}\|} = 0.$$

On the other hand, fixing a nonzero vector $\mathbf{v}\in\mathbb{R}^n$ and setting $\mathbf{h}=t\mathbf{v}$, we have

$$\begin{aligned}\lim_{t\to 0} \frac{\|A_1(t\mathbf{v})-A_2(t\mathbf{v})\|}{\|t\mathbf{v}\|} &= \lim_{t\to 0} \frac{\|tA_1(\mathbf{v})-tA_2(\mathbf{v})\|}{|t|\|\mathbf{v}\|}\\ &= \lim_{t\to 0} \frac{|t|\|A_1(\mathbf{v})-A_2(\mathbf{v})\|}{|t|\|\mathbf{v}\|}\\ &= \lim_{t\to 0} \frac{\|A_1(\mathbf{v})-A_2(\mathbf{v})\|}{\|\mathbf{v}\|},\end{aligned}$$

which is independent of t. Since the limit is zero, $\|A_1(\mathbf{v}) - A_2(\mathbf{v})\| = 0$ for all $\mathbf{v} \neq \mathbf{0}$ and so for all $\mathbf{v}$ (since a linear map is always zero on $\mathbf{0}$). Thus $A_1 = A_2$.

2. Write $\|F(\mathbf{x}) - F(\mathbf{a})\| \leq \|A(\mathbf{x}-\mathbf{a})\| + H(\mathbf{x})\|\mathbf{x}-\mathbf{a}\|$ (what variation on the triangle inequality are we using?) We must show that $\lim_{\mathbf{x}\to\mathbf{a}} F(\mathbf{x}) = F(\mathbf{a})$, or equivalently $\lim_{\mathbf{x}\to\mathbf{a}} \|F(\mathbf{x}) - F(\mathbf{a})\| = 0$. But since a linear map is continuous,

$$\lim_{\mathbf{x}\to\mathbf{a}} \|A(\mathbf{x}-\mathbf{a})\| = \lim_{\mathbf{x}\to\mathbf{a}} \|A(\mathbf{x}) - A(\mathbf{a})\| = 0,$$

and

$$\lim_{\mathbf{x}\to\mathbf{a}} H(\mathbf{x})\|\mathbf{x}-\mathbf{a}\| = 0 \cdot 0 = 0.$$

Thus $\lim_{\mathbf{x}\to\mathbf{a}} \|F(\mathbf{x}) - F(\mathbf{a})\| = 0$.

3. If $F(\mathbf{x}) = \mathbf{c}$ is a constant, take A to be the zero matrix $\mathbf{0}$. Then

$$\|F(\mathbf{x}) - F(\mathbf{a}) - \mathbf{0}(\mathbf{x}-\mathbf{a})\| = \|\mathbf{c} - \mathbf{c} - \mathbf{0}\| = 0 = 0\|\mathbf{x}-\mathbf{a}\|.$$

Thus the definition is satisfied with $H(\mathbf{x}) = 0$.

4. If $F(\mathbf{x}) = A(\mathbf{x})$ for a linear map A, then

$$\begin{aligned}\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\| &= \|A(\mathbf{x}) - A(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\| \\ &= \|A(\mathbf{x}-\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\| = 0 = 0\|\mathbf{x}-\mathbf{a}\|.\end{aligned}$$

Thus the definition is satisfied with $H(\mathbf{x}) = 0$.

5. Left as an exercise.

6. Note that we can use any norm we choose, and it will be convenient (as in general when we try to break a problem up into components) to use the $\|\cdot\|_\infty$ norm. In general, given a linear map $A\colon \mathbb{R}^n \to \mathbb{R}^m$, we can write $A = (\ell_1, \ldots, \ell_m)$, where each ℓ_i is a linear map from $\mathbb{R}^n$ to $\mathbb{R}$; here the $\ell_i(\mathbf{x})$ are the components of the vector $A(\mathbf{x})$. Now in this notation

$$\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\|_\infty = \max_i |f_i(\mathbf{x}) - f_i(\mathbf{a}) - \ell_i(\mathbf{x}-\mathbf{a})|.$$

Thus if each f_i is differentiable at $\mathbf{a}$ with derivative ℓ_i, then for all i

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{|f_i(\mathbf{x}) - f_i(\mathbf{a}) - \ell_i(\mathbf{x}-\mathbf{a})|}{\|\mathbf{x}-\mathbf{a}\|} = 0.$$

If we take $A = (\ell_1, \ldots, \ell_m)$, then it follows that

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\|_\infty}{\|\mathbf{x}-\mathbf{a}\|} = \lim_{\mathbf{x}\to\mathbf{a}} \max_i \frac{|f_i(\mathbf{x}) - f_i(\mathbf{a}) - \ell_i(\mathbf{x}-\mathbf{a})|}{\|\mathbf{x}-\mathbf{a}\|} = 0.$$

Thus F is differentiable at $\mathbf{a}$ and $DF_\mathbf{a} = ((Df_1)_\mathbf{a}, \ldots, (Df_m)_\mathbf{a})$. Conversely, if F is differentiable at $\mathbf{a}$, then for $A = DF_\mathbf{a}$ and $A = (\ell_1, \ldots, \ell_m)$, we have for all i,

$$\frac{|f_i(\mathbf{x}) - f_i(\mathbf{a}) - \ell_i(\mathbf{x}-\mathbf{a})|}{\|\mathbf{x}-\mathbf{a}\|} \le \frac{\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\|_\infty}{\|\mathbf{x}-\mathbf{a}\|}$$

and so

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{|f_i(\mathbf{x}) - f_i(\mathbf{a}) - \ell_i(\mathbf{x}-\mathbf{a})|}{\|\mathbf{x}-\mathbf{a}\|} = 0$$

for all i. Thus f_i is differentiable at $\mathbf{a}$ and its derivative is the i^{th} component of the derivative of F.

□

So far we have not said how to calculate the derivative in any interesting example. To calculate DF, we shall show that there is a relation between the derivative as we have defined it and partial derivatives. We begin with the easy direction, that if a function is differentiable at a point then the partial derivatives exist and compute the total derivative:

Lemma 8.18. *Let X be an open subset of $\mathbb{R}^n$ and $\mathbf{a} \in X$. Suppose that $f\colon X \to \mathbb{R}$ is differentiable, with derivative $A = (m_1, \ldots, m_n)$. Then $m_i = \partial f(\mathbf{a})/\partial x_i$ for every i.*

Proof. Let $\mathbf{h} = t\mathbf{e}_i$. Then

$$\frac{|f(\mathbf{a}+t\mathbf{e}_i) - f(\mathbf{a}) - A(t\mathbf{e}_i)|}{\|t\mathbf{e}_i\|} = \left| \frac{f(\mathbf{a}+t\mathbf{e}_i) - f(\mathbf{a})}{t} - A(\mathbf{e}_i) \right|.$$

Taking the limit as $t \to 0$, we see that the partial derivative $\partial f(\mathbf{a})/\partial x_i$ exists and equals $A(\mathbf{e}_i) = m_i$ for every i. □

Conversely, suppose that the partial derivatives exist for every i. Is it true that f is differentiable? In general the answer is no—in fact, there are examples where f fails to be continuous at a point (exercise)—but it is true provided that all the partials are defined and continuous at every point of X:

Proposition 8.19. *Let X be an open subset of $\mathbb{R}^n$ and let $f\colon X \to \mathbb{R}$ be a function. Suppose that the partial derivatives $\partial f(\mathbf{a})/\partial x_i$ exist for every i and for all points $\mathbf{a} \in X$ and that the functions $\partial f/\partial x_i$ are continuous in X. Then f is differentiable at $\mathbf{a}$ for all $\mathbf{a} \in X$ and moreover*

$$Df_{\mathbf{a}} = \left(\frac{\partial f}{\partial x_1}(\mathbf{a}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{a}) \right)$$

as a linear map $\mathbb{R}^n \to \mathbb{R}$, or equivalently as a $1 \times n$ matrix.

Corollary 8.20. *Let X be an open subset of $\mathbb{R}^n$ and let $F\colon X \to \mathbb{R}^m$ be a function, with $F = (f_1, \ldots, f_m)$. Suppose that, for each $\mathbf{a} \in X$, the partial derivatives $\partial f_i(\mathbf{a})/\partial x_j$ exist for all i, j, and that the functions $\partial f_i/\partial x_j$ are continuous functions in X. Then F is differentiable at $\mathbf{a}$ for all $\mathbf{a} \in X$ and moreover*

$$DF_{\mathbf{a}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \ldots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a}) & \ldots & \frac{\partial f_m}{\partial x_n}(\mathbf{a}) \end{pmatrix}$$

as a linear map $\mathbb{R}^n \to \mathbb{R}$, or equivalently as an $m \times n$ matrix.

The proposition essentially says that

$$f(x_1, \ldots, x_n) \approx f(a_1, \ldots, a_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a_1, \ldots, a_n)(x_i - a_i),$$

where the precise meaning of the approximation is the statement that the difference is $o(\|\mathbf{x} - \mathbf{a}\|)$. Another and suggestive way to write this is as:

$$\Delta f \approx \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) \cdot \Delta x_i.$$

Proof of the proposition. For simplicity we write out the proof when $n = 2$ and briefly describe how to write down the proof in general. We must show that the linear map given by the partials is a good linear approximation to f, in other words we must estimate the expression

$$|f(x_1, x_2) - f(a_1, a_2) - \frac{\partial f}{\partial x_1}(a_1, a_2)(x_1 - a_1) - \frac{\partial f}{\partial x_2}(a_1, a_2)(x_2 - a_2)|.$$

To do this, use the old trick of adding and subtracting an appropriate quantity to write this as

$$\begin{aligned}
&|f(x_1,x_2)-f(x_1,a_2)+f(x_1,a_2)-f(a_1,a_2)-\\
&\frac{\partial f}{\partial x_1}(a_1,a_2)(x_1-a_1)-\frac{\partial f}{\partial x_2}(a_1,a_2)(x_2-a_2)|\\
&\leq |f(x_1,x_2)-f(x_1,a_2)-\frac{\partial f}{\partial x_2}(a_1,a_2)(x_2-a_2)|\\
&+|f(x_1,a_2)-f(a_1,a_2)-\frac{\partial f}{\partial x_1}(a_1,a_2)(x_1-a_1)|.
\end{aligned}$$

(Geometrically, this means that we are breaking up the change from (x_1,x_2) to (a_1,a_2) into a horizontal and a vertical component.) Looking at the second term first, we have by hypothesis that $f(x,a_2)$ is a differentiable function of x and so has a good linear approximation: thus

$$|f(x_1,a_2)-f(a_1,a_2)-\frac{\partial f}{\partial x_1}(a_1,a_2)(x_1-a_1)|=h_1(x_1)|x_1-a_1|$$

with $\lim_{x_1\to a_1} h_1(x_1)=0$. This first term is more involved; using the mean value theorem, there exists a real number c between a_2 and x_2 such that $f(x_1,x_2)-f(x_1,a_2)=\partial f(x_1,c)/\partial x_2(x_2-a_2)$. Thus,

$$\begin{aligned}
&\left|f(x_1,x_2)-f(x_1,a_2)-\frac{\partial f}{\partial x_2}(a_1,a_2)(x_2-a_2)\right|=\\
&=\left|\frac{\partial f}{\partial x_2}(x_1,c)(x_2-a_2)-\frac{\partial f}{\partial x_2}(a_1,a_2)(x_2-a_2)\right|\\
&=\left|\frac{\partial f}{\partial x_2}(x_1,c)-\frac{\partial f}{\partial x_2}(a_1,a_2)\right||x_2-a_2|.
\end{aligned}$$

Let

$$\begin{aligned}
h_2(x_1,x_2)&=\frac{|f(x_1,x_2)-f(x_1,a_2)-\frac{\partial f}{\partial x_2}(a_1,a_2)(x_2-a_2)|}{|x_2-a_2|}\\
&=\left|\frac{\partial f}{\partial x_2}(x_1,c)-\frac{\partial f}{\partial x_2}(a_1,a_2)\right|
\end{aligned}$$

for some c in between a_2 and x_2. Then as the partials are continuous we see that $\lim_{(x_1,x_2)\to(a_1,a_2)} h_2(x_1,x_2)=0$. Let $H(\mathbf{x})=\max\{h_1,h_2\}$. We have showed that

$$\begin{aligned}
|f(x_1,x_2)-f(a_1,a_2)-&\frac{\partial f}{\partial x_1}(a_1,a_2)(x_1-a_1)-\frac{\partial f}{\partial x_2}(a_1,a_2)(x_2-a_2)|\\
&\leq H(\mathbf{x})\left(|x_1-a_1|+|x_2-a_2|\right)\\
&=H(\mathbf{x})\|\mathbf{x}\|_1.
\end{aligned}$$

Thus f is differentiable at $\mathbf{a}$, with derivative as claimed.

The proof in the case of n variables is similar: we write

$$f(x_1,\dots,x_n) - f(a_1,\dots,a_n) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a_1,\dots,a_n)(x_i - a_i) =$$
$$= f(x_1,\dots,x_n) - f(x_1,\dots,x_{n-1},a_n) + f(x_1,\dots,x_{n-1},a_n) - f(x_1,\dots,a_{n-1},a_n)$$
$$+f(x_1,\dots,a_{n-1},a_n) - \cdots - f(a_1,\dots,a_n) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a_1,\dots,a_n)(x_i - a_i).$$

We group these terms into a sum of terms of the form

$$f(x_1,\dots,x_{i-1},x_i,a_{i+1},\dots,a_n) - f(x_1,\dots,x_{i-1},a_i,a_{i+1},\dots,a_n) - \frac{\partial f}{\partial x_i}(a_1,\dots,a_n)(x_i - a_i)$$

and argue via the mean value theorem that there exists a c_i in between a_i and x_i such that as

$$f(x_1,\dots,x_{i-1},x_i,a_{i+1},\dots,a_n) - f(x_1,\dots,x_{i-1},a_i,a_{i+1},\dots,a_n) =$$
$$= \frac{\partial f}{\partial x_i}(x_1,\dots,x_{i-1},c_i,a_{i+1},\dots,a_n)(x_i - a_i).$$

The rest of the proof proceeds as in the case $n = 2$. □

Just as for one variable functions, we can think of the derivative of $f\colon \mathbb{R}^n \to \mathbb{R}$ as a new function $Df_{\mathbf{x}}$, which we shall sometimes write as $(Df)(\mathbf{x})$. Note that $Df\colon \mathbb{R}^n \to (\mathbb{R}^n)^*$, where $(\mathbb{R}^n)^*$ is the vector space of linear maps $\mathbb{R}^n \to \mathbb{R}$. However, a linear map $\ell\colon \mathbb{R}^n \to \mathbb{R}$ is identified with a $1\times n$ matrix, i.e. a vector $\mathbf{c}$, via the rule $\ell(\mathbf{x} = \langle \mathbf{c}, \mathbf{x}\rangle$. Hence we often think of Df as the **vector-valued** function $\left(\frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n}\right)$. We this vector valued function the *gradient* of f and denote it by ∇f or $\operatorname{grad} f$. The gradient is a basic example of a *vector field*, which we shall discuss shortly. Likewise, the derivative of a vector-valued function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is the matrix valued function $DF\colon \mathbb{R}^n \to \mathbb{R}^{nm}$. Clearly, then, second derivatives will be much more complicated, and we will discuss them later.

8.3 The chain rule and its consequences

Next we come to the chain rule. It looks simple when stated in terms of linear approximations and complicated when stated in terms of partial derivatives:

Proposition 8.21. *Let X be an open subset of $\mathbb{R}^n$. Suppose that $F\colon X \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is a function and that $G\colon Y \to \mathbb{R}^k$ is a function defined in an open set containing $F(X)$. Let $\mathbf{a} \in X$ and let $\mathbf{b} = F(\mathbf{a})$. Suppose that F is differentiable at $\mathbf{a}$ with derivative A and that G is differentiable at $\mathbf{b}$ with derivative $\mathbf{b}$. Then $G \circ F$ is differentiable at $\mathbf{a}$ and its derivative is $B \circ A$. Symbolically,*

$$D(G \circ F)_{\mathbf{a}} = DG_{F(\mathbf{a})} \circ DF_{\mathbf{a}}.$$

Before we crank out the proof, let us see what the chain rule means in practice, when the functions F and G have continuous partial derivatives. In this case DF is an $m \times n$ matrix built up from the partials of F and DG is a $k \times m$ matrix built up from the partials of G, and the recipe for the derivative of $G \circ F$ is to take the matrix product. The prototype for all such rules is the following: suppose that $\mathbf{r}\colon (a,b) \to \mathbb{R}^n$ is a differentiable path, and that $f\colon X \to \mathbb{R}$ is a differentiable function (with continuous partial derivatives) defined on an open set X containing $\mathbf{r}((a,b))$. Write $\mathbf{r}(t) = (x_1(t), \ldots, x_n(t))$. Then $f \circ \mathbf{r}$ is a single function of a single variable t and the chain rule says:

$$\frac{d}{dt} f \circ \mathbf{r}(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}.$$

This follows since Df is the row vector with entries $\partial f/\partial x_i$ and $D\mathbf{r}$ is a column vector with entries dx_i/dt. All of the other cases of the chain rule are built up from this special case. For example, if $f(x,y)$ is a function of two variables and $x = x(u,v,w)$, $y = x(u,v,w)$, so that there is a map $\mathbb{R}^3 \to \mathbb{R}^2$ sending (u,v,w) to $(x(u,v,w), y(u,v,w))$, then for example

$$\frac{\partial f}{\partial v} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v}.$$

We can verify this directly by writing down the matrix product, but another way to remember the rule is that we find $\partial f/\partial v$ by adding up all the products of partials of f with respect to variables which are functions involving v, times the partial of that variable with respect to v.

Proof of the chain rule. We begin with the following lemma:

Lemma 8.22. *Let $A\colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Then there exists a constant C such that $\|A(\mathbf{x})\| \leq C\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$.*

Proof. It suffices to prove the statement for the $\|\cdot\|_\infty$ norm on $\mathbb{R}^m$ and the $\|\cdot\|_1$ norm on $\mathbb{R}^n$. Using the $\|\cdot\|_\infty$ norm has the usual effect of reducing the problem to the case of a linear function $\ell\colon \mathbb{R}^n \to \mathbb{R}$: if $A(\mathbf{x}) =$

$(\ell_1(\mathbf{x}), \ldots, \ell_m(\mathbf{x}))$, then $\|A(\mathbf{x})\|_\infty = \max_i |\ell_i(\mathbf{x})|$, so if we know that there is a constant C_i such that $|\ell_i(\mathbf{x})| \leq C_i \|\mathbf{x}\|_1$ for all $\mathbf{x} \in \mathbb{R}^n$, then take $C = \max_i C_i$ to get $\|A(\mathbf{x})\|_\infty \leq C\|\mathbf{x}\|_1$ for all $\mathbf{x}$.

So we are reduced to the case of a linear function $\ell(\mathbf{x}) = \sum_{i=1}^n a_i x_i$. Take $C = \max_i |a_i|$. Then

$$|\ell(\mathbf{x})| = |\sum_{i=1}^n a_i x_i| \leq \sum_i |a_i||x_i| \leq C \sum_i |x_i| = C\|\mathbf{x}\|_1.$$

□

We will outline another proof of the lemma in the exercises. It is an important problem in the above, using for example the usual norm, to determine the best constant C. For example, in the case where $A\colon \mathbb{R}^n \to \mathbb{R}^n$ is a linear change of coordinates, the best constant C represents the maximum amount by which A stretches length.

Returning to the proof of the chain rule, suppose that F is differentiable at $\mathbf{a}$ and that G is differentiable at $\mathbf{b} = F(\mathbf{a})$. Let $A = DF_{\mathbf{a}}$ and $B = DG_{\mathbf{b}}$. Thus there are functions $H_1(\mathbf{x})$ and $H_2(\mathbf{y})$ (defined near $\mathbf{a}$ in $\mathbb{R}^n$ and near $\mathbf{b} \in \mathbb{R}^m$ respectively) such that

$$\begin{aligned} \|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\| &= H_1(\mathbf{x})\|\mathbf{x}-\mathbf{a}\|; \\ \|G(\mathbf{y}) - G(\mathbf{b}) - B(\mathbf{y}-\mathbf{b})\| &= H_2(\mathbf{y})\|\mathbf{y}-\mathbf{b}\|, \end{aligned}$$

where $\lim_{\mathbf{x}\to\mathbf{a}} H_1(\mathbf{x}) = \lim_{\mathbf{y}\to\mathbf{b}} H_2(\mathbf{y}) = 0$. Here the functions $H_1(\mathbf{x})$ and $H_2(\mathbf{y})$ are not defined at $\mathbf{x} = \mathbf{a}$ or $\mathbf{y} = \mathbf{b}$ respectively, but we can define $H_1(\mathbf{a}) = H_2(\mathbf{b}) = 0$, and then H_1 is continuous at $\mathbf{a}$ and similarly for H_2. Moreover the above equalities still hold true for $\mathbf{x} = \mathbf{a}$ or $\mathbf{y} = \mathbf{b}$ (since then both sides are zero).

We need to show that $B \circ A$ gives a good linear approximation to $G \circ F$ at $\mathbf{a}$. Doing the usual adding and subtracting, we have

$$\begin{aligned} &\|G \circ F(\mathbf{x}) - G \circ F(\mathbf{a}) - B \circ A(\mathbf{x}-\mathbf{a})\| = \\ &\|G \circ F(\mathbf{x}) - G \circ F(\mathbf{a}) - B(F(\mathbf{x}) - F(\mathbf{a})) + B(F(\mathbf{x}) - F(\mathbf{a})) - B \circ A(\mathbf{x}-\mathbf{a})\| \\ &\leq \|G \circ F(\mathbf{x}) - G \circ F(\mathbf{a}) - B(F(\mathbf{x}) - F(\mathbf{a}))\| + \|B(F(\mathbf{x}) - F(\mathbf{a})) - B \circ A(\mathbf{x}-\mathbf{a})\|. \end{aligned}$$

We analyze these terms separately. First consider the term $\|B(F(\mathbf{x}) - F(\mathbf{a})) - B \circ A(\mathbf{x}-\mathbf{a})\|$. Using the lemma applied to B, there is a constant C_1 such that

$$\begin{aligned} \|B(F(\mathbf{x}) - F(\mathbf{a})) - B \circ A(\mathbf{x}-\mathbf{a})\| &\leq C_1\|F(\mathbf{x}) - F(\mathbf{a}) - A(\mathbf{x}-\mathbf{a})\| \\ &= C_1 H_1(\mathbf{x})\|\mathbf{x}-\mathbf{a}\|. \end{aligned}$$

As for the term $\|G \circ F(\mathbf{x}) - G \circ F(\mathbf{a}) - B(F(\mathbf{x}) - F(\mathbf{a}))\|$, we have

$$\|G \circ F(\mathbf{x}) - G \circ F(\mathbf{a}) - B(F(\mathbf{x}) - F(\mathbf{a}))\| = H_2(F(\mathbf{x}))\|F(\mathbf{x}) - F(\mathbf{a})\|,$$

where we have used $\mathbf{b} = F(\mathbf{a})$ and replaced $\mathbf{y}$ by $F(\mathbf{x})$. Now one of the usual variations on the triangle inequality gives

$$\|F(\mathbf{x}) - F(\mathbf{a})\| \leq \|A(\mathbf{x} - \mathbf{a})\| + H_1(\mathbf{x})\|\mathbf{x} - \mathbf{a}\|.$$

Also, again by the lemma, there is a constant C_2 such that $\|A(\mathbf{x} - \mathbf{a})\| \leq C_2\|\mathbf{x} - \mathbf{a}\|$. If we combine all these terms, we see that

$$\begin{gathered}\|G \circ F(\mathbf{x}) - G \circ F(\mathbf{a}) - B \circ A(\mathbf{x} - \mathbf{a})\| \leq \\ (C_1 H_1(\mathbf{x}) + H_2(F(\mathbf{x}))(C_2 + H_1(\mathbf{x})))\|\mathbf{x} - \mathbf{a}\|.\end{gathered}$$

So it suffices to show that

$$\lim_{\mathbf{x}\to\mathbf{a}} (C_1 H_1(\mathbf{x}) + H_2(F(\mathbf{x}))(C_2 + H_1(\mathbf{x}))) = 0.$$

Clearly $\lim_{\mathbf{x}\to\mathbf{a}} C_1 H_1(\mathbf{x}) = 0$. Now recall that H_2 is continuous at $\mathbf{b}$ and that $H_2(\mathbf{b}) = 0$, and that F is continuous at $\mathbf{a}$. It follows that $\lim_{\mathbf{x}\to\mathbf{a}} H_2(F(\mathbf{x}))$ exists and is zero. Thus

$$\lim_{\mathbf{x}\to\mathbf{a}} H_2(F(\mathbf{x}))(C_2 + H_1(\mathbf{x})) = 0(C_2 + 0) = 0,$$

and we are done. □

Let us use the chain rule to reinterpret the gradient. We consider a function $f\colon \mathbb{R}^n \to \mathbb{R}$, or more generally a function f defined on an open subset X of $\mathbb{R}^n$. We shall always assume that f has as many derivatives as we need, and in particular that it has continuous partial derivatives. In this circumstance we have defined the gradient ∇f. The gradient has the following properties, whose proof is left as an exercise:

1. $\nabla(f + g) = \nabla f + \nabla g$;
2. $\nabla cf = c\nabla f$;
3. $\nabla(fg) = f\nabla g + g\nabla f$.

The chain rule connects the gradient to the one-variable function we obtain by restricting f to various paths $\mathbf{r}(t)$, where $\mathbf{r}$ is differentiable and $\mathbf{r}(0) = \mathbf{a}$:

$$\frac{d}{dt} f \circ \mathbf{r}(t)\bigg|_{t=0} = \left\langle \nabla f(\mathbf{a}), \frac{d}{dt}\mathbf{r}(t)\bigg|_{t=0} \right\rangle = \langle \nabla f(\mathbf{a}), \mathbf{r}'(0)\rangle.$$

Thus, the derivative of $f \circ \mathbf{r}(t)$ is the inner product of the gradient of f with the tangent vector to the curve. In particular, the directional derivative of f at the point $\mathbf{a}$ in the direction $\mathbf{v}$, which is the special case where $\mathbf{r}(t) = \mathbf{a}+t\mathbf{v}$, is

$$\langle \nabla f(\mathbf{a}), \mathbf{v} \rangle.$$

For example, suppose that $f(x_1, x_2, x_3) = x_1^2 x_2 - x_1^4 x_3^3$ and that $\mathbf{a} = (1, 1, 1)$. Then

$$\nabla f = (2x_1x_2 - 4x_1^3x_3^3, x_1^2, -3x_1^4x_3^2), \qquad \nabla f(1, 1, 1) = (-2, 1, -3),$$

and the directional derivative of f at $(1, 1, 1)$ in the direction $(2, -3, 1)$ is

$$\langle (-2, 1, -3), (2, -3, 1) \rangle = -4 - 3 - 3 = -10.$$

One problem, given a function f, is to determine the direction where f is increasing the fastest, in other words where the directional derivative is the largest. Of course if we replace $\mathbf{v}$ by $t\mathbf{v}$ we just multiply the directional derivative. Thus it is natural to consider only unit directions $\mathbf{v}$. If $\mathbf{v}$ is a unit direction, $\|\mathbf{v}\| = 1$, then we can apply Cauchy-Schwarz:

$$|\langle \nabla f(\mathbf{a}), \mathbf{v} \rangle| \leq \|\nabla f(\mathbf{a})\| \|\mathbf{v}\| = \|\nabla f(\mathbf{a})\|.$$

Moreover equality holds only if $\mathbf{v}$ and $\nabla f(\mathbf{a})$ lie on the same line through the origin. Of course if $\nabla f(\mathbf{a}) = \mathbf{0}$ then the directional derivative is always zero. Otherwise for $\mathbf{v}$ to be a unit vector on the same line through the origin as $\nabla f(\mathbf{a})$ says exactly that

$$\mathbf{v} = \pm \frac{\nabla f(\mathbf{a})}{\|\nabla f(\mathbf{a})\|}.$$

In case the sign is positive, $\mathbf{v}$ and $\nabla f(\mathbf{a})$ point in the same direction and the value of the directional derivative is

$$\left\langle \nabla f(\mathbf{a}), \frac{\nabla f(\mathbf{a})}{\|\nabla f(\mathbf{a})\|} \right\rangle = \|\nabla f(\mathbf{a})\|,$$

whereas if the sign is negative the directional derivative is $-\|\nabla f(\mathbf{a})\|$. Summarizing:

Proposition 8.23. *If $\nabla f(\mathbf{a}) \neq \mathbf{0}$, the gradient points in the direction of steepest increase of f, in other words the unit direction $\mathbf{u}$ for which the directional derivative of f at the point $\mathbf{a}$ in the direction $\mathbf{v}$, is $\dfrac{\nabla f(\mathbf{a})}{\|\nabla f(\mathbf{a})\|}$ and the value of the directional derivative is $\|\nabla f(\mathbf{a})\|$.* □

For example, for the function $f(x_1, x_2, x_3) = x_1^2 x_2 - x_1^4 x_3^3$ and the point $(1, 1, 1)$ above, the unit direction of steepest increase is

$$\frac{(-2, 1, -3)}{\|(-2, 1, -3)\|} = \left(\frac{-2}{\sqrt{14}}, \frac{1}{\sqrt{14}}, \frac{-3}{\sqrt{14}} \right),$$

and the value of the directional derivative in this direction is $\sqrt{14}$.

The chain rule formula also tells us the following: let

$$L_c = f^{-1}(c) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = c\}$$

be the *level set* of level c (the intersection of the graph of f with the hyperplane $x_{n+1} = c$). Thus the level sets L_{c_1} and L_{c_2} are disjoint unless $c_1 = c_2$ (if $c_1 \neq c_2$ then $L_{c_1} \cap L_{c_2} = \emptyset$) and every point in $\mathbb{R}^n$ is in one and only one level set. Suppose that $\mathbf{r}(t)$ is a path contained in a fixed level set L_c, with $\mathbf{r}(0) = \mathbf{a}$. Then $f \circ \mathbf{r}(t) = c$ is constant and so

$$\left\langle \nabla f(\mathbf{a}), \left. \frac{d\mathbf{r}(t)}{dt} \right|_{t=0} \right\rangle = 0.$$

This says that $\nabla f(\mathbf{a})$ is orthogonal to the tangent vector at $\mathbf{a}$ of every curve contained in the level set and passing through $\mathbf{a}$—we say that $\nabla f(\mathbf{a})$ is *orthogonal to the level set* at $\mathbf{a}$. It is of course intuitively reasonable that the direction of steepest ascent is orthogonal to the tangents of the curves in the level set (where there is no change in the value of f). For example, the sphere S^{n-1} is the level set of the function $f(\mathbf{x}) = \|\mathbf{x}\|^2 = x_1^2 + \cdots + x_n^2$ at level 1. Here, we use the square of the norm as it is simpler to take the gradient, and in fact

$$\nabla \|\mathbf{x}\|^2 = 2\mathbf{x}.$$

Thus, if $\mathbf{r}$ is a curve contained in S^{n-1} and such that $\mathbf{r}(t) = \mathbf{a}$, then the tangent vector $\mathbf{r}'(t)$ is orthogonal to $\mathbf{a}$, which we have already computed by a different method in Corollary 8.14.

As we have seen, the condition $\nabla f(\mathbf{a}) \neq \mathbf{0}$ is a natural condition to impose if we want to analyze directional derivatives in a meaningful way. Hence we make the following definition:

Definition 8.24. The point $\mathbf{a}$ is a *critical point of* f if $\nabla f(\mathbf{a}) = \mathbf{0}$. The point $\mathbf{a}$ is a *regular point of* f if it is not a critical point.

If $\mathbf{a}$ is not a critical point of f, and $f(\mathbf{a}) = c$, we define the *tangent hyperplane* to the level set L_c at $\mathbf{a}$ to be the set

$$\{\mathbf{x} \in \mathbb{R}^n : \langle \nabla f(\mathbf{a}), \mathbf{x} - \mathbf{a} \rangle = 0.$$

Up to a translation, the tangent hyperplane is the vector subspace $\{\nabla f(\mathbf{a})\}^\perp$ of dimension $n-1$, but it is translated so that it passes through the point $\mathbf{a}$. An important special case is when we look at the graph of a function f: in this case, the graph is the level set L_0 in $\mathbb{R}^{n+1}$ of the function $g(x_1,\ldots,x_{n+1}) = x_{n+1} - f(x_1,\ldots,x_n)$. Note that the gradient of the function g, i.e. $\nabla g(a_1,\ldots,a_{n+1})$, is **not** the same as $\nabla f(a_1,\ldots,a_n)$; in fact,

$$\nabla g(a_1,\ldots,a_{n+1}) = (-\nabla f(a_1,\ldots,a_n), 1).$$

In particular it is never $\mathbf{0}$. In this case, the tangent hyperplane through the point $(a_1,\ldots,a_{n+1})$ is given by

$$\langle(-\nabla f(a_1,\ldots,a_n),1),(x_1,\ldots,x_{n+1}) - (a_1,\ldots,a_{n+1})\rangle = 0.$$

The point $(a_1,\ldots,a_{n+1})$ satisfies $g(a_1,\ldots,a_{n+1}) = 0$, i.e. $(a_1,\ldots,a_{n+1})$ lies on L_0, if and only if $a_{n+1} = f(a_1,\ldots,a_n)$. Thus the equation for the tangent hyperplane is

$$x_{n+1} = f(a_1,\ldots,a_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a_1,\ldots,a_n)\cdot(x_i - a_i),$$

which is the equation for the linear approximation to f.

We should also look at points for which the tangent hyperplane fails to exist, namely critical points. The main importance of critical points is the following: if $\mathbf{a}$ is a local maximum or minimum of f, then it is a critical point. This is clear, since the function $f(\mathbf{a} + t\mathbf{e}_i)$ must then have a local maximum or minimum at zero, and by applying one-variable calculus.

We should expect a second derivative test for when a function has a local maximum or minimum, neither, or we cannot say, at a critical point $\mathbf{a}$. We will discuss this later, once we have introduced higher partial derivatives.

The gradient is an example of a *vector field*:

Definition 8.25. Let X be an open subset of $\mathbb{R}^n$. A *vector field* $\mathbf{E}\colon X \to \mathbb{R}^n$ is a function (usually assumed differentiable) from $X \subseteq \mathbb{R}^n$ to $\mathbb{R}^n$. A vector field $\mathbf{E}$ is a gradient vector field if it is of the form ∇f for some differentiable function f.

In components a vector field $\mathbf{E}$ looks like $\mathbf{E} = (h_1,\ldots,h_n)$. Intuitively, we think of $\mathbf{E}$ as a function which assigns to each point $\mathbf{x}$ of $\mathbb{R}^n$ a vector $\mathbf{E}(\mathbf{x})$ which we locate at $\mathbf{x}$ and think of as a force acting on a test particle, say, at that point. To say that $\mathbf{E}$ is a gradient vector field is to say that there is an unknown function f such that $h_i = \partial f/\partial x_i$ for every i. For example

if $\mathbf{E} = \mathbf{r}$, i.e. $h_i = x_i$ for every i, then we can take $f(\mathbf{x}) = \|\mathbf{x}\|^2/2$. But a general vector field is not a gradient vector field. The main condition is the following: if a function f has continuous second partial derivatives, then, as we shall show when we discuss the equality of mixed partials, for all $i \neq j$,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Thus if $\mathbf{E} = (h_1, \ldots, h_n)$ is a gradient vector field, where the h_i are reasonable functions (continuous partial derivatives, for example), then we have the conditions

$$\frac{\partial h_i}{\partial x_j} = \frac{\partial h_j}{\partial x_i}$$

for all $i \neq j$. For example, take the vector field $\mathbf{E} = (-x_2, x_1)$. Then

$$\frac{\partial h_1}{\partial x_2} = \frac{\partial}{\partial x_2}(-x_2) = -1 \neq \frac{\partial h_2}{\partial x_1} = \frac{\partial}{\partial x_1}(x_1).$$

Thus $\mathbf{E}$ is not a gradient vector field.

In the language of physics, finding a function f such that $\mathbf{E} = \nabla f$ is the same as finding a *potential* for $\mathbf{E}$, and vector fields which have potentials are called *conservative* (because potential energy, suitably interpreted, is a conserved quantity, not because of any political orientation). The level sets are then called equipotential curves or surfaces, since the potential energy does not change along the level set. There is also a connection with the concept of work, which we shall discuss later.

8.4 Higher derivatives

Let U be an open subset of $\mathbb{R}^n$. We shall just consider a single real-valued function $f\colon U \to \mathbb{R}$. In this case, the derivative of f is really n functions $\partial f/\partial x_i$ from U to $\mathbb{R}$. Of course, we can repeat this process, if the derivatives in question exist, to define

$$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

We also abbreviate

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_i^2}.$$

Higher expressions such as

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}$$

are similarly defined. Such expressions are called the k^{th} partials or partial derivatives of order k. Partials of order 2 can be described by the $n \times n$ matrix $H(f) = \left(\dfrac{\partial^2 f}{\partial x_j \partial x_i} \right)$ which is called the *Hessian* of f. The collection of all higher partials of order k is described by n^k numbers and is an example of a *tensor*.

We say that the function f is *of class* $C^k(U)$ if all such expressions are defined and continuous on U (we omit the reference to the open set U if it is clear from the context). Thus, if $f \in C^k(U)$, then $f \in C^\ell(U)$ for all $\ell \leq k$. Finally, we set $C^\infty(U) = \bigcap_{k \geq 1} C^k(U)$. Thus $C^\infty(U)$ is the set of all functions which have partial derivatives of every order. The sets $C^k(U)$ and $C^\infty(U)$ are (infinite dimensional) vector spaces.

Just as a function has n first partial derivatives, it has n^2 second partial derivatives and n^k partial derivatives of order k. But in fact under reasonable circumstances it has fewer. The following is called **equality of mixed partials**.

Theorem 8.26. *Suppose that $f \in C^2(U)$. Then for all i and j, we have*

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Proof. It is enough to consider the case where $i = 1$ and $j = 2$, and to ignore the remaining variables. Fix a point (a_1, a_2) and let as usual $x_1 - a_1 = h_1, x_2 - a_2 = h_2$. Consider the following expression, where the variables x_1 and x_2 play symmetric roles:

$$S(h_1, h_2) = f(a_1 + h_1, a_2 + h_2) - f(a_1 + h_1, a_2) - f(a_1, a_2 + h_2) + f(a_1, a_2).$$

We claim that

$$\lim_{\substack{(h_1,h_2) \to (0,0) \\ h_1 \neq 0, h_2 \neq 0}} \frac{S(h_1, h_2)}{h_1 h_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}(a_1, a_2) = \frac{\partial^2 f}{\partial x_1 \partial x_2}(a_1, a_2),$$

which clearly proves the theorem. Fix for the moment a_2, h_2 and define

$$g(x_1) = f(x_1, a_2 + h_2) - f(x_1, a_2).$$

By construction $S(h_1, h_2) = g(a_1 + h_1) - g(a_1)$. Now since $g(x_1)$ is differentiable, the mean value theorem says that there exists a c_1 between a_1 and $a_1 + h_1$ such that

$$S(h_1, h_2) = g'(c_1) \cdot h_1 = \left[\frac{\partial f}{\partial x_1}(c_1, a_2 + h_2) - \frac{\partial f}{\partial x_1}(c_1, a_2) \right] h_1.$$

We can also apply the mean value theorem to the expression

$$\frac{\partial f}{\partial x_1}(c_1, a_2 + h_2) - \frac{\partial f}{\partial x_1}(c_1, a_2);$$

thus there is a c_2 between a_2 and $a_2 + h_2$ such that

$$\frac{\partial f}{\partial x_1}(c_1, a_2 + h_2) - \frac{\partial f}{\partial x_1}(c_1, a_2) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(c_1, c_2) h_1 h_2.$$

Dividing by $h_1 h_2$, which makes sense as long as h_1 and h_2 are nonzero, we see that

$$\frac{S(h_1, h_2)}{h_1 h_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}(c_1, c_2).$$

Now since $\dfrac{\partial^2 f}{\partial x_2 \partial x_1}$ is continuous, the limit of $\dfrac{\partial^2 f}{\partial x_2 \partial x_1}(c_1, c_2)$ as $(h_1, h_2) \to (0, 0)$ is $\dfrac{\partial^2 f}{\partial x_2 \partial x_1}(a_1, a_2)$, since c_1 lies between a_1 and $a_1 + h_1$ and c_2 lies between a_2 and $a_2 + h_2$. This shows that

$$\lim_{\substack{(h_1,h_2)\to(0,0) \\ h_1 \neq 0, h_2 \neq 0}} \frac{S(h_1, h_2)}{h_1 h_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}(a_1, a_2),$$

and the other limit

$$\lim_{\substack{(h_1,h_2)\to(0,0) \\ h_1 \neq 0, h_2 \neq 0}} \frac{S(h_1, h_2)}{h_1 h_2} = \frac{\partial^2 f}{\partial x_1 \partial x_2}(a_1, a_2)$$

is proved by a symmetric argument. □

As a result, in any expression

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}},$$

we can rearrange the order of the x_{i_j} arbitrarily, and the resulting expression will be the same as the original one provided that f is of class C^k. (This result fails if f is only assumed to have partial derivatives of order k, without assuming that they are continuous.) This says that the tensor defined by the $\partial^k f / \partial x_{i_1} \cdots \partial x_{i_k}$ is *symmetric*.

8.5 Taylor's theorem

Next we turn to Taylor's theorem in several variables. To begin we recall the statement and proof of Taylor's theorem in one variable. We will not give the strongest statement here but just one suited to our purposes.

Theorem 8.27 (Taylor's theorem with remainder)**.** *Let $g(t)$ be a function defined in some open interval I containing a and x, and suppose that the higher derivatives $g^{(i)}$ exist for $i = 1, 2, \ldots, m+1$. Then there exists a real number c between a and x such that*

$$g(x) = g(a) + g'(a)(x-a) + \cdots + \frac{g^{(m)}(a)}{m!}(x-a)^m + \frac{g^{(m+1)}(c)}{(m+1)!}(x-a)^{m+1}.$$

Here the expression

$$P_m(x) = g(a) + g'(a)(x-a) + \cdots + \frac{g^{(m)}(a)}{m!}(x-a)^m$$

is called the m^{th} order Taylor polynomial for g at a and the difference

$$g(x) - P_m(x) = R_m(x)$$

is called the m^{th} order remainder term; the statement of Taylor's theorem is just a description of the remainder term in terms of the unknown number c. For example, $P_1(x) = g(a) + g'(a)(x-a)$ is just the tangent line approximation to $g(x)$. But even in this case we will see that the theorem says something new.

Proof of Taylor's theorem. The proof is in a certain sense very similar to the proof of the Mean Value Theorem. Suppose for example that $x > a$, and set $x = b$ to emphasize the similarity with the proof of the mean value theorem. Define a function $H(t)$ by the formula

$$H(t) = g(t) - \left(P_m(t) + \frac{k}{(m+1)!}(t-a)^{m+1} \right),$$

where as above $P_m(t) = g(a) + g'(a)(t-a) + \cdots + \dfrac{g^{(m)}(a)}{m!}(t-a)^m$ is the Taylor polynomial and k is chosen to be the unique real number such that $H(b) = 0$. We can always find the number k since $b - a \neq 0$:

$$k = \frac{(m+1)!(g(b) - P_m(b))}{(b-a)^{m+1}}.$$

Thus

$$g(b) = P_m(b) + \frac{k}{(m+1)!}(b-a)^{m+1}$$

and so the theorem is equivalent to the statement that $k = g^{(m+1)}(c)$ for some c between a and b. Note that the mean value theorem is the case $m = 0$ of Taylor's theorem and in this case $k = \dfrac{g(b) - g(a)}{b-a}$. If we look at the statement of Taylor's theorem for $x = b$, using the fact that by construction $H(b) = 0$, it is just the statement that there is a c between a and b such that $k = g^{(m+1)}(c)$. To prove this, note that $g(a) + g'(a)(t - a) + \cdots + \dfrac{g^{(m)}(a)}{m!}(t-a)^m$ has been chosen to have the same derivatives at a as $g(t)$ up through order m, whereas $\dfrac{k}{(m+1)!}(t-a)^{m+1}$ has its first m derivatives at a equal to zero. Thus the first m derivatives of $H(t)$ at a are all zero, and the $(m+1)^{\text{st}}$ derivative of $H(t)$ at a point t is $g^{(m+1)}(t) - k$, since the $(m+1)^{\text{st}}$ derivative of a polynomial of degree m is zero and the $(m+1)^{\text{st}}$ derivative of $(t-a)^{m+1}$ is the constant $(m+1)!$.

Now since $H(a) = H(b) = 0$, we can apply Rolle's theorem to find a c_1, $a < c_1 < b$, such that $H'(c_1) = 0$. Likewise, if $m \geq 1$, since by construction $H'(a) = 0$, we can apply Rolle's theorem again, using the fact that $H'(a) = H'(c_1) = 0$, to find a c_2, with $a < c_2 < c_1 < b$, such that $H''(c_2) = 0$. Continuing in this way, we can find $a < c_m < c_{m-1} < \cdots < c_1$ such that $H^{(m)}(c_m) = 0$. Finally, using $H^{(m)}(a) = H^{(m)}(c_m) = 0$, we can find a c with $a < c < c_m < b$, such that $H^{(m+1)}(c) = 0$. But by the remarks above,

$$0 = H^{(m+1)}(c) = g^{(m+1)}(c) - k,$$

and thus $k = g^{(m+1)}(c)$ for some c with $a < c < b$. This concludes the proof of Taylor's theorem. □

Note that we don't actually need to assume that $g \in C^{m+1}(I)$ in the above argument, only that g has $m+1$ derivatives. However, if we want to be able to say anything meaningful about the behavior of g, we would like to say that since g^{m+1} is continuous it is bounded near a, and thus

$$|g(x) - P_m(x)| = |R_m(x)| \leq C_m|x-a|^{m+1}$$

for some constant C_m, depending on m, where $P_m(x)$ is the degree m Taylor polynomial at a. In this case $g(x) - P_m(x) = R_m(x)$ is $O(|x-a|^{m+1})$ and in particular it is $o(|x-a|^m)$. Moreover, an explicit bound for g^{m+1} in an interval leads to an explicit bound, valid over the whole interval. Thus,

for example, in the case $m = 1$, the theorem says that is $g \in C^2(I)$, then $g(x) - g(a) - g'(a)(x - a)$ is $O(|x - a|^2)$, i.e. bounded by $C(x - a)^2$. This is of course a better statement than just saying $g(x)$ is $o(|x - a|)$.

Note that in the above discussion we have been talking about **Taylor polynomials** and how well they approximate a function but not about **Taylor series**, in other words the power series $\sum_{i=0}^{\infty} \frac{g^i(a)}{i!}(x - a)^i$. The issue of whether or not the power series converges to the original function is a question of whether the remainder goes to zero for some range of x with $|x - a| < r$ as $n \to \infty$. There exist C^∞ functions where the Taylor series has a radius of convergence equal to zero (in other words, converges only for $x = a$) and other functions for which the series converges but not to the original function. However it takes some work to construct such functions. For example, the function $g(x)$ defined by

$$g(x) = \begin{cases} e^{-1/x^2}, & \text{if } x > 0; \\ 0, & \text{if } x \leq 0, \end{cases}$$

can be checked to be C^∞, but its derivatives at 0 are all 0 and hence its Taylor series is identically zero. On the other hand, a function where we can bound $g^m(x)$ independently of m, for example $\sin x$ or $\cos x$, where $|g^m(x)| \leq 1$, or e^x, where $g^m(x) = g(x)$ for all m, will have a Taylor series which converges everywhere to the original function, by looking at the remainder term. In general, a function g defined and C^∞ in an interval I whose Taylor series converges everywhere in I to g is called *analytic*; this is a stronger property than being C^∞.

We turn now to the several variables version of Taylor's theorem. In fact. this version is no more difficult than the one variable version, but it is notationally more complicated. Assume that $f(x_1, \ldots, x_n)$ is of class C^{m+1} in an open ball B around $\mathbf{a} = (a_1, \ldots, a_n)$ (or more generally in an open convex set containing $\mathbf{a}$). We suppose that $\mathbf{x} = (x_1, \ldots, x_n) \in B$ and write as usual $\mathbf{h} = (h_1, \ldots, h_n) = \mathbf{x} - \mathbf{a}$. The idea is that we can reduce Taylor's theorem for f to a one variable problem by looking at the line segment $\mathbf{a}+t\mathbf{h}$ joining $\mathbf{a}$ to $\mathbf{x}$, and evaluating at $t = 1$. To do this, we will need to take the m^{th} derivatives of the function $f(\mathbf{a} + t\mathbf{h})$. However, this will just be a repeated application of the chain rule. Thus:

$$\frac{d}{dt} f(\mathbf{a} + t\mathbf{h}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a} + t\mathbf{h}) \cdot \frac{d}{dt}(th_i) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a} + t\mathbf{h}) \cdot h_i.$$

To find the second derivative, we take the first derivative of each of the terms $\frac{\partial f}{\partial x_i}(\mathbf{a}+t\mathbf{h})$, and each such term gives

$$\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}+t\mathbf{h})h_j.$$

Thus, if we multiply this term by h_i and sum over i, we see that

$$\frac{d^2}{dt^2}f(\mathbf{a}+t\mathbf{h}) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}+t\mathbf{h})h_i h_j.$$

The general case is similar:

$$\frac{d^k}{dt^k}f(\mathbf{a}+t\mathbf{h}) = \sum_{i_1,\ldots,i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_1}\cdots\partial x_{i_k}}(\mathbf{a}+t\mathbf{h})h_{i_1}\cdots h_{i_k}.$$

If we apply Taylor's theorem to the function $f(\mathbf{a}+t\mathbf{h})$ and take $t=1$, we see that we have the following:

Theorem 8.28. *Let $f(x_1,\ldots,x_n)$ be of class C^{m+1} in an open ball B around $\mathbf{a}$. Then*

$$f(\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a})(x_i-a_i) + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a})(x_i-a_i)(x_j-a_j) + \cdots +$$
$$+\frac{1}{m!}\sum_{i_1,\ldots,i_m=1}^{n} \frac{\partial^m f}{\partial x_{i_1}\cdots\partial x_{i_m}}(\mathbf{a})(x_{i_1}-a_{i_1})\cdots(x_{i_m}-a_{i_m}) +$$
$$+\frac{1}{(m+1)!}\sum_{i_1,\ldots,i_{m+1}=1}^{n} \frac{\partial^{m+1} f}{\partial x_{i_1}\cdots\partial x_{i_{m+1}}}(\mathbf{c})(x_{i_1}-a_{i_1})\cdots(x_{i_{m+1}}-a_{i_{m+1}}),$$

where $\mathbf{c}$ is some vector on the line segment between $\mathbf{x}$ and $\mathbf{a}$. □

Just as in the one variable case we will think of the formula above as saying that

$$f(\mathbf{x}) = P_m(\mathbf{x}) + R_m(\mathbf{x}),$$

where $P_m(\mathbf{x})$ is a polynomial in several variables and $R_m(\mathbf{x})$ is the remainder term. We can also ask about power series in several variables and define analytic functions, but we shall not do so here.

Note that, by the equality of mixed partials, many of the terms in the sums above are the same. For example, for $i \neq j$ the degree two term has the

two equal terms $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a})(x_i - a_i)(x_j - a_j)$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a})(x_i - a_i)(x_j - a_j)$ and so the full degree two term is just

$$\frac{1}{2}\left(\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(\mathbf{a})(x_i - a_i)^2 + 2\sum_{i<j} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a})(x_i - a_i)(x_j - a_j)\right).$$

We give some notation (which we shall not use) to write the Taylor series efficiently without repeating the mixed partials. Let $\mathbf{k} = (k_1, \ldots, k_n)$ be a vector, all of whose entries are nonnegative integers. We set

$$\begin{aligned} |\mathbf{k}| &= k_1 + \cdots + k_n; \\ \mathbf{k}! &= k_1! \cdots k_n!; \\ (\mathbf{x} - \mathbf{a})^{\mathbf{k}} &= (x_1 - a_1)^{k_1} \cdots (x_n - a_n)^{k_n}; \\ \frac{\partial^{|\mathbf{k}|}}{\partial \mathbf{x}^{\mathbf{k}}} &= \frac{\partial^{|\mathbf{k}|}}{\partial x_1^{k_1} \cdots \partial x_n^{k_n}}. \end{aligned}$$

With this notation the degree m Taylor polynomial is given by

$$P_m(\mathbf{x}) = \sum_{|\mathbf{k}| \leq m} \frac{1}{\mathbf{k}!} \frac{\partial^{|\mathbf{k}|}}{\partial \mathbf{x}^{\mathbf{k}}} f(\mathbf{a})(\mathbf{x} - \mathbf{a})^{\mathbf{k}}$$

and the remainder term is

$$R_m(\mathbf{x}) = \sum_{|\mathbf{k}| = m+1} \frac{1}{\mathbf{k}!} \frac{\partial^{|\mathbf{k}|}}{\partial \mathbf{x}^{\mathbf{k}}} f(\mathbf{c})(\mathbf{x} - \mathbf{a})^{\mathbf{k}},$$

where $\mathbf{c}$ is on the line segment between $\mathbf{a}$ and $\mathbf{x}$.

Note: the procedure of passing from the symmetric tensor $\frac{\partial^m f}{\partial x_{i_1} \cdots \partial x_{i_m}}(\mathbf{a})$ to an associated polynomial is really the generalization of the discussion of symmetric multilinear forms and their relation to homogeneous polynomials.

In a concrete example, one does not find out a Taylor polynomial by computing derivatives and plugging these into the formula, but rather by using what is known from one variable calculus and then plugging in. For example, suppose that we wish to work out the Taylor polynomial of degree two for $f(x_1, x_2) = x_1 e^{x_1 x_2} + (1 + x_1^2 + x_2^3)\cos(x_1 + x_2)$ at $(0, 0)$. Plugging into the usual formulas, we find that

$$\begin{aligned} f(x_1, x_2) &= x_1(1 + x_1 x_2 + \frac{(x_1 x_2)^2}{2!} + \ldots) + (1 + x_1^2 + x_2^3)(1 - \frac{(x_1 + x_2)^2}{2!} + \ldots) \\ &= x_1 + x_1^2 x_2 + 1 - \frac{1}{2}(x_1 + x_2)^2 + x_1^2 + \ldots \end{aligned}$$

The terms of degree at most two in this power series give $1 + x_1 + \frac{1}{2}x_1^2 - x_1x_2 - \frac{1}{2}x_2^2$, and so this is the Taylor polynomial $P_2(\mathbf{x})$. Working backward, this information also tells us that $f(0,0) = 0$,

$$\frac{\partial f}{\partial x_1}(0,0) = 1, \frac{\partial f}{\partial x_2}(0,0) = 0,$$
$$\frac{\partial^2 f}{\partial x_1^2}(0,0) = 1, \frac{\partial^2 f}{\partial x_1 \partial x_2}(0,0) = \frac{\partial^2 f}{\partial x_2^2}(0,0) = -1.$$

We turn now to the size of the remainder $R_m(\mathbf{x})$. Assuming that $f \in C^{m+1}$, all of the partials of order $m+1$ are continuous and thus bounded in an open set around $\mathbf{a}$. Thus

$$|R_m(\mathbf{x})| \leq C \sum_{i_1,\ldots,i_{m+1}=1}^{n} |h_{i_1} \cdots h_{i_{m+1}}|.$$

Now each expression $|h_{i_1} \cdots h_{i_{m+1}}|$ is an example of a continuous homogeneous function of degree $m+1$, in other words a function H such that $H(t\mathbf{h}) = t^{m+1}H(\mathbf{h})$ for all $t \in \mathbb{R}$, $t > 0$, and $\mathbf{h} \in \mathbb{R}^n$. To estimate the size of such a function, for small $\mathbf{h}$, we have the following lemma:

Lemma 8.29. *Let $H\colon \mathbb{R}^n \to \mathbb{R}$ be a continuous homogeneous function of degree k, where k is a positive real number, in other words $H(t\mathbf{h}) = t^k H(\mathbf{h})$ for all $t \in \mathbb{R}$, $t > 0$, and $\mathbf{h} \in \mathbb{R}^n$. Then there exists a constant C such that, for all $\mathbf{h} \in \mathbb{R}^n$,*

$$|H(\mathbf{h})| \leq C\|\mathbf{h}\|^k.$$

Proof. First suppose that $\|\mathbf{h}\| = 1$, so that $\mathbf{h}$ lies on the unit sphere in $\mathbb{R}^n$, which is compact. Then since H is continuous, it is bounded on the unit sphere, and thus there is a constant C such that $H(\mathbf{h}) \leq C$. Now, just as in the arguments on norms, we can always multiply a nonzero vector $\mathbf{h}$ by the scalar $1/\|\mathbf{h}\|$ so that $\mathbf{h}/\|\mathbf{h}\|$ has norm one. Thus

$$H\left(\frac{\mathbf{h}}{\|\mathbf{h}\|}\right) = \frac{H(\mathbf{h})}{\|\mathbf{h}\|^k} \leq C,$$

so that $|H(\mathbf{h})| \leq C\|\mathbf{h}\|^k$ as long as $\mathbf{h} \neq 0$. On the other hand, since H is continuous and $k > 0$, it is easy to see by taking $\mathbf{h} \to 0$ that $H(0) = 0$ and thus the inequality $|H(\mathbf{h})| \leq C\|\mathbf{h}\|^k$ is satisfied for $\mathbf{h} = 0$ as well. □

Thus the remainder term satisfies

$$|R_m(\mathbf{x})| \leq C\|\mathbf{x} - \mathbf{a}\|^{m+1},$$

and in particular $|R_m(\mathbf{x})|$ is $o(\|\mathbf{x} - \mathbf{a}\|^m)$.

8.6 Classification of critical points

Let us apply Taylor's formula with remainder to the study of local maxima and minima in several variables. We are after a generalization of the second derivative test in one variable. First let us note the analogue of the first derivative test.

Lemma 8.30. *Suppose that $f \in C^1(U)$ and that $\mathbf{a}$ is a local maximum or minimum for f. Then $\mathbf{a}$ is a critical point for f, in other words $\nabla f(\mathbf{a}) = \mathbf{0}$.*

Proof. If we restrict f to a line $\mathbf{a} + t\mathbf{v}$, it will still have a local maximum or minimum on this line. So by the one variable result

$$\frac{df}{dt}(\mathbf{a} + t\mathbf{v})\bigg|_{t=0} = 0.$$

But this derivative is just $\langle \nabla f(\mathbf{a}), \mathbf{v} \rangle$. It follows that $\nabla f(\mathbf{a})$ is orthogonal to every vector $\mathbf{v}$ and thus is zero, as claimed. □

Now suppose that f is a C^3 function and look at the second order Taylor polynomial for f. Using the fact that $\nabla f(\mathbf{a}) = 0$, we see that this gives

$$f(\mathbf{x}) = f(\mathbf{a}) + \frac{1}{2}\left(\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(\mathbf{a}) h_i^2 + 2 \sum_{i<j} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) h_i h_j \right) + R_2(\mathbf{x}),$$

where $|R_2(\mathbf{x})| \leq C\|\mathbf{h}\|^3$. Let us write the second expression as $Q(\mathbf{h})$, where $Q(\mathbf{h})$ has the general form

$$Q(\mathbf{h}) = \sum_i a_{ii} h_i^2 + 2 \sum_{i<j} a_{ij} h_i h_j$$

for some constants a_{ij}. Thus Q is a quadratic form whose associated matrix is

$$H(f) = \left(\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) \right),$$

the *Hessian* of f. (Strictly speaking we only define the Hessian at a critical point.) The fact that the Hessian is symmetric is just a restatement of the equality of mixed partials.

Let Q be a quadratic form such that $Q(\mathbf{h}) = 0$ if and only if $\mathbf{h} = \mathbf{0}$. We call such a form Q *definite*. In fact, since for $n > 1$ $\mathbb{R}^n - \{\mathbf{0}\}$ is connected, it follows from the intermediate value theorem that a definite quadratic form Q on $\mathbb{R}^n$ for $n > 1$ must satisfy: for $\mathbf{h} \neq \mathbf{0}$, $Q(\mathbf{h})$ is always > 0 or always

< 0. In fact, the same is true for a one variable quadratic form as well since in this case by definition $Q(h) = ah^2$ with $a \neq 0$. Thus Q is either positive definite or negative definite as previously defined.

Definite forms have the following property, which once again is reminiscent of our discussion of norms:

Lemma 8.31. *Suppose that $Q(\mathbf{h})$ is a positive definite form. Then there exists a constant $m > 0$ such that, for all $\mathbf{h}$,*

$$Q(\mathbf{h}) \geq m\|\mathbf{h}\|^2.$$

Likewise, if Q is negative definite, then there exists a constant $M < 0$ such that, for all $\mathbf{h}$,

$$Q(\mathbf{h}) \leq M\|\mathbf{h}\|^2.$$

Proof. We will just do the positive definite case. Consider the restriction of Q to the unit sphere $\{\mathbf{h} : \|\mathbf{h}\| = 1\}$. Then since the unit sphere is compact, Q attains its minimum value m there. Since Q is positive definite, $m > 0$. By definition, for all $\mathbf{h}$ such that $|\mathbf{h}\| = 1$, $Q(\mathbf{h}) \geq m$. Now for a general $\mathbf{h}$ with $\mathbf{h} \neq 0$, $\mathbf{h}/|\mathbf{h}\|$ lies on the unit sphere, and thus

$$Q\left(\frac{\mathbf{h}}{|\mathbf{h}\|}\right) = \frac{Q(\mathbf{h})}{\|\mathbf{h}\|^2} \geq m.$$

It follows that $Q(\mathbf{h}) \geq m\|\mathbf{h}\|^2$ as claimed. □

Suppose now that the Hessian of f is positive definite at the critical point $\mathbf{a}$, and look at the Taylor polynomial for f. It follows that this has the general form

$$f(\mathbf{x}) = f(\mathbf{a}) + Q(\mathbf{h}) + R_2(\mathbf{x}) \geq f(\mathbf{a}) + m\|\mathbf{h}\|^2 - C\|\mathbf{h}\|^3,$$

at least for small $\|\mathbf{h}\|$. If $\|\mathbf{h}\|$ is small, then

$$m\|\mathbf{h}\|^2 - C\|\mathbf{h}\|^3 = \|\mathbf{h}\|^2(m - C\|\mathbf{h}\|) > 0$$

and thus $f(\mathbf{x}) > f(\mathbf{a})$ for all $\mathbf{x}$ such that $\mathbf{h} = \mathbf{x} - \mathbf{a}$ is sufficiently small. In other words, $\mathbf{a}$ is a local minimum for f. Likewise if the Hessian of f is negative definite at $\mathbf{a}$, then $\mathbf{a}$ is a local maximum for f. Summarizing:

Theorem 8.32. *Suppose f is of class C^3 and that $\mathbf{a}$ is a critical point for f. If the Hessian of f is positive definite at $\mathbf{a}$, then $\mathbf{a}$ is a local minimum for f. If the Hessian of f is negative definite at $\mathbf{a}$, then $\mathbf{a}$ is a local maximum for f.*

In case Q is indefinite then $\mathbf{a}$ is neither a local maximum nor a local minimum for f, and such a point is loosely called a *saddle point* for f. We will try to describe these points more precisely below. It might be that Q is just semidefinite (or even zero) and in these cases no information can be obtained from the second derivative.

The question now becomes: how do we decide if the Hessian of f is definite at $\mathbf{a}$? This is just a linear algebra question about the quadratic form f, and we have already described the answer. Combining this with Theorem 8.32 above gives the following in case $n = 2$:

Theorem 8.33. *Let $f(x_1, x_2) \in C^3$ and let $\mathbf{a}$ be a crtical point for f. Let $H(f)(\mathbf{a})$ be the Hessian of f at $\mathbf{a}$.*

1. *If $\dfrac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) > 0$ and $\det H(f)(\mathbf{a}) > 0$, then $\mathbf{a}$ is a local minimum for f.*
2. *If $\dfrac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) < 0$ and $\det H(f)(\mathbf{a}) > 0$, then $\mathbf{a}$ is a local maximum for f.*
3. *If $\det H(f)(\mathbf{a}) < 0$, then $\mathbf{a}$ is a saddle point for f.*
4. *If $\det H(f)(\mathbf{a}) = 0$, then the second derivative test gives no information.*

Anything can happen if $\det H(f)(\mathbf{a}) = 0$. If $f(x_1, x_2) = x_1^2$, then $(0, 0)$ is a critical point but it is not a local minimum since the minimum value is attained all along the x_2-axis. If $f(x_1, x_2) = x_1^2 + x_2^4$, then $(0, 0)$ is a critical point and a local minimum. If $f(x_1, x_2) = x_1^2 + x_2^3$, then $(0, 0)$ is a critical point and f can be either positive or negative for values of (x_1, x_2) close to $(0, 0)$. We usually reserve the term saddle point for a critical point $\mathbf{a}$ such that $\det H(f)(\mathbf{a}) < 0$, and any critical point $\mathbf{a}$ such that $\det H(f)(\mathbf{a}) \neq 0$ (either a local maximum, a local minimum, or a saddle point) is called a *nondegenerate critical point*.

In higher dimensions, we have seen the following criterion for when a quadratic form corresponding to a symmetric matrix A is positive or negative definite: let $A = (a_{ij})$ be an $n \times n$ symmetric matrix, and let A_r be the $r \times r$ symmetric matrix given by

$$A_r = (a_{ij})_{\substack{1 \leq i \leq r \\ 1 \leq j \leq r}}.$$

Then A is positive definite if and only if

$$\det A_1 > 0, \det A_2 > 0, \ldots, \det A_n > 0,$$

and A is negative definite if and only if

$$\det A_1 < 0, \det A_2 > 0, \det A_3 < 0, \ldots,$$

and $\det A_n$ is either positive or negative depending on whether n is even or odd. Thus, if $A = H(f)$ satisfies the above condition for being positive or negative definite, then $\mathbf{a}$ is either a local minimum for f (if $H(f)$ is positive definite) or a local maximum for f (if $H(f)$ is negative definite). In the case of a critical point of a function f, if $\det H(f)(\mathbf{a}) \neq 0$ but $H(f)(\mathbf{a})$ is not definite we call $\mathbf{a}$ a saddle point again, and refer to every critical point $\mathbf{a}$ such that $\det H(f)(\mathbf{a}) \neq 0$ (either a local maximum, a local minimum, or a saddle point) as a *nondegenerate critical point*.

We conclude by mentioning a more real world type of maximum/minimum question. Just as in one variable, where we realistically have to look at functions defined on closed intervals and check the end points, in several variables we must look at functions defined on *compact* sets if we want a guarantee that extreme points will exist. Typically such a compact set will have an open part, the interior, plus a boundary which is a union (in the case of a compact subset of $\mathbb{R}^2$, say) of finitely many curves. For example, the unit square $\{(x_1, x_2) : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$ is such a compact set. The square consists of the open set $\{(x_1, x_2) : 0 < x_1 < 1, 0 < x_2 < 1\}$, together with the four boundary curves, the edges (which are straight lines in this case), and the four points where the boundary curves come together. In checking for an actual maximum/minimum value, there are three types of points we have to consider:

1. maximum/minimum values in the interior, which give critical points of f and where we can apply the second derivative test;

2. maximum/minimum values in the interior of a boundary curve, which give critical points of the one variable function which is the restriction of f to that boundary curve, and where we can apply the usual second derivative test of one variable calculus;

3. endpoints of the boundary curves, where we must evaluate the function directly and compare it with all of the other potential maximum or minimum values which we found in (1) and (2).

In general, issues such as (2) often lead to the problem of maximizing or minimizing a function on a level set of another function. For example, if A is a linear map from $\mathbb{R}^n$ to itself, the real-valued function $\|A(\mathbf{x})\|^2$ has no maximum value. But it is more enlightening to look at the values $\|A(\mathbf{x})\|^2$

for $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$, i.e. just at the values of $\|A(\mathbf{x})\|^2$ on the unit sphere S^{n-1}, which is a level set. Since S^{n-1} is compact and $\|A(\mathbf{x})\|^2$ is continuous, it will have a maximum and a minimum value on S^{n-1}, and we would like to describe these.

One computational method for dealing with the maximum and minimum values functions restricted to level sets is the method of *Lagrange multipliers*. This problem is sometimes called *constrained optimization* since the values $\mathbf{x}$ are constrained to lie in a level set. We will discuss this method in the next chapter.

Chapter 9

Inverse and implicit function theorems

9.1 Coordinates and k-manifolds

In this chapter, we want to discuss two related questions:

1. What do we mean by coordinates and coordinate changes on $\mathbb{R}^n$ or on open subsets of $\mathbb{R}^n$?

2. What are the (nonlinear) geometric objects we would like to study in $\mathbb{R}^n$ (e.g. curves or surfaces) and what do we mean by coordinates on them? Also, how might such objects arise naturally in studying solutions of nonlinear equations?

We begin with a general discussion of the linear case. We have discussed linear changes of coordinates in detail: if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathbb{R}^n$, then every vector $\mathbf{x}$ in $\mathbb{R}^n$ can be uniquely written as $\sum_i t_i \mathbf{v}_i$ for $t_i \in \mathbb{R}^n$, and we think of the t_i as the coordinates of $\mathbf{x}$ in the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$. One way to think of this is that the linear function $\mathbb{R}^n \to \mathbb{R}^n$ defined by $G(t_1, \ldots, t_n) = \sum_i t_i \mathbf{v}_i$ is a bijection. Let F be its inverse. Thus F is a vector valued function on $\mathbb{R}^n$, and its component functions, which we call t_i, are n real-valued functions on $\mathbb{R}^n$. The meaning of $F(\mathbf{x})$ is that it tells us how to write $\mathbf{x}$ as a linear combination of the $\mathbf{v}_i$. We also know how to recognize the change of coordinate maps F and G: they are linear maps with inverses. More generally, if G corresponds to a matrix A, then the condition that G is invertible is: $\det A \neq 0$. From this point of view, if $G = F$ is the identity map, then the x_i are just the coordinate functions on $\mathbb{R}^n$, so that

we can view them as either the entries of a fixed vector $\mathbf{x}$ or as defining linear functions from $\mathbb{R}^n$ to $\mathbb{R}$.

The geometric objects of study in linear algebra are then the vector (or affine) subspaces of $\mathbb{R}^n$. If $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are linearly independent in $\mathbb{R}^n$, then we can parametrize $V = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ via the linear function $H(t_1, \ldots, t_d) = \sum_{i=1}^{d} t_i \mathbf{v}_i$. It is a bijection from $\mathbb{R}^d$ to V with inverse H^{-1}. We can also describe V as a level set: there exists a surjective linear map $L\colon \mathbb{R}^n \to \mathbb{R}^{n-d}$ whose kernel, i.e. the level set $L^{-1}(\mathbf{0})$, is exactly V. Other level sets would give affine subspaces of the form $V + \mathbf{p}$.

Note that some coordinates might look better than others. For example an orthonormal basis has many nice properties, one of which is that we can measure distance in the x_i coordinates or the t_i coordinates and get the same answer (F is an isometry). However, even in the case of a vector subspace V of $\mathbb{R}^n$, V will in general have no best basis, even if we want to choose an orthonormal basis. We might also want to enlarge the possible F's we use in a mild way, by allowing translation through a point $\mathbf{p}$ as well, so that we consider linear-affine maps F from $\mathbb{R}^n$ to itself, with an inverse.

For many purposes, we shall also want to understand nonlinear coordinates on $\mathbb{R}^n$. What this should mean is clear from the above discussion: we seek a map $G\colon \mathbb{R}^n \to \mathbb{R}^n$, say $G(t_1, \ldots, t_n) = (x_1, \ldots, x_n)$, so that every $\mathbf{x}$ can be uniquely described by the t_i's. Thus given $(t_1, \ldots, t_n)$ we describe a unique point of $\mathbb{R}^n$. Often we will phrase this differently: we want to find the t_i's given $\mathbf{x}$, so that we want the inverse function $F = G^{-1}$. By the definition of an inverse function, if $F(\mathbf{x}) = \mathbf{t} = (t_1, \ldots, t_n)$, then $G(t_1, \ldots, t_n) = \mathbf{x}$, so that F tells us how to write $\mathbf{x}$ in terms of the t_i-coordinates. In particular, the t_i appear in this point of view as the component functions of F, so that the t_i are functions of $x_1, \ldots, x_n$ (and vice-versa). We shall also require that F and G be continuous, since it seems reasonable not to allow changes of coordinates to tear apart point in $\mathbb{R}^n$ that are close together. (It is not usually enough to require just that F be continuous.) We call such a function F from $\mathbb{R}^n$ to $\mathbb{R}^n$ a *homeomorphism*. If both F and G are differentiable, we call F a *diffeomorphism*. The minimal requirement we will make is that F and G are C^1, but we will feel free to impose the condition that they have as many derivatives as needed whenever appropriate. More generally, we have:

Definition 9.1. Let $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ be two subsets. Then a *homeomorphism* F from X to Y is a function $F\colon X \to Y$ such that F is a bijection, and both F and F^{-1} are continuous. In this case we say that X and Y are *homeomorphic*.

A *diffeomorphism* F from an open subset U of $\mathbb{R}^n$ to an open subset V

of $\mathbb{R}^m$ is a bijection $: U \to V$ such that F and $F-1$ are both C^1. If $\mathbf{a} \in U$ and $\mathbf{b} = F(\mathbf{a}) \in V$, then the chain rule implies that $DF_{\mathbf{a}} \colon \mathbb{R}^n \to \mathbb{R}^m$ and $D(F^{-1})_{\mathbf{b}} \colon \mathbb{R}^m \to \mathbb{R}^n$ are linear maps which are inverses of each other, and so $n = m$. Diffeomorphisms for more general subsets of $\mathbb{R}^n$ will be discussed later.

Remark 9.2. It is natural to ask whether there can exist a homeomorphism from an open set $U \subseteq \mathbb{R}^n$ to an open set $V \subseteq \mathbb{R}^m$, $m \neq n$. In fact the answer to this question is no, and this fact is referred to as "invariance of domain." However, the proof of this fact is much more difficult than the proof in the case of a diffeomorphism which we outlined above.

From this point of view, one of the main problems of topology is to decide when two subsets of $\mathbb{R}^n$, $\mathbb{R}^m$ are homeomorphic and in general to classify all possible sets up to homeomorphism.

Back to the problem of describing different coordinates on $\mathbb{R}^n$, we will only be concerned with the case of diffeomorphisms. One reason is a practical one: in real life, most meaningful operations are indeed differentiable, and so it is reasonable to concentrate on these. Another reason is the following: given a general continuous function $F \colon \mathbb{R}^n \to \mathbb{R}^n$, we want to decide when it is a change of coordinates (homeomorphism), at least in some sense, and there is no practical way of deciding this question for general continuous functions F. However, as we will see, we will be able to give a satisfactory partial answer to this question in case F is a C^1 map.

Before we continue to discuss the general problem, let us describe the most important coordinate changes that actually arise in examples. The first is *polar coordinates*:

$$\begin{aligned} x &= r\cos\theta; \\ y &= r\sin\theta. \end{aligned}$$

Here we should think of polar coordinates as a function $P \colon \mathbb{R}^2 \to \mathbb{R}^2$, where the coordinates on the first $\mathbb{R}^2$ are called r and θ and $P(r,\theta) = (r\cos\theta, r\sin\theta)$. We are also interested in trying to find an inverse map, which would tell us how to write r and θ in terms of x and y. We let $r = x^2 + y^2$, so that $r \geq 0$ always. Thus r is the distance from (x,y) to the origin. As for θ, we could set $\theta = \tan^{-1}(y/x)$ or $\theta = \cos^{-1}(x/r)$ or $\theta = \sin^{-1}(y/r)$. The first definition runs into trouble if $x = 0$, whereas the other two are meaningful if $r \neq 0$, in other words except at the origin. Note however that we run into continuity or consistency problems somewhere, which we

could see from the formulas since the inverse trigonometric functions are not everywhere defined. In terms of the (x, y)-plane, we want θ (traditionally) to be the angle made by the line through the origin and (x, y) with the positive x-axis, and require that $0 \leq \theta < 2\pi$. In particular as we approach the positive x-axis from below, the limiting value of θ is $2\pi \neq 0$, so that θ is not a continuous function of x and y at the positive x-axis. There is nothing special here about the positive x-axis: we could define θ so as to be continuous there, by letting θ become negative (but small in absolute value) as we go below the positive x-axis, but then we will run into consistency problems somewhere else, where as we approach points in two different directions the values of θ we get will differ by $\pm 2\pi$. Notice also that we will never be able to define θ at the origin, since there is absolutely no reasonable definition for the angle at that point. Thus P is surjective but not injective. We can restrict P to the infinite half open rectangle

$$\{(r, \theta) : r \geq 0, 0 \leq \theta < 2\pi\}.$$

Here P is **nearly** injective (the only problem is in the half-open line segment $r = 0$, $0 \leq \theta < 2\pi$) but its inverse is not continuous along the positive x-axis. If we throw away the point $(0, 0)$, P is a bijection to $\mathbb{R}^2 - \{(0, 0)\}$, but its inverse is not continuous. If we throw away the entire positive x-axis, then P defines a homeomorphism from $(0, \infty) \times (0, 2\pi)$ to $\mathbb{R}^2 - \{(x, y) : x \geq 0\}$. A more natural procedure would be to define θ as taking values in the unit circle $S^1 = \{(x, y) : x^2 + y^2 = 1\}$. We could do this by defining

$$\theta(x, y) = \frac{1}{r}(x, y), (x, y) \neq (0, 0),$$

or by noting that $\theta \in [0, 2\pi) \mapsto (\cos\theta, \sin\theta) \in S^1$ defines a continuous bijection map $[0, 2\pi) \to S^1$ which is not a homeomorphism but has the effect of gluing 2π to 0, and so deals with the problem that θ is not continuous along the positive x-axis but jumps by 2π. Note that as a function of (x, y), the map $(x, y) \mapsto \frac{1}{r}(x, y)$ is the same as the map $(x, y) \mapsto (\cos\theta, \sin\theta)$. This tells us that P defines a homeomorphism from $(0, \infty) \times S^1 \subset \mathbb{R}^3$ to $\mathbb{R}^2 - \{(0, 0)\}$. Explicitly, the map is given by (here $r > 0$ and (a, b) is a point of the unit circle) $(r, (a, b)) \mapsto (ra, rb)$ and the inverse map is given by $(x, y) \mapsto (r, (x/r, y/r))$, where as usual $r = \sqrt{x^2 + y^2}$. You should stare at the picture of polar coordinates until the idea that $\mathbb{R}^2 - \{(0, 0)\}$ is really the same as an open infinite cylinder becomes plausible. Notice however that the measurement of length is very different in the two pictures.

Another way of dealing with the fact that P doesn't have a well-defined inverse (is not injective) is to say that *locally* P has an inverse, at least away from $\{r = 0\}$. In other words, for all $(r, \theta) \in \mathbb{R}^2$ with $r \neq 0$, there exists an open set $U \subset \mathbb{R}^2$ with $(r, \theta) \in U$ such that P is a homeomorphism from U to $P(U) = V \subset \mathbb{R}^2$. We say that P is a *local homeomorphism*. This amounts to saying that we can define θ in a small enough disk around any point (x, y) which is not the origin, in fact in any open set not containing the origin provided that we do not go all the way around the origin. In this case P^{-1} is defined and continuous, and in fact from our explicit formulas P^{-1} is differentiable. Since P and P^{-1} are differentiable, a homework problem based on the chain rule says that $DP_{(r,\theta)}$ is an invertible matrix at every point (r, θ), as long as $r \neq 0$, or equivalently that $\det DP_{(r,\theta)} \neq 0$. We can see this directly:

$$DP_{(r,\theta)} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

and so $\det DP_{(r,\theta)} = r(\cos^2\theta + \sin^2\theta) = r \neq 0$.

Before we leave this example, note that the lines $r = c$ and $\theta = c$ (c a constant) have interesting geometry too. The subset of $\mathbb{R}^2$ defined by $r = c$ is a circle of radius r, and the equation $r = c$ translates into the equation $F(x, y) = 0$, where $F(x, y) = \sqrt{x^2 + y^2} - c$. If we wanted, we could replace this equation by the simpler equation $x^2 + y^2 - c^2$, or even solve explicitly for y as a function of x, except at $x = \pm c, y = 0$: $y = \pm\sqrt{c^2 - x^2}$ where the choice of the square root is determined by whether $y > 0$ or $y < 0$. Thus in a small disk around (x, y), the set $r = c$ is the graph of a function. At the exceptional points $y = 0, x = \pm c$, we can instead solve for x as a function of y. Note that there is a natural choice of a parameter on the set $r = c$, namely θ. The parameter θ has a lot of geometric meaning in this example, via its interpretation as an angle and also because it is arc length up to the factor c. On the other hand, given θ and the parametrization $x = c\cos\theta, y = c\sin\theta$, we can solve for y as a function of x as follows: write $\theta = \cos^{-1}(x/c)$ and set $y = c\sin\left(\cos^{-1}(x/c)\right)$. A simple picture of right triangles shows that $\sin\left(\cos^{-1}(x/c)\right) = \sqrt{1 - (x/c)^2}$, as long as $y > 0$, and so $y = c\sqrt{1 - (x/c)^2} = \sqrt{c^2 - x^2}$. Note that there are problems when $x = \pm c$, in taking $\cos^{-1}(\pm 1)$, and also we need to be careful about the signs.

A second example of coordinates, which is very similar to polar coordinates, is *cylindrical coordinates*. These are coordinates (r, θ, z) on $\mathbb{R}^3$, defined by $x = r\cos\theta$, $y = r\sin\theta$, and $z = z$. So we just take polar coordinates in the plane and add the extra coordinate z (the height above the (x, y)-plane). We must make the same restrictions on r and θ to find inverse func-

tions, and always run into trouble if $r = 0$. If $C(r, \theta, z) = (r\cos\theta, r\sin\theta, z)$, then the derivative of C is given by

$$\begin{pmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and the determinant is again r. The sets where one of the coordinates is held constant are also interesting: for example, $r = c$ is a cylinder, infinite at both ends, in $\mathbb{R}^3$, and it is homeomorphic to $\mathbb{R} \times S^1$. This set is parametrized by the remaining two variables θ, z. You should be able to describe the other sets $\theta = c$, $z = c$, as well as the sets where *two* of the variables are held constant. In this case the sets in question are curves and are parametrized by the remaining variable.

A more interesting variation is given by *spherical coordinates*. In this case, we parametrize (x, y, z) by variables ρ, θ, ϕ. Here ρ is the distance from (x, y, z) to the origin, θ is the angle made by the vector which is the projection of (x, y, z) to the (x, y)-plane (in other words $(x, y, 0)$) with the positive x-axis, and ϕ is the angle between the positive z-axis and (x, y, z). As a function, we have $S\colon \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$\begin{aligned} x &= \rho\cos\theta\sin\phi; \\ y &= \rho\sin\theta\sin\phi; \\ z &= \rho\cos\phi. \end{aligned}$$

The natural restrictions on ρ, θ, ϕ are: $\rho \geq 0$, $0 \leq \theta < 2\pi$, $0 \leq \phi < \pi$. We leave it as an exercise to calculate the derivative $DS_{(\rho,\theta,\phi)}$ of S and to verify that $\det DS_{(\rho,\theta,\phi)} = -\rho^2\sin\phi$. In particular $DS_{(\rho,\theta,\phi)}$ has an inverse except when $\rho = 0$ or $\phi = 0, \pi$.

One can also use spherical coordinates to parametrize interesting geometric objects, in this case spheres. The set $\{\rho = c\}$ defines a sphere of radius c. Taking for example $c = 1$, the unit sphere is parametrized by coordinates θ, ϕ (except when $\phi = 0$). On this fixed sphere, the curves $\theta = \text{const}$ correspond to longitude lines from the north to the south pole, whereas the curves $\phi = \text{const}$ correspond to latitude circles. Note that $\phi = 0$ or $\phi = \pi$ are not actually curves, but are just the north and south poles respectively. We think of the function

$$F(\theta, \phi) = (\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi)$$

as a function from $[0, 2\pi] \times [0, \pi]$ to S^2, which parametrizes S^2 aside frrom some mild problems when $\theta = 0, 2\pi$ or $\phi = 0, \pi$. To visualize the topological effect of F, note that F glues the two line segments $\{0\} \times [0, \pi]$ and

$\{2\pi\} \times [0, \pi]$ together by identifying $(0, t)$ with $(2\pi, t)$, and collapses the line segments $[0, 2\pi] \times \{0\}$ and $[0, 2\pi] \times \{\pi\}$ to points (the north and south poles).

We could more generally describe all of the subsets of $\mathbb{R}^3$ where any one or two of the coordinates ρ, θ, ϕ are held fixed. For example, as in the case of cylindrical coordinates, the set $\{\theta = c\}$ is an infinite cylinder. On the other hand, the set $\{\phi = c\}$ is a cone whose vertex is the origin and such that the z-axis is the axis of symmetry of the cone (except for $\phi = \pi/2$, where we get the xy-plane).

It is easy to generalize this discussion inductively to the analogue of spherical coordinates in $\mathbb{R}^n$. We just write out the answer:

$$\begin{aligned}
x_1 &= r \cos\phi_1 \sin\phi_2 \sin\phi_3 \cdots \sin\phi_{n-1};\\
x_2 &= r \sin\phi_1 \sin\phi_2 \sin\phi_3 \cdots \sin\phi_{n-1};\\
x_3 &= r \cos\phi_2 \sin\phi_3 \sin\phi_4 \cdots \sin\phi_{n-1};\\
&\vdots\\
x_{n-1} &= r \cos\phi_{n-2} \sin\phi_{n-1};\\
x_n &= r \cos\phi_{n-1}.
\end{aligned}$$

Here as usual $r^2 = \|\mathbf{x}\|^2 = x_1^2 + \cdots + x_n^2$, $0 \le \phi_1 \le 2\pi$ and $0 \le \phi_k \le \pi$ for $k \ge 2$.

Having said a few words about coordinates, we give a description of the geometric objects in $\mathbb{R}^n$ we want to study.

Definition 9.3. The subset $X \subseteq \mathbb{R}^n$ is *locally homeomorphic to* $\mathbb{R}^k$ if for every point $\mathbf{x} \in X$, there exists a ball B in $\mathbb{R}^n$ centered at $\mathbf{x}$ such that $X \cap B$ is homeomorphic to an open set in $\mathbb{R}^k$. A closed subset of $\mathbb{R}^n$ locally homeomorphic to $\mathbb{R}^k$ will be called a *k-manifold* in $\mathbb{R}^n$. Sometimes we shall just be interested in closed subsets X of an open subset U of $\mathbb{R}^n$; by definition this means that X is the intersection of U with a closed subset of $\mathbb{R}^n$. In this case, if locally homeomorphic to $\mathbb{R}^k$, then we will call X a k-manifold in U.

In practice, we will also be interested in maps with derivatives. For U an open subset of $\mathbb{R}^n$, we say that a closed subset X of U is a *smooth k-manifold* in U if for every point $\mathbf{x} \in X$, there exists a ball B in U centered at $\mathbf{x}$ and a diffeomorphism $F\colon U \cap B \to V$, where V is an open subset of $\mathbb{R}^n$, such that $F(X \cap B)$ is equal to the intersection $\mathbb{R}^k \cap V$, where $\mathbb{R}^k$ is viewed as a subset of $\mathbb{R}^n$ by setting the last $n - k$ coordinates equal to zero. In other words, there is a diffeomorphism (= nonlinear change of coordinates) which "straightens out" the set X.

For example, a point is a 0-manifold, a circle is a 1-manifold, and the torus and the sphere are two (different) 2-manifolds. An open subset of $\mathbb{R}^n$ is an n-manifold. Given $k \neq \ell$, it follows from invariance of domain that a k-manifold cannot simultaneously be an ℓ-manifold. Hence, if X is a k-manifold, the integer k is well-defined and is called the *dimension* of X. In the differentiable case, this follows from what we know about diffeomeorphisms.

On the other hand, the following are **not** manifolds: the union of two intersecting lines in $\mathbb{R}^2$, or a full "two-sided" cone $z^2 = x^2 + y^2$ in $\mathbb{R}^3$.

One standard way to produce k-manifolds or smooth k-manifolds is via graphs of continuous or C^1 functions, which we shall also call *graph manifolds*. Given a continuous function $F\colon \mathbb{R}^k \to \mathbb{R}^j$, we have the graph of F, which is a subset of $\mathbb{R}^{k+j} = \mathbb{R}^n$, say, where we let $n = k + j$: namely the set

$$X = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{k+j} : \mathbf{y} = F(\mathbf{x})\}.$$

Since F is continuous, X is closed. Moreover X is homeomorphic to $\mathbb{R}^k$. For example, define $G\colon \mathbb{R}^k \to X$ by $G(\mathbf{x}) = (\mathbf{x}, F(\mathbf{x}))$ with inverse map $H\colon X \to \mathbb{R}^k$ defined by $G(\mathbf{x}, \mathbf{y}) = \mathbf{x}$. Thus G associates to a point $\mathbf{x} \in \mathbb{R}^k$ the point on the graph lying "above" $\mathbf{x}$, and H associates to a point $(\mathbf{x}, \mathbf{y})$ on the graph the projection to the point $\mathbf{x}$ lying "below" it. Note that G is **not** the same as F. For example F could be constant but G never is. It is easy to see that G is a homeomorphism from R^n to X, i.e. that G and H are both continuous and are inverses of each other. The variables $x_1, \ldots, x_k$ are the natural parameters on X from this point of view. We could do the same with a function F only defined on an open subset U of $\mathbb{R}^k$, in which case the graph of F is a closed subset of $U \times \mathbb{R}^j \subseteq \mathbb{R}^n$ homeomorphic to U.

In the differentiable case, a graph manifold is a smooth k-manifold, although we have to work a little harder. Setting $n = k + j$ as above, so that $j = n - k$. a function $F\colon \mathbb{R}^k \to \mathbb{R}^{n-k}$ is an $(n-k)$-tuple of functions $(f_1, \ldots, f_{n-k})$ and the graph manifold X is given by

$$\begin{aligned} X &= \{(x_1, \ldots, x_k, f_1(x_1, \ldots, x_k), \ldots, f_{n-k}(x_1, \ldots, x_k)) : (x_1, \ldots, x_k) \in \mathbb{R}^k\} \\ &= \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_{k+1} = f_1(x_1, \ldots, x_k), \ldots, x_n = f_{n-k}(x_1, \ldots, x_k))\}. \end{aligned}$$

Define $T\colon \mathbb{R}^n \to \mathbb{R}^n$ by the formula

$$T(x_1, \ldots, x_n) = (x_1, \ldots, x_k, x_{k+1} - f_1(x_1, \ldots, x_k), \ldots, x_n - f_{n-k}(x_1, \ldots, x_k)).$$

Then it is easy to see that T is a C^1 function if the f_i are C^1, and that then T has a C^1 inverse given by

$$T^{-1}(x_1, \ldots, x_n) = (x_1, \ldots, x_k, x_{k+1} + f_1(x_1, \ldots, x_k), \ldots, x_n + f_{n-k}(x_1, \ldots, x_k)).$$

Clearly $T(X)$ is exactly the set $\mathbb{R}^k \times \{\mathbf{0}\} \subseteq \mathbb{R}^n$. Hence X is a smooth k-manifold.

Most of the time, an interesting k-manifold will not consist of a single open set homeomeorphic to an open subset of $\mathbb{R}^k$, but rather will be in several pieces. For example, consider the 2-sphere $S^2 = \{(x_1, x_,x_3) : x_1^2 + x_2^2 + x_3^2 = 1\}$. The open northern hemisphere where $x_3 > 0$ is the graph manifold for the C^∞ function $f(x_1, x_2) = \sqrt{1 - x_1^2 - x_2^2}$, and hence is homeomorphic and diffeomeorphic to the open unit disk in $\mathbb{R}^2$. A similar statement holds for the open southern hemisphere $x_3 < 0$, which is the graph manifold for the function $-\sqrt{1 - x_1^2 - x_2^2}$. Every point not on the equator is in either the northern or southern hemisphere. To account for the points on the equator, we could use the four hemispheres $x_1 > 0$. $x_1 < 0$, $x_2 > 0$, $x_2 < 0$. Each such hemisphere is the graph of some function $x_i = f(x_j, x_k)$; the only difference is that we have not singled out the last variable to a function of the others. For example, the hemisphere $x_1 > 0$ is the graph $x_1 = \sqrt{1 - x_2^2 - x_3^2}$. In this way we have expressed S^2 as a union of 6 graph manifolds, and so it is a 2-manifold, indeed a smooth 2-manifold. A similar construction shows that S^n is a union of $2n + 2$ smooth graph manifolds.

In fact, one can write S^2 (or S^n) more efficiently as a union of **two** pieces, each of which is homeomorphic (and in fact diffemorphic) to $\mathbb{R}^2$ (or $\mathbb{R}^n$ in general). This is by what is called *stereographic projection*: imagine the 2-sphere as resting on the plane P parallel to the xy-plane defined by $x_3 = -1$ (so that only the south pole $(0, 0, -1)$ touches P). A line through the north pole $(0, 0, 1)$ and another point $\mathbf{x} \in S^2$ will meet P in exactly one point, and this sets up a homeomorphism $F \colon S^2 - \{(0, 0, 1)\} \to \mathbb{R}^2$. In fact, it is not hard to give an explicit formula for F. In this way, we can view S^2 and $\mathbb{R}^2$ with an extra point, the north pole, added "at infinity." A similar statement works for any n. Doing the same construction with the roles of the north pole and the south pole reversed gives a homeomorphism from $S^2 - \{(0, 0, -1)\}$ to $\mathbb{R}^2$, so that we have covered all of S^2 (or more generally S^n) with just two pieces.

One of the main problems in topology is to decide when two k-manifolds are homeomorphic or diffeomorphic, and if possible to classify k-manifolds. Although this problem is actually unsolvable, a great deal is known about it. For example, it is easy to classify connected 1-manifolds. (It is reasonable to only consider connected manifolds, since anything which is not connected can be written in terms of connected pieces.) A compact connected 1-manifold is homeomorphic to the circle S^1, and a noncompact connected 1-manifold is homeomorphic to $\mathbb{R}$ (or equivalently to an open interval in $\mathbb{R}$).

For a 2-manifold the situation is already much more complicated, although one can classify compact connected 2-manifolds. For example, the sphere and the torus are not homeomorphic. In addition, there are many non-homeomorphic non-compact 2-manifolds. One can show that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^2 - \{(0,0)\}$, and in general the complement of n points in $\mathbb{R}^2$ is homeomorphic to the complement of m points if and only if $n = m$. One phenomenon which happens for 2-manifolds but not for 1-manifolds is that there exist "one-sided" 2-manifolds. For example, the Möbius strip is a noncompact one-sided 2-manifold. There are also compact examples, of which one of the most well-known examples is the Klein bottle. A 2-manifold with two sides is called *orientable*; we will discuss the meaning of this later. The classification of compact 2-manifolds has been known for a long time. For example, a compact orientable 2-manifold is the 2-sphere, a torus, or an analogue of the torus with more holes (such as a pretzel), and two such 2-manifolds are homeomorphic (and even diffeomorphic) $\iff$ they have the same number of holes.

Note that, if X is a k-manifold in $\mathbb{R}^n$, then by definition for every point $\mathbf{x} \in X$ there exists a homeomorphism F from an open subset of $\mathbb{R}^k$ to $X \cap B$, where B is an open ball containing $\mathbf{x}$. More generally, we will call a k-manifold X, together with a homeomorphism F from an open subset of $\mathbb{R}^k$ to X, a *parametrized manifold*, with the function F being the parametrization. Thus every manifold is built up from pieces which are parametrized manifolds. From now on we will just look at the differentiable case. In this case, there is a diffeomorphism F from an open subset U of $\mathbb{R}^n$ to B such that F induces a bijection from $U \cap (\mathbb{R}^k \times \{\mathbf{0}\})$ to $X \cap B$. If V is the open subset of $\mathbb{R}^k$ corresponding to $U \cap (\mathbb{R}^k \times \{\mathbf{0}\})$, then there is an induced differentiable function $V \to \mathbb{R}^n$ whose image is $X \cap B$, which we can think of as a set of coordinates on the piece $X \cap B$ of X. More generally, we will think of an injective C^1 function $H\colon V \subseteq \mathbb{R}^k \to \mathbb{R}^n$, whose image is X or the intersection of X with an open subset of X, as a *parametrization* of X or as giving coordinates on X. We can think of this as a higher dimensional analogue of a parametrized curve (the case $k = 1$). As we shall see, it will be reasonable to put some additional requirements on H. For example, it is natural to require that the derivative $DH_{\mathbf{x}}$ is injective for all $\mathbf{x} \in V$; this is the higher dimensional analogue of saying that the parametrized curve is required to have a nonzero tangent vector at all points. To tie this in with our earlier discussion, polar or spherical coordinates enable us to parametrize a circle or a sphere (except at the north or south pole). Frequently we need a choice of parametrization to work with the manifold, although there is usually no best choice. In general, as in the example of the sphere, we will

not be able to describe all of the manifold X with a single parametrization, and must instead think of assembling X from smaller pieces which we can imagine as glued or sewn together.

Let us try to summarize the questions about k-manifolds as follows, just looking at the smooth case. There are three ways we might try to find smooth k-manifolds:

1. Via paramtrizations: given an injective C^1 function $H\colon V \to \mathbb{R}^n$, where V is an open subset of $\mathbb{R}^k$, and with some assumptions on the derivatives of H such as: $DH_{\mathbf{x}}$ is injective for all $\mathbf{x} \in V$.

2. Via level sets: given a C^1 function $G\colon U \to \mathbb{R}^{n-k}$, where U is an open subset of $\mathbb{R}^n$, we seek conditions on G so that the level set $G^{-1}(\mathbf{c})$ is a smooth k-manifold. If $G = (g_1, \ldots, g_{n-k})$ is written in components, then
$$G^{-1}(\mathbf{c}) = \{\mathbf{x} \in U : g_1(\mathbf{x}) = c_1, \ldots, g_{n-k}(\mathbf{x}) = c_{n-k}\}$$
for some real numbers $c_1, \ldots, c_{n-k}$ (the components of $\mathbf{c}$). So from this point of view we are asking if the set of simultaneous solutions to the $n-k$ nonlinear equations $g_i(\mathbf{x}) = c_i$ has some nice geometric properties.

3. Via graph manifolds X, the graph of the C^1 function $G\colon V \to \mathbb{R}^n$, where V is an open subset of $\mathbb{R}^k$. Note that a graph manifold is **both** a parametrized manifold and a level set. As we shall see, under the right extra assumptions, conversely (at least in small open balls) parametrized manifolds and level sets can also be described as graph manifolds.

After this long preliminary discussion concerning coordinates and smooth k-manifolds, let us move on to state some results. We begin with one concerning coordinates:

Theorem 9.4 (Inverse Function Theorem)**.** *Let U be an open subset of $\mathbb{R}^n$ and let $F\colon U \to \mathbb{R}^n$ be a C^1 function, such that $DF_{\mathbf{a}}$ is invertible for some $\mathbf{a} \in U$. Then there exist open sets $U' \subseteq U$ and $V \subseteq \mathbb{R}^n$ such that $\mathbf{a} \in U'$, $F(U') = V$, and the restricted map $F\colon U' \to V$ is a bijection. If we let F^{-1} be the inverse map, then F^{-1} is C^1 also, and $D(F^{-1})_{\mathbf{y}} = (DF_{F^{-1}(\mathbf{y})})^{-1}$ for all $\mathbf{y} \in V$.*

Before we state the implicit function theorem in general, let us state it for a single function f of n variables:

Theorem 9.5 (Implicit Function Theorem for one function)**.** *Let U be an open subset of $\mathbb{R}^n$ and let $f\colon U \to \mathbb{R}$ be a C^1 function. Suppose that $\mathbf{a} = (a_1, \ldots, a_n) \in U$, with $f(\mathbf{a}) = 0$, and that $\dfrac{\partial f}{\partial x_n}(\mathbf{a}) \neq 0$. Then there exists an open set $U' \subseteq \mathbb{R}^{n-1}$ containing $(a_1, \ldots, a_{n-1})$, an open interval $I \subseteq \mathbb{R}$ containing a_n, with $U' \times I \subseteq U$, and a C^1 function $g\colon U' \to I$ with with $g(a_1, \ldots, a_{n-1}) = a_n$ such that:*

1. *For all $(x_1, \ldots, x_{n-1}, x_n) \in U' \times I$, $f(x_1, \ldots, x_{n-1}, x_n) = 0 \iff x_n = g(x_1, \ldots, x_{n-1})$. In other words, the level set $f^{-1}(0) \cap (U' \times I)$ is exactly the graph of g.*

2. *The partial derivatives of g can be found by implicit differentiation: for $i < n$, if we let $\mathbf{x}' = (x_1, \ldots, x_{n-1})$, then*

$$\begin{aligned} 0 &= \frac{\partial f(\mathbf{x}', g(\mathbf{x}'))}{\partial x_i} \\ &= \frac{\partial f}{\partial x_i}(\mathbf{x}', g(\mathbf{x}')) + \frac{\partial f}{\partial x_n}(\mathbf{x}', g(\mathbf{x}'))\frac{\partial g}{\partial x_i}(\mathbf{x}'), \end{aligned}$$

so that

$$\frac{\partial g}{\partial x_i}(\mathbf{x}') = -\frac{\partial f}{\partial x_i}(\mathbf{x}', g(\mathbf{x}'))\Big/\frac{\partial f}{\partial x_n}(\mathbf{x}', g(\mathbf{x}')).$$

Before we state the general implicit function theorem, let us make some comments on the above. (1) says that the solution to $f = 0$, at least near $(a_1, \ldots, a_n)$, is the graph of the function $x_n = g(x_1, \ldots, x_{n-1})$. Notice if we had originally taken f to be a "graph equation" $x_n - g(x_1, \ldots, x_{n-1})$, so that the zero set is clearly the graph, then $\partial f/\partial x_n = 1$ which is always nonzero, and in fact ∇f would then be the vector $(-\nabla g, 1)$. The theorem says that for a general f, provided $\partial f/\partial x_n \neq 0$, we can replace f by a (usually different) equation which has the same zero set but which comes from the graph of a function. (2) tells us how to find the derivative of g by the usual mechanism of implicit differentiation. Note that in the statement of the theorem, the last variable x_n is singled out. But we could make a more general statement by assuming that $\partial f(\mathbf{a})/\partial x_i \neq 0$ for some i, and then the theorem would say that we could write x_i as a function of the other variables. To say that $\partial f(\mathbf{a})/\partial x_i \neq 0$ for some i is to say that $\nabla f(\mathbf{a}) \neq \mathbf{0}$. Thus if $\nabla f(\mathbf{a}) \neq \mathbf{0}$, then the zero set $\{f = 0\}$ defines a smooth $(n-1)$-manifold in U. For example, for the $(n-1)$-sphere in $\mathbb{R}^n$, we take $f(\mathbf{x}) = \|\mathbf{x}\|^2 - 1$. Then $\nabla f = 2\mathbf{x}$, and this is never zero **on the zero set of** f. (Of course it is zero elsewhere, namely at the origin.) Thus $f = 0$ is an $(n-1)$-manifold. We can solve

for x_n as a function of $x_1, \ldots, x_{n-1}$ except when $x_n = 0$, and there we can solve for x_i as a function of the remaining variables except when $x_i = 0$. Of course, on the sphere the coordinates are never all simultaneously zero. We can also write x_n explicitly as a function of $x_1, \ldots, x_{n-1}$ in this case (but usually we will not be able to do so).

In general, we have defined a point $\mathbf{a}$ to be a regular point for f if $\nabla f(\mathbf{a}) \neq \mathbf{0}$ and a critical point otherwise. At a regular point, the level set containing $\mathbf{a}$ is then a smooth $(n-1)$-manifold. We say that a critical point $\mathbf{a}$ is a *singular point* of the level set. For example, the critical point $(0,0)$ of $f(x_1, x_2) = x_1 x_2$ lives on the level set $x_1 x_2 = 0$, and is the point where two lines intersect; the level set is clearly not a manifold at this point. The critical point $(0,0)$ of $f(x_1, x_2) = x_1^2 + x_2^2$ lives on a level set which is a single point, and so is not even one-dimensional. Of course, the same is true for any strict local maximum or minimum, in any dimension.

Let us see how to derive this preliminary version of the implicit function theorem from the inverse function theorem. The idea is to first show that $x_1, \ldots, f$ are a system of coordinates in an open set containing $\mathbf{a}$, which we can take to be of the form $U' \times I$. Let $F \colon U \to \mathbb{R}^n$ be defined by

$$F(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-1}, f(x_1, \ldots, x_n)).$$

Then $DF_{(a_1,\ldots,a_n)}$ corresponds to the matrix

$$\begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ \frac{\partial f}{\partial x_1}(\mathbf{a}) & \frac{\partial f}{\partial x_2}(\mathbf{a}) & \ldots & \frac{\partial f}{\partial x_{n-1}}(\mathbf{a}) & \frac{\partial f}{\partial x_n}(\mathbf{a}) \end{pmatrix}.$$

Then $\det F(\mathbf{a}) = \dfrac{\partial f}{\partial x_n}(\mathbf{a}) \neq 0$, so that F defines a system of coordinates on some open set $U' \times I$, i.e. if V is the image $F(U' \times I)$, then F defines a diffeomorphism $U' \times I \to V$. There is then the inverse map $F^{-1} \colon V \to U' \times I$. We can write $F^{-1} = (h_1, \ldots, h_n)$. The zero set of f is the same as

$$F^{-1}(\{(y_1, \ldots, y_n) \in V : y_n = 0\}.$$

Now write

$$(x_1, \ldots, x_n) = F^{-1}(y_1, \ldots, y_n) = (h_1(y_1, \ldots, y_n), \ldots, h_n(y_1, \ldots, y_n)),$$

where the $x_i = h_i(y_1, \ldots, y_n)$ are C^1 functions of the y_i's and use

$$(y_1, \ldots, y_n) = F(F^{-1}(y_1, \ldots, y_n)) = F(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-1}, f).$$

Thus says that $y_i = x_i$, $i < n$, and so

$$x_n = h_n(y_1, \ldots, y_{n-1}, y_n) = h_n(x_1, \ldots, x_{n-1}, y_n)$$

for some C^1 function h_n. Thus the set where $y_n = 0$ is exactly the set where $x_n = h_n(x_1, \ldots, x_{n-1}, 0)$. Set $g(x_1, \ldots, x_{n-1}) = h_n(x_1, \ldots, x_{n-1}, 0)$. Then the graph of g is the place where $f = 0$, at least in $U' \times I$. The statements on the differentiability of g and the formula for the derivative follow by implicitly differentiating as in the statement of (2) of the theorem. □

Now let us state the general case of the implicit function theorem:

Theorem 9.6 (Implicit Function Theorem, general case)**.** *Let $F\colon U \to \mathbb{R}^m$ be a C^1 function, where U is an open subset of $\mathbb{R}^n \times \mathbb{R}^m$. Write $F = (f_1, \ldots, f_m)$. Given $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m$, where $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_m)$, suppose that $F(\mathbf{a}, \mathbf{b}) = \mathbf{0}$ and that the $m \times m$-matrix*

$$\Big(\frac{\partial f_i}{\partial x_{n+j}}(\mathbf{a}, \mathbf{b})\Big), \qquad 1 \le i, j \le m,$$

has an inverse.

Then there are open sets $V \subseteq \mathbb{R}^n$ containing $\mathbf{a}$ and $W \subseteq \mathbb{R}^m$ containing $\mathbf{b}$ and a C^1 function $G\colon V \to W \subset \mathbb{R}^m$ such that $G(\mathbf{a}) = \mathbf{b}$ and with the following properties:

1. *For all $\mathbf{x} \in V$ and $\mathbf{y} \in W$, $F(\mathbf{x}, \mathbf{y}) = \mathbf{0} \iff \mathbf{y} = G(\mathbf{x})$. Thus the intersection of $F^{-1}(\mathbf{0})$ with the open set $V \times W$ about $(\mathbf{a}, \mathbf{b})$ is exactly the graph of G.*

2. *The derivative of G can be found by the formula*

$$DG_{\mathbf{x}} = -(D_2 F^{-1}_{(\mathbf{x}, G(\mathbf{x}))} \circ D_1 F_{(\mathbf{x}, G(\mathbf{x}))}.$$

Here $D_2 F_{(\mathbf{x}, G(\mathbf{x}))}$ stands for the $m \times m$ matrix

$$\Big(\frac{\partial f_i}{\partial x_{n+j}}(\mathbf{x}, G(\mathbf{x}))\Big), \qquad 1 \le i, j \le m,$$

and part of the statement is that this matrix has an inverse for $\mathbf{x} \in V$, and $D_1 F_{(\mathbf{x}, G(\mathbf{x}))}$ denotes the $m \times n$ matrix

$$\Big(\frac{\partial f_i}{\partial x_j}(\mathbf{x}, G(\mathbf{x}))\Big), \qquad 1 \le i \le m, 1 \le j \le n.$$

Let us see that the inverse function theorem implies the implicit function theorem. The proof is very similar to the special case of a single function and we shall be brief. Given $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ as in the statement of the implicit function theorem, consider the map $H\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ defined by

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

If we work out the derivative $DH_{(\mathbf{a},\mathbf{b})}$, we see that it corresponds to the $(n+m) \times (n+m)$ matrix

$$\begin{pmatrix} \mathrm{Id} & \left(\frac{\partial f_i}{\partial x_k}(\mathbf{a}, \mathbf{b})\right)_{\substack{1 \le i \le m \\ 1 \le k \le n}} \\ 0 & \left(\frac{\partial f_i}{\partial x_{n+j}}(\mathbf{a}, \mathbf{b})\right)_{\substack{1 \le i \le m \\ 1 \le j \le m}} \end{pmatrix}.$$

Now it is easy to see that a matrix of the form

$$\begin{pmatrix} \mathrm{Id} & A \\ 0 & T \end{pmatrix},$$

where Id is $n \times n$, T is $m \times m$, and A and 0 are $n \times m$ and $m \times n$ respectively, has an inverse exactly when T has an inverse (explicit solution). In fact, the inverse must be of the form

$$\begin{pmatrix} \mathrm{Id} & B \\ 0 & T^{-1} \end{pmatrix},$$

for some $n \times m$ matrix B. So we apply the implicit function theorem to find an inverse to H near $(\mathbf{a}, \mathbf{b})$. Using $H \circ H^{-1}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})$, it is easy to see that $H^{-1}(\mathbf{x}, \mathbf{y})$ is of the form $(\mathbf{x}, \widetilde{G}(\mathbf{x}, \mathbf{y}))$. Define a new function $G(\mathbf{x})$ by $H^{-1}(\mathbf{x}, \mathbf{0}) = (\mathbf{x}, G(\mathbf{x}))$, so that $\widetilde{G}(\mathbf{x}, \mathbf{0}) = G(\mathbf{x})$. Thus clearly (after shrinking all the open sets in question) $F(\mathbf{x}, \mathbf{y}) = \mathbf{0} \iff \mathbf{y} = G(\mathbf{x})$. The formula for the derivative of G follows similarly from the fact that the derivative of H^{-1} is the inverse of the derivative of H. The details are much the same as for a single function, but we will not write them all out.

We can restate the implicit function theorem in a way that does not single out the last m variables in the domain of F. With F as in the statement of the implicit function theorem, let $\mathbf{v}_i = (\partial f_1/\partial x_i(\mathbf{a}), \ldots, \partial f_m/\partial x_i(\mathbf{a}))$ be the i^{th} column of $DF_{\mathbf{a}}$. The hypothesis of the version of the implicit function theorem given above is the statement that $\mathbf{v}_{n+1}, \ldots, \mathbf{v}_{n+m}$ is a basis of $\mathbb{R}^m$, and in particular that $\mathbf{v}_1, \ldots, \mathbf{v}_{n+m}$ span $\mathbb{R}^m$. More generally, suppose that F is any function as in the statement of the implicit function theorem such that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{n+m}$ span $\mathbb{R}^m$, or equivalently such that the linear

map $DF_{\mathbf{a}}\colon \mathbb{R}^{n+m} \to \mathbb{R}^m$ is surjective. Then some subset of the $\mathbf{v}_i$ is a basis of $\mathbb{R}^m$, and after reordering the coordinates of $\mathbb{R}^{n+m}$ we could assume that it is the last m of the $\mathbf{v}_i$. Thus in general, provided that $DF_{\mathbf{a}}\colon \mathbb{R}^{n+m} \to \mathbb{R}^m$ is surjective, we can write the level set $F^{-1}(\mathbf{0})$ as a graph manifold, in the sense that m of the variables can be written as functions in the remaining n variables, and in particular $F^{-1}(\mathbf{0})$ is a manifold.

There is yet another version of the implicit function theorem for functions whose derivative is injective, as opposed to the case where the derivative is surjective. This version goes as follows: let U be an open subset of $\mathbb{R}^k$ and let $F\colon U \to \mathbb{R}^n$ be a function with continuous partials such that there exists a point $\mathbf{a} \in U$ with $DF_{\mathbf{a}}$ is injective. Thus necessarily $k \leq n$. For example, suppose that we have a "graph map" $F\colon U \to \mathbb{R}^n$ defined by $F(\mathbf{x}) = (\mathbf{x}, f_{k+1}(\mathbf{x}), \ldots, f_n(\mathbf{x}))$. Then DF looks like

$$\begin{pmatrix} \mathrm{Id} & & 0 \\ \partial f_{k+1}/\partial x_1 & \cdots & \partial f_{k+1}/\partial x_n \\ \vdots & \vdots & \vdots \\ \partial f_n/\partial x_1 & \cdots & \partial f_n/\partial x_n \end{pmatrix},$$

and it is 1-1. The implicit function in this version will say again that every function F can be rearranged into this form by choosing coordinates on $U \subseteq \mathbb{R}^k$ appropriately. We state it in a slightly more special form as follows:

Theorem 9.7. *Let U be an open subset of $\mathbb{R}^k$ and let $F\colon U \to \mathbb{R}^n$ be a function with continuous partials, let $\mathbf{a} \in U$, and let*

$$\mathbf{v}_i = \left(\frac{\partial f_1}{\partial x_i}(\mathbf{a}), \ldots, \frac{\partial f_n}{\partial x_i}(\mathbf{a})\right)$$

be the i^{th} column of $DF_{\mathbf{a}}$. Let $p\colon \mathbb{R}^n \to \mathbb{R}^k$ be the linear projection map defined by $p(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$. Suppose that the vectors $p(\mathbf{v}_1), \ldots, p(\mathbf{v}_k)$ are linearly independent and hence a basis of $\mathbb{R}^k$. (In particular $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent.) Then there exists an open set $U' \subseteq U$ with $\mathbf{a} \in U'$, an open set $V \subseteq \mathbb{R}^n$ with $p(f(\mathbf{a})) \in V$ and C^1 functions $g_{k+1}, \ldots, g_n$ defined in V such that $F(U')$ is equal to the graph of the function $G = (g_{k+1}, \ldots, g_n)$.

Proof. With $p\colon \mathbb{R}^n \to \mathbb{R}^k$ the linear map $p(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$, our hypothesis is that $p \circ DF_{\mathbf{a}}\colon \mathbb{R}^k \to \mathbb{R}^k$ is an isomorphism. It follows by the chain rule (or easy explicit calculation) that $D(p \circ F)_{\mathbf{a}} = p \circ DF_{\mathbf{a}}$ and so $p \circ F$ has an inverse near $\mathbf{a}$—call the inverse map H. In other words, there exists an open set $U' \subseteq U$ with $\mathbf{a} \in U'$, an open set $V \subseteq \mathbb{R}^k$ with $p(F(\mathbf{a})) \in V$

and a C^1 function $H\colon V \to U'$ such that, for all $\mathbf{y} = (y_1, \ldots, y_k) \in V$ and $\mathbf{x} \in U'$, $\mathbf{y} = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))$ if and only if $\mathbf{x} = H(\mathbf{y})$. Thus, for $\mathbf{x} \in U'$, $(f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}), f_{k+1}(\mathbf{x}), \ldots, f_n(\mathbf{x})) = (y_1, \ldots, y_k, g_{k+1}(\mathbf{y}), \ldots, g_n(\mathbf{y}))$, where $g_j(\mathbf{y}) = f_j \circ H(\mathbf{y})$. Thus, near $\mathbf{a}$, the image of F is described as the graph of the function $G\colon V \to \mathbb{R}^{n-k}$, where $G(\mathbf{y}) = (f_{k+1} \circ H(\mathbf{y}), \ldots, f_n \circ H(\mathbf{y}))$. □

Remark 9.8. 1) If we do not shrink the open set U, it is possible that the image of F might come back and cross itself later. Even if we rule out this possibility by requiring that F is injective, and furthermore a homeomorphism onto its image, we cannot assume that that the image of F is everywhere the graph of a function; it is essential to shrink the set U.

2) In general, with $F = (f_1, \ldots, f_n)$ and $\mathbf{v}_i = (\partial f_1/\partial x_i(\mathbf{a}), \ldots, \partial f_n/\partial x_i(\mathbf{a}))$, the assumption that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent means that we can put the $\mathbf{v}_i$ in row echelon form so that the $\mathbf{v}_i$ are a linear combination of

$$(1, 0, \ldots, 0, *, \ldots, *), \ldots, (0, 0, \ldots, 1, *, \ldots, *).$$

Of course, this is the typical case. But in some exceptional cases we might not be able to get the standard row echelon form, but would still be able to say that **some** of the variables y_i in the image of F could be written as a function of the others. Thus the above theorem is really a statement about the image of a C^1 map F whose derivative is injective at some point $\mathbf{a}$: it says that, after possibly shrinking the domain of F, the image of F is a graph manifold, in the sense that it is the graph of a function but possibly after reordering the variables.

Chapter 10

Integration

10.1 Functions of one variable

We begin with a discussion of integration in one variable. Let $f\colon [a,b] \to \mathbb{R}$ be a bounded function. We will mostly be interested in the case where f is continuous, in which case it is bounded by the extreme value theorem. The basic idea is to approximate the integral of f over $[a,b]$ by a finite sum. To do so, define a *partition* P of $[a,b]$ to be an increasing sequence $t_0 = a < t_1 < \cdots < t_n = b$. We can think of P as just a finite subset $\{t_0, \ldots, t_n\}$ of $[a,b]$ containing a and b or geometrically as a way of partitioning $[a,b]$ into a union of subintervals $[t_{i-1}, t_i]$ which just meet at endpoints. The *trivial partition* P_0 then corresponds to the set $\{a,b\}$. The *size* of P is the maximum value of $t_i - t_{i-1}$. Usually we will be interested in partitions whose size is small. Given the bounded function f, let m_i be the greatest lower bound of f on $[t_{i-1}, t_i]$ and let M_i be the least upper bound of f on $[t_{i-1}, t_i]$. Thus, for all $x_i \in [t_{i-1}, t_i]$,

$$m_i \leq f(x_i) \leq M_i,$$

and if f is continuous, then m_i and M_i are actually values of f on $[t_{i-1}, t_i]$. Define the *lower sum*

$$L(f,P) = \sum_{i=1}^{n} m_i(t_i - t_{i-1})$$

and the *upper sum*

$$U(f,P) = \sum_{i=1}^{n} M_i(t_i - t_{i-1}).$$

For example, for the trivial partition P_0, $L(f, P_0) = m(b-a)$ and $U(f, P_0) = M(b-a)$, where m is the greatest lower bound of f on $[a, b]$ and M is the least upper bound of f on $[a, b]$.

For all choices of $x_i \in [t_{i-1}, t_i]$, we have

$$L(f, P) \leq \sum_{i=1}^{n} f(x_i)(t_i - t_{i-1}) \leq U(f, P).$$

In case $f(x) \geq 0$ for all $x \in [a, b]$, we can think of $L(f, P)$ as a lower bound on the area under the graph of f, and likewise $U(f, P)$ is an upper bound.

Definition 10.1. If P and P' are two partitions with $P = \{t_0, \ldots, t_n\}$ and $P' = \{t'_0, \ldots, t'_m\}$, we say that P' is a *refinement* of P if, for every i with $0 \leq i \leq n$, there exists a j such that $t_i = t'_j$, or equivalently if $P \subseteq P'$. In other words, we get P from P' by further subdividing the intervals $[t_{i-1}, t_i]$. Note that every two partitions P and Q have a common refinement P', in other words there exists a partition P' which is a refinement of both P and Q. For example, thinking of P and Q as finite subsets of $[a, b]$, one can just take $P \cup Q$.

For example, every partition is a refinement of the trivial partition.

Lemma 10.2. *If P' is a refinement of P, then*

$$L(f, P) \leq L(f, P') \leq U(f, P') \leq U(f, P).$$

Proof. Let $t_{i-1} < t_i$ be two consecutive elements of the partition P, and suppose that the elements of P' that lie in the interval $[t_{i-1}, t_i]$ are written as $s_1 = t_{i-1} < s_2 < \cdots < s_{n_i} = t_i$. Since $[s_{j-1}, s_j]$ is a subinterval of $[t_{i-1}, t_i]$, if m'_j is the greatest lower bound of f on $[s_{j-1}, s_j]$, then $m'_j \geq m_i$, where m_i is the greatest lower bound of f on $[t_{i-1}, t_i]$, and similarly for the upper bounds. Thus

$$\sum_{j=1}^{n_i} m'_j(s_j - s_{j-1}) \geq \sum_{j=1}^{n_i} m_i(s_j - s_{j-1}) = m_i \sum_{j=1}^{n_i} (s_j - s_{j-1}) = m_i(t_i - t_{i-1}).$$

Summing over i then shows that $L(f, P) \leq L(f, P')$, and the argument that $U(f, P') \leq U(f, P)$ is similar. □

Corollary 10.3. *If P and Q are any two partitions of $[a, b]$, then $L(f, P) \leq U(f, Q)$.*

Proof. Choosing a partition P' which is a refinement of P and Q, we see from the previous lemma that $L(f,P) \leq L(f,P') \leq U(f,P') \leq U(f,Q)$. $\square$

Since the numbers $L(f,P)$ are bounded above, independently of the choice of the partition P, we can define

$$L(f) = \text{lub}\{L(f,P) : P \text{ is a partition of } [a,b]\}.$$

Likewise set

$$U(f) = \text{glb}\{U(f,P) : P \text{ is a partition of } [a,b]\}.$$

Definition 10.4. We say that f is *integrable* if $L(f) = U(f)$. In this case we define the *integral* of f over $[a,b]$ as

$$\int_a^b f = L(f) = U(f).$$

We also use the notation $\displaystyle\int_a^b f(x)dx$, where the $(x)dx$ part of the notation is conceptually meaningless but computationally useful.

We sometimes call the number $L(f)$ the *lower integral* of f and $U(f)$ the *upper integral* of f.

Example 10.5. Let $f\colon [0,1] \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 0, & \text{if } x \in \mathbb{Q} \cap [0,1]; \\ 1, & \text{otherwise.} \end{cases}$$

In other words, $f(x) = 0$ for all **rational** $x \in [0,1]$ and $f(x) = 1$ for all **irrational** $x \in [0,1]$. Then it is easy to see that $L(f) = 0$ and $U(f) = 1$. In particular, f is not integrable.

The following lemma gives another description of what it means for f to be integrable:

Lemma 10.6. *The bounded function f on $[a,b]$ is integrable if and only if, for every real number $\epsilon > 0$, there exists a partition P such that*

$$U(f,P) - L(f,P) < \epsilon.$$

Proof. Suppose that f is integrable. Given $\epsilon > 0$, there exists a partition P such that $L(f) - L(f,P) < \epsilon/2$ and a partition Q such that $U(f,Q) - U(f) < \epsilon/2$. Choose a refinement P' of both P and Q. Since $L(f) = U(f)$, we see that

$$\begin{aligned} U(f,P') - L(f,P') &= U(f,P') - U(f) + L(f) - L(f,P) \\ &\leq U(f,Q) - U(f) + L(f) - L(f,P) < \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

Thus, for the partition P', we have satisfied the inequality in the lemma.

Conversely, suppose that, for every real number $\epsilon > 0$, there exists a partition P such that $U(f,P) - L(f,P) < \epsilon$. Since $L(f,P) \leq L(f) \leq U(f) \leq U(f,P)$, it follows that

$$0 \leq U(f) - L(f) < \epsilon$$

for every positive real number ϵ and so $U(f) = L(f)$ as claimed. □

Using this lemma, we can prove the following basic fact about continuous functions:

Theorem 10.7. *If f is continuous on $[a,b]$, it is integrable.*

Proof. Recall that a continuous function on a closed bounded interval is uniformly continuous (this was a homework problem). Thus, given $\epsilon > 0$, there is a $\delta > 0$ such that, if $|x - y| < \delta$, then $|f(x) - f(y)| < \epsilon/(b-a)$. (We have renamed ϵ here to make the proof come out neatly.) Let $P = \{t_0, \ldots, t_n\}$ be any partition with size less than δ (for example you can divide $[a,b]$ up into subintervals of equal length less than δ). If m_i is as defined above, then $m_i = f(x_i)$ for some $x_i \in [t_{i-1}, t_i]$, by the extreme value theorem applied to $[t_{i-1}, t_i]$. Likewise $M_i = f(y_i)$ for some $y_i \in [t_{i-1}, t_i]$. Then $|x_i - y_i| \leq t_i - t_{i-1} < \delta$, and so $M_i - m_i < \epsilon/(b-a)$. It follows that

$$U(f,P) - L(f,P) = \sum_{i=1}^{n} (M_i - m_i)(t_i - t_{i-1}) < \epsilon/(b-a) \sum_{i=1}^{n} (t_i - t_{i-1}).$$

Now $\sum_{i=1}^{n} (t_i - t_{i-1}) = b - a$, which you can see algebraically (it is the sum

$$t_1 - t_0 + t_2 - t_1 + \cdots + t_{n-1} - t_{n-2} + t_n - t_{n-1},$$

where all of the terms cancel off in pairs except the terms $t_n = b$ and $-t_0 = -a$), or geometrically (it is the sum of the lengths of all the subintervals of

$[a, b]$, which thus add up to the length of $[a, b]$, namely $b - a$). In any case, we see that

$$U(f, P) - L(f, P) < \frac{\epsilon}{b-a}(b-a) = \epsilon,$$

and by the lemma above this says that f is integrable. □

A function need not be continuous to be integrable. For example, as we shall see below, a function which is discontinuous at only finitely many points of $[a, b]$ but is bounded is integrable. We call such a function a (bounded) *piecewise continuous function*. Such functions include step functions, or functions which are defined by two different rules: for example $f(x) = g_1(x), a \leq x < t_0$ and $f(x) = g_2(x)$ for $t_0 < x \leq b$, where g_1, g_2 are continuous functions. The value of the integral is unchanged, however we define $f(t_0)$. However, unbounded functions, or functions on unbounded intervals, lead to the study of improper integrals and we shall not discuss them.

Lemma 10.8. *Let f be a bounded function on $[a, b]$. Let $c \in (a, b)$, and suppose that, for all $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq [a, b]$, the function f is integrable on $[a, c - \delta]$ and on $[c + \delta, b]$. Then f is integrable. A similar statement holds if $c = a$ or $c = b$.*

Proof. We may assume that f is not constant. Let m be the greatest lower bound of f on $[a, b]$ and let M be the least upper bound, so that $m < M$. Given $\epsilon > 0$, choose $\delta < \epsilon/6(M-m)$, and choose partitions P_1 of $[a, c-\delta]$ and P_2 of $[c+\delta, b]$ such that $U(f, P_1) - L(f, P_1) < \epsilon/3$ and $U(f, P_2) - L(f, P_2) < \epsilon/3$. Then for the partition P of $[a, b]$ given by $P_1 \cup P_2$ (which clearly contains $\{c - \delta, c + \delta\}$), we clearly have

$$U(f, P) - L(f, P) = U(f, P_1) - L(f, P_1) + U(f, P_2) - L(f, P_2) + (M_0 - m_0)(2\delta),$$

where m_0 is the greatest lower bound of f on $[c - \delta, c + \delta]$ and M_0 the least upper bound. Since $m \leq m_0 \leq M_0 \leq M$, we have

$$U(f, P) - L(f, P) < \epsilon/3 + \epsilon/3 + (M - m)2(\epsilon/6(M - m)) = \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon.$$

Thus f is integrable by Lemma 10.6. □

Corollary 10.9. *Suppose that f is bounded and piecewise continuous on $[a, b]$ (i.e. there exists a finite set $X \subseteq [a, b]$ such that f is continuous on $[a, b] - X$). Then f is integrable.*

Proof. This follows from the previous corollary and induction on the number of elements of X. □

The integral as defined above has the following standard properties:

Proposition 10.10. (i) *If f is a bounded integrable function on $[a,b]$, m is the greatest lower bound of f on $[a,b]$ and M is the least upper bound of f on $[a,b]$, then*

$$m(b-a) \leq \int_a^b f \leq M(b-a).$$

(ii) *If f and g are integrable on $[a,b]$, then so is $f+g$, and moreover*

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

(iii) *If f is integrable on $[a,b]$ and c is a constant, then cf is also integrable, and moreover*

$$\int_a^b (cf) = c\int_a^b f.$$

(iv) *If f is integrable and $f \geq 0$ on $[a,b]$ (in other words $f(x) \geq 0$ for all $x \in [a,b]$), then $\int_a^b f \geq 0$. More generally, if f_1 and f_2 are integrable and $f_1 \geq f_2$ on $[a,b]$ (in other words $f_1(x) \geq f_2(x)$ for all $x \in [a,b]$), then $\int_a^b f_1 \geq \int_a^b f_2$.*

(v) *If f is integrable on $[a,b]$, then so is $|f|$, and $\left|\int_a^b f\right| \leq \int_a^b |f|$.*

(vi) *Given real numbers $a < b < c$, if f is a bounded function on $[a,c]$, then f is integrable on $[a,c]$ if and only if it is integrable on both $[a,b]$ and $[b,c]$, and in this case*

$$\int_a^c f = \int_a^b f + \int_b^c f.$$

Proof. (i): Follows from $L(f,P_0) \leq \int_a^b f \leq U(f,P_0)$, where P_0 is the trivial partition. (ii): If $[t_{i-1},t_i] \subseteq [a,b]$, m_i is the greatest lower bound of f on

$[t_{i-1}, t_i]$, n_i is the greatest lower bound of g on $[t_{i-1}, t_i]$, M_i is the least upper bound of f on $[t_{i-1}, t_i]$, and N_i is the least upper bound of g on $[t_{i-1}, t_i]$, then it is easy to see that, for ℓ_i the greatest lower bound of $f+g$ on $[t_{i-1}, t_i]$ and L_i the least upper bound of $f+g$ on $[t_{i-1}, t_i]$, that we have

$$m_i + n_i \leq \ell_i \leq L_i \leq M_i + N_i.$$

It follows that, if P is a partition of $[a, b]$, then

$$L(f, P) + L(g, P) \leq L(f + g, P) \leq U(f + g, P) \leq U(f, P) + U(g, P).$$

In particular, given $\epsilon > 0$, if P is a partition such that $U(f, P) - L(f, P) < \epsilon/2$ and $U(g, P) - L(g, P) < \epsilon/2$, then

$$\begin{aligned} U(f + g, P) - L(f + g, P) &< U(f, P) + U(g, P) - (L(f, P) + L(g, P)) \\ &= (U(f, P) - L(f, P)) + (U(g, P) - L(g, P)) < \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

Thus $f+g$ is integrable. Since $\int_a^b (f+g) \leq U(f+g, P) \leq U(f, P)+U(g, P)$ for every partition P, we see that $\int_a^b f + \int_a^b g \leq \int_a^b (f + g)$. Arguing similarly with the lower integrals gives

$$\int_a^b (f + g) \leq \int_a^b f + \int_a^b g \leq \int_a^b (f + g),$$

so that we must have $\int_a^b (f+g) = \int_a^b f + \int_a^b g$, proving (ii). (iii): Immediate from the definition. (iv): The first statement (that $f \geq 0$ on $[a, b] \implies \int_a^b f \geq 0$) is immediate from (i), since clearly $m \geq 0$ in the notation of (i), and the second part is clear by applying the first part to the function $f_1 - f_2 \geq 0$ and using (i) and (ii). (iv): To see that $|f|$ is integrable, let P be a partition of $[a, b]$ and as usual let m_i be the greatest lower bound of f on $[t_{i-1}, t_i]$ and M_i the least upper bound of f on $[t_{i-1}, t_i]$. If n_i is the greatest lower bound of $|f|$ on $[t_{i-1}, t_i]$ and N_i the least upper bound of $|f|$ on $[t_{i-1}, t_i]$. then it is easy to see that $N_i \in \{|m_i|, |M_i|\}$ and that $n_i \in \{0, |m_i|, |M_i|\}$. Moreover, if $n_i = 0$, then $m_i \leq 0 \leq M_i$. It follows by checking the various cases that

$$N_i - n_i = |N_i - n_i| \leq M_i - m_i.$$

Thus, for every partition P of $[a, b]$,

$$U(|f|, P) - L(|f|, P) \leq U(f, P) - L(f, P).$$

In particular, given $\epsilon > 0$, we can choose a partition P such that $U(f, P) - L(f, P) < \epsilon$. Hence $U(|f|, P) - L(|f|, P) < \epsilon$ as well, so that $|f|$ is integrable. The inequality $\left|\int_a^b f\right| \leq \int_a^b |f|$ then follows from the triangle inequality. (vi): Follows easily since every partition P of $[a, c]$ defines partitions $[a, b]$ and $[b, c]$, by adding in the point b, and conversely, given two partitions P_1 of $[a, b]$ and P_2 of $[b, c]$, the union $P_1 \cup P_2$ defines a partition of $[a, c]$. □

Remark 10.11. (ii) above says that the integral is a linear map from the vector space of integrable functions on $[a, b]$ to $\mathbb{R}$.

Remark 10.12. Motivated by (vi) and by the fact that $\int_a^a f$ should be defined to be zero, we define $\int_b^a f$, in case $b > a$, to be $-\int_a^b f$. The integral is thus an *oriented* integral; it depends not just on the interval but also upon a choice of direction.

Finally we have the following, which is how we actually compute an integral:

Theorem 10.13 (Fundamental Theorem of Calculus)**.** *Let f be continuous on $[a, b]$ and define*

$$F(x) = \int_a^x f.$$

Then F is continuous on $[a, b]$ and differentiable on (a, b) and $F'(x) = f(x)$.

Proof. Since f is bounded on $[a, b]$, we suppose that $|f(x)| \leq C$ for all $x \in [a, b]$. Then, for all $x, y \in [a, b]$ with, say $x \leq y$, if $|x - y| < \epsilon/C$, then

$$|F(x) - F(y)| = \left|\int_x^y f\right| \leq \int_x^y |f| \leq C|y - x| < \epsilon.$$

Thus F is continuous. To see that F is differentiable on (a, b) and that $F'(x) = f(x)$, we compute: given $x \in (a, b)$ and $h > 0$, we have

$$\frac{F(x + h) - F(x)}{h} = \frac{1}{h} \int_x^{x+h} f,$$

and thus, if $m(h)$ is the minimum value of f on $[x, x+h]$ and $M(h)$ the maximum value, then

$$m(h) = \frac{1}{h}(m(h) \cdot h) \leq \frac{1}{h}\int_x^{x+h} f \leq \frac{1}{h}(M(h) \cdot h) = M(h).$$

Since f is continuous at x, for all $\epsilon > 0$, there exists a $\delta > 0$ such that, if $|t - x| < \delta$, then $|f(t) - f(x)| < \epsilon$. In particular, for such a δ, if $h < \delta$, then $|m(h) - f(x)| < \epsilon$ and $|M(h) - f(x)| < \epsilon$, since both $m(h)$ and $M(h)$ are values of f in $[x, x+h]$. Since $m(h) \leq f(x) \leq M(h)$, it follows in fact that

$$f(x) - \epsilon < m(h) \leq \frac{1}{h}\int_x^{x+h} f \leq M(h) < f(x) + \epsilon,$$

i.e. that, for all $\epsilon > 0$, there exists a $\delta > 0$ such that, if $h \geq 0$ and $h < \delta$, then

$$\left|\frac{F(x+h) - F(x)}{h} - f(x)\right| < \epsilon.$$

Thus $\lim_{h\to 0^+} \dfrac{F(x+h) - F(x)}{h} = f(x)$. The case $\lim_{h\to 0^-} \dfrac{F(x+h) - F(x)}{h}$, i.e. the case $h < 0$ is similar. Hence

$$\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = f(x).$$

□

10.2 Multiple integrals

Next we consider integrals in two variables; the generalization to n variables is similar. Given a rectangle $R = [a, b] \times [c, d]$, suppose that we have chosen partitions $P = \{t_0, \ldots, t_n\}$ for $[a, b]$ and $Q = \{s_0, \ldots, s_m\}$ for $[c, d]$. We then have the product partition $P \times Q = \{[t_{i-1}, t_i] \times [s_{j-1}, s_j] : 1 \leq i \leq n, 1 \leq j \leq m\}$. We refer to each rectangle $R_{ij} = [t_{i-1}, t_i] \times [s_{j-1}, s_j]$ as a *subrectangle of the partition* $P \times Q$. A refinement of the partition $P \times Q$ is given as $P' \times Q'$, where P' is a refinement of P and Q' is a refinement of Q. If f is bounded on R, we let m_{ij} be the greatest lower bound of f on $[t_{i-1}, t_i] \times [s_{j-1}, s_j]$ and let M_{ij} be the least upper bound there, and define

$$L(f, P \times Q) = \sum_{i,j} m_{ij}(t_i - t_{i-1})(s_j - s_{j-1}) = \sum_{i,j} m_{ij} \operatorname{area}(R_{ij});$$

$$U(f, P \times Q) = \sum_{i,j} M_{ij}(t_i - t_{i-1})(s_j - s_{j-1}) = \sum_{i,j} M_{ij} \operatorname{area}(R_{ij}),$$

noting that $(t_i - t_{i-1})(s_j - s_{j-1})$ is the area of the rectangle $R_{ij} = [t_{i-1}, t_i] \times [s_{j-1}, s_j]$. Thus in case $f \geq 0$ on R we can view $L(f, P \times Q)$ as a lower bound on the volume under the graph and $U(f, P \times Q)$ as an upper bound. If $P' \times Q'$ is a refinement of $P \times Q$, then

$$L(f, P \times Q) \leq L(f, P' \times Q') \leq U(f, P' \times Q') \leq L(f, P \times Q).$$

The numbers $L(f)$ and $U(f)$ are defined as in the one-variable case, and are called the lower and upper integrals of f respectively, and we say that f is *integrable* on R if $L(f) = U(f)$. Just as in the one-variable case, this is equivalent to: for all $\epsilon > 0$, there exists a partition $P \times Q$ such that $U(f, P \times Q) - L(f, P \times Q) < \epsilon$. In this case, we denote the integral by

$$\iint_R f \qquad \text{or} \qquad \iint_R f(x, y)dxdy.$$

An argument similar to the argument for functions of one variable, using the uniform continuity of f on the compact set R, shows:

Theorem 10.14. *If f is a continuous function on R, then f is integrable.* □

Integrals in n variables are defined similarly. We call a product $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ a *generalized rectangle* or simply a *rectangle* and define the quantity $\text{vol}(R) = (b_1 - a_1) \cdots (b_n - a_n)$ the *n-volume* or simply the *volume* of R. Of course, when $R = 1$ we call 1-volume length and when $n = 2$ we call 2-volume area. Note that volume is *translation invariant*: For every $\mathbf{p} \in \mathbb{R}^n$, if we define the translation

$$R + \mathbf{p} = \{\mathbf{x} + \mathbf{p} : \mathbf{x} \in R\},$$

then $R + \mathbf{p}$ is again a generalized rectangle and $\text{vol}(R + \mathbf{p}) = \text{vol}(R)$. Similarly, volume behaves under change of scale as follows: given $t \in \mathbb{R}$, $t \neq 0$, if we define

$$tR = \{t\mathbf{x} : \mathbf{x} \in R\},$$

then tR is again a generalized rectangle and $\text{vol}(tR) = |t|^n \text{vol}(R)$.

Upper and lower sums and integrability are defined as in the case $n = 2$, where we denote an integral of a function of n-variables over a generalized rectangle R by

$$\int \cdots \int_R f \qquad \text{or} \qquad \int \cdots \int_R f(x_1, \ldots, x_n)dx_1 \cdots dx_n,$$

or sometimes simply as $\int_R f$ when it is clear that we are dealing with a function of n variables. We can view the integral as a generalized kind of volume. For example, if $f(x_1, \ldots, x_n)$ is a continuous, nonnegative function of n variables over a generalized rectangle $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$, then

$$\{(x_1, \ldots, x_n, x_{n+1}) : (x_1, \ldots, x_n) \in R, 0 \leq x_{n+1} \leq f(x_1, \ldots, x_n)\}$$

is a region in $\mathbb{R}^{n+1}$ which has an $(n+1)$-volume given by the integral $\int_R f$.

For $n > 1$, there are many interesting subsets of $\mathbb{R}^n$ which are not rectangles. We will define a large class of such sets shortly. In general, if D is a bounded subset of $\mathbb{R}^n$, so that D is contained in some rectangle R, we will want to integrate over D, not over the whole rectangle R. Given a bounded function $f\colon R \to \mathbb{R}$, where $D \subseteq R$, we can make f nonzero only on D by multiplying f by the characteristic function χ_D of D ($\chi_D(\mathbf{x}) = 1$ if $\mathbf{x} \in D$ and $\chi_D(\mathbf{x}) = 0$ if $\mathbf{x} \notin D$). Note that, even if f is continuous, the function $\chi_D \cdot f$ will usually not be continuous. More generally, if f is just defined on D, there is a natural procedure for turning f into a function on a larger rectangle:

Definition 10.15. Let R be a rectangle in $\mathbb{R}^n$, let D be a subset of R and let $f\colon D \to \mathbb{R}$ be a function defined on D. Then we define the *the extension of f by 0* to be the function $\tilde{f}$ defined by:

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \text{if } \mathbf{x} \in D; \\ 0, & \text{if } \mathbf{x} \in R - D. \end{cases}$$

Note that $\tilde{f}$ is usually discontinuous, even if $f\colon D \to \mathbb{R}$ is continuous. Finally, if $\tilde{f}$ is integrable, then we define

$$\int_D f = \int_R \tilde{f}.$$

It is easy to see that this definition does not depend on the choice of rectangle containing D.

By analogy with the case of rectangles, if $f\colon D \to \mathbb{R}$ is a nonnegative function such that $\int_D f$ exists, we will interpret the integral as the volume of the subset of $\mathbb{R}^{n+1}$ defined by

$$(x_1, \ldots, x_{n+1}) : (x_1, \ldots, x_n) \in D, 0 \leq x_{n+1} \leq f(x_1, \ldots, x_n)\}.$$

Other physical applications of integration are as follows: suppose that $D \subseteq \mathbb{R}^3$ and that f represents some physical quantity associated to points of D. Typical examples might include:

1. $f(x, y, z)$ is the density of D at the point (x, y, z).
2. $f(x, y, z)$ is the temperature of D at the point (x, y, z).
3. $f(x, y, z)$ is the (electric) charge density of D at the point (x, y, z).

Then $\iiint_D f$ represents the total amount of the quantity involved: density in case (1), total temperature or temperature of D in case (2), total charge in case (3). Another example comes from probability. A *probability density function* $f\colon \mathbb{R}^3 \to \mathbb{R}$ is a function satisfying $f(x, y, z) \geq 0$ for all (x, y, z), and $\iiint_{\mathbb{R}^3} f = 1$. Here of course we must correctly interpret the improper integral. For example, f could be the probability density of an electron. In this case, the probability that the electron is in the region D is given by $\iiint_D f$.

Returning to general properties of the integral, we have the following analogue of Proposition 10.10:

Proposition 10.16. (i) *If f is a bounded integrable function on R, m is the greatest lower bound of f on R and M is the least upper bound of f on R, then*

$$m \operatorname{vol}(R) \leq \int_R f \leq M \operatorname{vol}(R).$$

(ii) *If f and g are integrable on R, then so is $f + g$, and moreover*

$$\int_R (f + g) = \int_R f + \int_R g.$$

(iii) *If f is integrable on R and c is a constant, then cf is also integrable, and moreover*

$$\int_R (cf) = c \int_R f.$$

(iv) *If f is integrable and $f \geq 0$ on R (in other words $f(x) \geq 0$ for all $x \in R$), then $\int_R f \geq 0$. More generally, if f_1 and f_2 are integrable and $f_1 \geq f_2$ on R (in other words $f_1(x) \geq f_2(x)$ for all $x \in R$), then $\int_R f_1 \geq \int_R f_2$.*

(v) *If f is integrable on R, then so is $|f|$, and* $\left|\int_R f\right| \leq \int_a^b |f|$.

We will discuss the analogue of (vi) shortly.

The main problem now is how to compute the integral. The basic result is:

Theorem 10.17 (Fubini's theorem)**.** *Suppose that $R = [a,b] \times [c,d]$ is a rectangle and that $f(x,y)$ is continuous on R. Then*

$$\iint_R f = \int_a^b \left\{ \int_c^d f(x,y)dy \right\} dx = \int_c^d \left\{ \int_a^b f(x,y)dx \right\} dy.$$

A similar statement holds for n variables. In practice we omit the braces when we write the integrals. The notation then tells us in which order we should do the integrals: we should work from the inner to the outer variable. In other words, $\int_a^b \int_c^d f(x,y)\,dydx$ means: first integrate f as a function of y, treating x as a constant, and substituting in the limits $y = c$ and $y = d$ to get a function of x alone, then integrate this function of x from a to b.

In case $f \geq 0$, Fubini's theorem may be viewed as a special case of Cavalieri's principle: it says that the volume under the graph of f is obtained by integrating the cross-sectional areas.

The simplest example is the case where $f(x,y) = g(x)h(y)$ is a product of two functions of the separate variables. In this case

$$\iint_R (g(x)h(y))dxdy = \left(\int_a^b g \right) \left(\int_c^d h \right).$$

For a function $f(x,y)$ which is not of the form $g(x)h(y)$, it may turn out that it is easy to calculate the double integral in one order, but much harder the other way.

Example 10.18.

$$\begin{aligned} \int_{-1}^1 \int_0^1 xye^{x^2y} dxdy &= \int_{-1}^1 \left\{ \frac{1}{2} e^{x^2y} \Big]_{x=0}^{x=1} \right\} dy \\ &= \frac{1}{2} \int_{-1}^1 (e^y - 1)dy \\ &= \frac{1}{2}(e - e^{-1} - 2). \end{aligned}$$

If we had tried to do this integral in the other order, we would have found (using integration by parts)

$$\int_0^1 \int_{-1}^1 xye^{x^2y}dxdy = \int_0^1 \left(\frac{ye^{x^2y}}{x} - \frac{e^{x^2y}}{x^3}\right)\Bigg]_{y=-1}^{y=1} dy$$
$$= \int_0^1 \left(\frac{e^{x^2}}{x} + \frac{e^{-x^2}}{x} - \frac{e^{x^2}}{x^3} + \frac{e^{-x^2}}{x^3}\right) dx.$$

Here, it is not even obvious that the improper integral exists! For a similar example, which can be handled by a judicious integration by parts, we have:

Example 10.19.

$$\int_1^2 \int_0^1 x\cos(xy)dxdy = \int_1^2 \left\{\left(x\frac{\sin(xy)}{y} + \frac{\cos(xy)}{y^2}\right)\Bigg]_0^1\right\} dy$$
$$= \int_1^2 \left(\frac{\sin y}{y} + \frac{\cos y}{y^2} - \frac{1}{y^2}\right) dy.$$

Here we have used integration by parts (details left as an exercise). The individual terms in the y-integrand above cannot be integrated in terms of elementary functions. However, if we try to integrate the first one, $\sin y/y$, by parts, we find that (with $u = 1/y$ and $dv = \sin y dy$):

$$\int \frac{\sin y}{y} dy = -\frac{\cos y}{y} - \int \frac{\cos y}{y^2} dy,$$

so that we can replace the term $\displaystyle\int \left(\frac{\sin y}{y} + \frac{\cos y}{y^2}\right) dy$ by $-\dfrac{\cos y}{y}$. Thus the integral becomes

$$\int_1^2 \left(\frac{\sin y}{y} + \frac{\cos y}{y^2} - \frac{1}{y^2}\right) dy = \int_1^2 \left(\frac{\sin y}{y} + \frac{\cos y}{y^2}\right) dy - \int_1^2 \frac{dy}{y^2}$$
$$= \left(-\frac{\cos y}{y} + \frac{1}{y}\right)\Bigg]_1^2 = -\frac{\cos 2}{2} + \frac{1}{2} + \cos 1 - 1.$$

If we try to do the integral by reversing the order, it is much easier! Here

$$\int_0^1 \int_1^2 x\cos(xy)dydx = \int_0^1 \sin(xy)]_1^2 dx$$
$$= \int_0^1 (\sin 2x - \sin x)dx = -\frac{\cos 2x}{2} + \cos x\Bigg]_0^1$$
$$= -\frac{\cos 2}{2} + \cos 1 + \frac{1}{2} - 1.$$

Of course, we get the same answer.

We will also want to be able to integrate functions over more general regions D. One basic kind of region in $\mathbb{R}^2$, which we shall call a *Type I region* is a region defined by inequalities:

$$D = \{(x, y) : a \leq x \leq b, f_1(x) \leq y \leq f_2(x)\}.$$

Here f_1 and f_2 are continuous functions on $[a, b]$ such that $f_1(x) \leq f_2(x)$ for all $x \in [a, b]$. One way to interpret the integral $\iint_D f(x, y)$ is to think of it as an integral of a *discontinuous* function $\tilde{f}$ over a rectangle $[a, b] \times [m, M]$, where m is a lower bound for f_1 on $[a, b]$ and M an upper bound for f_2 on $[a, b]$, and we define $\tilde{f}$ by to be the extension of f by zero:

$$\tilde{f}(x, y) = \begin{cases} f(x, y), & \text{if } f_1(x) \leq y \leq f_2(x); \\ 0, & \text{otherwise.} \end{cases}$$

Thus $\tilde{f}$ is almost always discontinuous along the curves $y = f_1(x)$, $y = f_2(x)$, but we will show that it is integrable on $[a, b] \times [m, M]$ as long as f is continuous and that the integral does not depend on the choice of the bounds m, M. In this case we can just define

$$\iint_D f = \iint_{[a,b]\times[m,M]} \tilde{f}.$$

Again, we need a way to compute the integral, and this is given by Fubini's theorem, suitably modified:

Theorem 10.20. *Let $D = \{(x, y) : a \leq x \leq b, f_1(x) \leq y \leq f_2(x)\}$ be a Type I region, where f_1 and f_2 are continuous on $[a, b]$. Suppose that R is some rectangle containing D and that f is continuous on R. Then $\chi_D f$ is integrable on R and*

$$\iint_D f = \iint_R \chi_D f = \int_a^b \left\{ \int_{f_1(x)}^{f_2(x)} f(x, y) dy \right\} dx.$$

More generally, if $f \colon D \to \mathbb{R}$ is continuous, and $\tilde{f} \colon D \to \mathbb{R}$ is the extension to R by zero, then $\tilde{f}$ is integrable on R and its integral is given by the formula above.

Of course, there are also *Type II regions* which are described by inequalities of the form

$$\begin{aligned} c \leq & y \leq d; \\ g_1(y) \leq & x \leq g_2(y), \end{aligned}$$

where g_1 and g_2 are continuous functions of y defined on $[c, d]$ and $g_1(y) \leq g_2(y)$ for all $y \in [c, d]$. It might happen that a region is both of Type I and of Type II, in which case we can again reverse the order of integration. But in general we are forced to do the integration as given, and in particular the last set of limits have to be constants, not functions, so that the final result is a number.

Example 10.21. Let T be the right triangle bounded by the lines $y = 0$, $x = 1$, and $y = 2x$, so that the three vertices are $(0, 0)$, $(1, 0)$, and $(1, 2)$. Then as a Type I region $T = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 2x$. The triangle T is also described as a Type II region (this is a special property of right triangles) via $T = \{(x, y) : 0 \leq y \leq 2, y/2 \leq x \leq 1\}$. Picking an integrand x^2y at random, we compute:

$$\begin{aligned} \iint_T x^2 y \, dxdy = \int_0^1 \int_0^{2x} x^2 y \, dydx = \int_0^1 \left. \frac{x^2 y^2}{2} \right|_{y=0}^{y=2x} dx = \frac{1}{2} \int_0^1 4x^4 \, dx \\ = \left. \frac{2}{5} x^5 \right|_{x=0}^{x=1} = \frac{2}{5}. \end{aligned}$$

In the other order, we get

$$\begin{aligned} \iint_T x^2 y \, dxdy = \int_0^2 \int_{y/2}^1 x^2 y \, dxdy = \int_0^2 \left. \frac{x^3 y}{3} \right|_{x=y/2}^{x=1} dy = \frac{1}{3} \int_0^1 \left(y - \frac{y^4}{8} \right) dy \\ = \frac{1}{3} \left(\frac{y^2}{2} - \frac{y^5}{40} \right) \Bigg|_{y=0}^{y=2} = \frac{1}{3} \left(2 - \frac{32}{40} \right) = \frac{1}{3} \left(2 - \frac{4}{5} \right) = \frac{1}{3} \cdot \frac{6}{5} = \frac{2}{5}. \end{aligned}$$

Note that, in this example, although the integrand x^2y is a product, we cannot write the integral as a product, because the region over which we integrate is not a rectangle.

Example 10.22. Consider the region D defined by $0 \leq x \leq 2$; $0 \leq y \leq x^2$. This Type I region can also be described by the Type II region $0 \leq y \leq 4$; $\sqrt{y} \leq x \leq 2$. Integrating the function x over D in two different ways, we get first

$$\int_0^2 \int_0^{x^2} x dydx = \int_0^2 xy \Big]_{y=0}^{y=x^2} dx = \int_0^2 x^3 dx = \frac{1}{4} x^4 \Big]_0^2 = 4.$$

Doing the integral the other way also gives

$$\int_0^4 \int_{\sqrt{y}}^2 x dx dy = \int_0^4 \frac{x^2}{2}\Big]_{x=\sqrt{y}}^{x=2} dy = \frac{1}{2}\int_0^4 (2-y)dy = \frac{1}{2}\left(4y - \frac{y^2}{2}\right)\Big]_0^4 = 4.$$

Example 10.23. Let D be the region in $\mathbb{R}^2$ enclosed by the line $y = x$ and the parabola $y = x^2$. Since the line intersects the parabola at the points $(0,0)$ and $(1,1)$, we can write D as the Type I region $\{(x,y) : 0 \leq x \leq 1, x^2 \leq y \leq x\}$ (note that, for $0 \leq x \leq 1$, the parabola always lies below the line). The region D is also a Type II region, given by $\{(x,y) : 0 \leq y \leq 1, y \leq x \leq \sqrt{y}\}$. Choosing a random integrand such as xy^2, we compute:

$$\begin{aligned}\iint_D xy^2\, dxdy &= \int_0^1 \int_{x^2}^x xy^2\, dydx = \int_0^1 \frac{xy^3}{3}\Big|_{y=x^2}^{y=x} dx = \int_0^1 \left(\frac{x^4}{3} - \frac{x^7}{3}\right) dx \\ &= \frac{1}{3}\left(\frac{x^5}{5} - \frac{x^8}{8}\right)\Big|_{x=0}^{x=1} = \frac{1}{3}\left(\frac{1}{5} - \frac{1}{8}\right) = \frac{1}{40}.\end{aligned}$$

Doing the integral in the other order gives

$$\begin{aligned}\iint_D xy^2\, dxdy &= \int_0^1 \int_y^{\sqrt{y}} xy^2\, dxdy = \int_0^1 \frac{x^2y^2}{3}\Big|_{x=y}^{x=\sqrt{y}} dy = \int_0^1 \left(\frac{y^3}{2} - \frac{y^4}{2}\right) dy \\ &= \frac{1}{2}\left(\frac{y^4}{4} - \frac{y^5}{5}\right)\Big|_{y=0}^{y=1} = \frac{1}{2}\left(\frac{1}{4} - \frac{1}{5}\right) = \frac{1}{40},\end{aligned}$$

so that as expected we get the same answer.

Similarly, for triple integrals, we can consider integrals over *standard regions* D in $\mathbb{R}^3$, i.e. subsets defined by inequalities of the form

$$\begin{aligned}a \leq& x \leq b;\\ h_1(x) \leq& y \leq h_2(x);\\ f_1(x,y) \leq& z \leq f_2(x,y),\end{aligned}$$

and similarly for any permutation of the variables, where h_1 and h_2 are two continuous functions on $[a,b]$ such that $h_1(x) \leq h_2(x)$ for all $x \in [a,b]$, and similarly for f_1, f_2. In this case, the triple integral is computed by taking the iterated integral

$$\int_a^b \int_{h_1(x)}^{h_2(x)} \int_{f_1(x,y)}^{f_2(x,y)} f(x,y,z) dz dy dx,$$

and similarly for other choices of the order of the variables (if this is possible).

As mentioned above, in case $z = f(x, y) \geq 0$, we can interpret $\iint_D f$ as the volume under the part of the graph $z = f(x, y)$ which lies over the region D. Similarly, given two continuous functions functions $h_1(x, y)$ and $h_2(x, y)$ on D with $h_1(x, y) \leq h_2(x, y)$ for all $(x, y) \in D$, the integral $\iint_D (h_2 - h_1)$ is the volume of the region between the two graphs lying over D. A special case of this is when $f = 1$, in which case $\iint_D 1$ is just the area of D. Note that, for example if D is a Type I region, then

$$\iint_D 1 = \int_a^b \int_{f_1(x)}^{f_2(x)} 1 dy dx = \int_a^b (f_2(x) - f_1(x)) dx,$$

which agrees with the usual first year calculus procedure. Likewise, the integral of the function 1 over a standard region D in $\mathbb{R}^3$ defined by the inequalities

$$D = \{(x, y, z) : a \leq x \leq b, h_1(x) \leq y \leq h_2(x), f_1(x, y) \leq z \leq f_2(x, y)\}$$

is the volume of D, and this gives the formula

$$\text{volume}(D) = \int_a^b \int_{h_1(x)}^{h_2(x)} (f_2(x, y) - f_1(x, y)) dy dx.$$

Example 10.24. Let us find the volume of a sphere of radius a (in other words, a ball of radius a in our terminology). The top and bottom of the sphere are described by the functions $z = \pm\sqrt{a^2 - x^2 - y^2}$ and the region in the xy-plane between these two graphs is the circle of radius a. Thus the volume in question is

$$\iint_{x^2+y^2 \leq a^2} 2\sqrt{a^2 - x^2 - y^2} dx dy = \int_{-a}^{a} \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} 2\sqrt{a^2 - x^2 - y^2} dy dx.$$

Now the inner integral $\int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} 2\sqrt{a^2 - x^2 - y^2} dy$ is the integral we would do to work out the area of a circle of radius $\sqrt{a^2 - x^2}$, and we know this area to be $\pi(\sqrt{a^2 - x^2})^2 = \pi(a^2 - x^2)$. Thus the total integral is

$$\int_{-a}^{a} \pi(a^2 - x^2) dx = \pi \left(a^2 x - \frac{x^3}{3} \right) \Big]_{-a}^{a} = \frac{4\pi a^3}{3}.$$

Note that the above argument reduced the problem of finding the volume of thge ball to integrating out the cross-sectional area (Cavalieri's principle). Of course, you can interpret Fubini's theorem as saying that the same is true for every volume integral.

Let us write down the analogues of Type I and II regions in all dimensions:

Definition 10.25. We define a *standard region* in $\mathbb{R}^n$ to be the closed set defined by the inequalities

$$\begin{aligned} a \leq & x_1 \leq b; \\ f_{11}(x_1) \leq & x_2 \leq f_{21}(x_1); \\ f_{12}(x_1, x_2) \leq & x_3 \leq f_{22}(x_1, x_2); \\ & \vdots \\ f_{1,n-1}(x_1, x_2, \ldots, x_{n-1}) \leq & x_n \leq f_{2,n-1}(x_1, x_2, \ldots, x_{n-1}), \end{aligned}$$

where the functions f_{1i} and f_{2i} are continuous functions on the standard region $D_i \subseteq R^i$ defined by only looking at the first i inequalities. We will also allow regions defined by permuting the variables (the analogues of Type II regions). Note that a standard region in $\mathbb{R}^n$ is successively built up from the standard regions $D_i \subseteq \mathbb{R}^i$ defined by the first i inequalities. An inductive argument shows that a standard region is closed and bounded and hence compact (by applying the extreme value theorem to the continuous function f_{i-1} defined on D_{i-1}, which is compact by induction).

For a standard region D given by inequalities as above, we define the *interior* $\operatorname{int}(D)$ to be the open set defined by the corresponding strict inequalities

$$\begin{aligned} a < & x_1 < b; \\ f_{11}(x_1) < & x_2 < f_{21}(x_1); \\ f_{12}(x_1, x_2) < & x_3 < f_{22}(x_1, x_2); \\ & \vdots \\ f_{1,n-1}(x_1, x_2, \ldots, x_{n-1}) < & x_n < f_{2,n-1}(x_1, x_2, \ldots, x_{n-1}), \end{aligned}$$

and the *boundary* ∂D to be the set $D - \operatorname{int}(D)$. Thus ∂D is a closed subset of D, hence compact, and it is defined by the inequalities defining D, where at least one inequality is actually an equality.

For example, given a Type I region in the plane defined by $D = \{(x, y) : a \leq x \leq b, f_1(x) \leq y \leq f_2(x)\}$, by definition ∂D consists of the two vertical line segments $\{a\} \times [f_1(a), f_2(a)]$ and $\{b\} \times [f_1(b), f_2(b)]$ as well as the two graphs $\{(x, y) : y = f_1(x), a \leq x \leq b\}$ and $\{(x, y) : y = f_2(x), a \leq x \leq b\}$.

Usually, ∂D will then consist of 4 curves, but as we have seen in examples there may be fewer if the graphs of f_1 and f_2 intersect at $x = a$ or $x = b$.

Given a standard region

$$D = \{(x, y, z) : a \leq x \leq b, h_1(x) \leq y \leq h_2(x), f_1(x, y) \leq z \leq f_2(x, y)\}$$

in $\mathbb{R}^3$, its boundary typically will consist of 6 pieces:

1. The two subsets of the two planes parallel to the yz-plane defined by $x = a, h_1(a) \leq y \leq h_2(a), f_1(a, y) \leq z \leq f_2(a, y)$ and $x = b, h_1(b) \leq y \leq h_2(b), f_1(b, y) \leq z \leq f_2(b, y)$. These are standard regions in the planes $x = a$ and $x = b$.

2. The two graphs of the two functions $z = f_1(x, y)$ and $z = f_2(x, y)$, for $(x, y) \in D_1$, where D_1 is the standard (Type I) region defined by $a \leq x \leq b, h_1(x) \leq y \leq h_2(x)$.

3. The two sets given by $a \leq x \leq b, y = h_1(x), f_1(x, h_1(x)) \leq z \leq f_2(x, h_1(x))$ and $a \leq x \leq b, y = h_2(x), f_1(x, h_2(x)) \leq z \leq f_2(x, h_2(x))$. If E is the standard region in the xz-plane defined by $a \leq x \leq b, f_1(x, h_1(x)) \leq z \leq f_2(x, h_1(x))$, then the set defined by $a \leq x \leq b, y = h_1(x), f_1(x, h_1(x)) \leq z \leq f_2(x, h_1(x))$ is the graph of the function $y = h_1(x)$, viewed as a function of x and y, over the standard region E, so it is of the same form as (2).

The boundary ∂D is hard to draw, even in $\mathbb{R}^3$. However, it is easy to see that, for a "general" standard region D in $\mathbb{R}^n$ and for every point $\mathbf{x} \in \partial D$, there is an open ball B containing $\mathbf{x}$ and a homeomorphism $F\colon B \to U$, where U is an open subset of $\mathbb{R}^n$, such that the image $F(\partial D)$ is equal to $\partial R \cap U$, where R is a rectangle. In other words, every point of ∂D looks topologically like a point in a cube of dimension $n-1$. In fact, if the f_{ij} are C^1 functions in an open set containing D_i, then we can assume that the F above is a diffeomorphism.

10.3 Proof of Fubini's theorem

We now want to prove theorems of Fubini type under fairly general circumstances.

Definition 10.26. Let D be a compact subset of $\mathbb{R}^n$. Then we say that D is *measurable* if the function χ_D is integrable (in some rectangle R containing D) and we define the *n-volume* or simply the *volume* $\mathrm{vol}(D)$ of D to be the

integral $\int_R \chi_D$. It is easy to see that these definitions are independent of the choice of rectangle R containing D. In case $n = 1$ we call $\text{vol}(D)$ the *length* of D and in case $n = 2$ we call $\text{vol}(D)$ the *area* of D.

Remark 10.27. 1) If D is contained in a rectangle R and $P_1 \times \cdots \times P_n$ is a partition of R, then $L(\chi_D, P_1 \times \cdots \times P_n)$ is the sum of the volumes of all of the rectangles in the partition which are contained in D and $U(\chi_D, P_1 \times \cdots \times P_n)$ is the sum of the volumes of all of the rectangles in the partition which have a nonempty intersection with D. In other words, $\text{vol}(D)$ is the least upper bound over all partitions P of the largest union of the rectangles in P which can be inscribed in D and it is also the greatest lower bound of the smallest union of rectangles in P which circumscribes D.

2) It is easy to see that volume of an arbitrary has the same properties as volume of a rectangle: for every translation $D + \mathbf{p}$, $\text{vol}(D + \mathbf{p}) = \text{vol}(D)$, and, for every real number t (including 0) $\text{vol}(tD) = |t|^n \text{vol}(D)$.

Definition 10.28. Let X be a subset of $\mathbb{R}^n$. Then X has *zero volume* if, for every $\epsilon > 0$, there exists a cover of X by a finite number of open rectangles $U_i = (a_{i1}, b_{i1}) \times \cdots \times (a_{in}, b_{in})$, $i = 1, \ldots, N$ (i.e. $X \subseteq \bigcup_{i=1}^N U_i$), such that

$$\sum_{i=1}^N \text{vol}(U_i) = \sum_{i=1}^N (b_{i1} - a_{i1}) \cdots (b_{in} - a_{in}) < \epsilon.$$

It is not hard to see that, if we had instead required the U_i to be **closed** rectangles, we would produce an equivalent definition (since every closed rectangle R is contained in the interior of a slightly larger rectangle cR, with $c = (1 + \epsilon)^{1/n}$, say). Finally, it is straightforward from the definitions that a compact subset X has zero volume $\iff$ it is measurable in the sense of the previous definition and its volume is 0.

Warning: We are using the terms measurable, integrable, and zero volume in a **very** naive sense. A careful treatment involves the theory of Lebesgue measure and integration.

An easy argument involving the definitions shows:

Lemma 10.29. *Let R be a rectangle and suppose that $X \subseteq R$ has zero volume. Let $f\colon R \to \mathbb{R}$ be a bounded function such that, if $f(\mathbf{x}) \neq 0$, then $\mathbf{x} \in X$, i.e. $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in R - X$. Then f is integrable and $\int_R f = 0$.* □

We can then state a (weak) analogue of Property (vi) of Proposition 10.10 as follows:

Proposition 10.30. *Let R be a rectangle in $\mathbb{R}^n$ and let D_1 and D_2 be two subsets of R such that $D_1 \cap D_2$ has zero volume. Let f be a bounded real-valued function defined on R (or more generally on $D_1 \cup D_2$). If $\chi_{D_1} f$ and $\chi_{D_2} f$ are both integrable, then $\chi_{D_1 \cup D_2} f$ is integrable, and in this case*

$$\int_{D_1 \cup D_2} f = \int_{D_1} f + \int_{D_2} f.$$

In particular, if D_1 and D_2 are measurable, then $D_1 \cup D_2$ is measurable, and in this case $\operatorname{vol}(D_1 \cup D_2) = \operatorname{vol}(D_1) + \operatorname{vol}(D_2)$.

Proof. Clearly $\chi_{D_1 \cup D_2} = \chi_{D_1} + \chi_{D_2} - \chi_{D_1 \cap D_2}$, and similarly

$$\chi_{D_1 \cup D_2} f = \chi_{D_1} f + \chi_{D_2} f - \chi_{D_1 \cap D_2} f.$$

By hypothesis, $\chi_{D_1} f$ and $\chi_{D_2} f$ are both integrable, and $\chi_{D_1 \cap D_2} f$ is integrable since it is nonzero only on the set $D_1 \cap D_2$ which has volume zero. Hence $\chi_{D_1 \cup D_2} f$ is integrable and

$$\begin{aligned} \int_{D_1 \cup D_2} f = \int_R \chi_{D_1 \cup D_2} f &= \int_R \chi_{D_1} f + \int_R \chi_{D_2} f - \int_R \chi_{D_1 \cap D_2} f \\ &= \int_{D_1} f + \int_{D_2} f - 0 = \int_{D_1} f + \int_{D_2} f. \end{aligned}$$

The final statement follows by taking f to be the constant function 1. □

Next we prove that a continuous function on a standard region is integrable. The first step is:

Theorem 10.31. *Let R be a rectangle in $\mathbb{R}^n$ and let X be a subset of R of zero volume. Suppose that $f\colon R \to \mathbb{R}$ is a bounded function and that f is continuous on $R - X$. Then f is integrable.*

Proof. Suppose that $|f(\mathbf{x})| \leq C$ for all $\mathbf{x} \in R$. Given $\epsilon > 0$, choose a partition $P = P_1 \times \cdots \times P_n$ of R such that the sum of all of the subrectangles in P which have a nonempty intersection with X has volume less than $\epsilon/4C$. Let U be the union of all of the open rectangles in P which meet X. Then U is open and so $R - U$ is a closed subset of R, hence compact. Thus, f is uniformly continuous on $R - U$. Hence, we can choose δ such that, if $\mathbf{x}, \mathbf{y} \in R - U$ and $\|\mathbf{x} - \mathbf{y}\| < \delta$, then $|f(\mathbf{x}) - f(\mathbf{y})| < \epsilon/2 \operatorname{vol}(R)$. Now, after

possibly refining the partition P, we obtain a new partition P' where we can assume that the diameter of every subrectangle in P' is less than δ. The difference $U(f,P)-L(f,P)$ is a sum of rectangles which either live in $R-U$ or are subrectangles of a rectangle of P meeting X. Summing over the first set of rectangles gives a contribution of at most $(\epsilon/2\,\mathrm{vol}(R))\cdot\mathrm{vol}(R)=\epsilon/2$, since the least upper bounds and greatest lower bounds of f on such a rectangle differ by at most $\epsilon/2\,\mathrm{vol}(R)$ and the sum of all the volumes of the subrectangles is at most the volume of R. Summing over the second set of rectangles gives a contribution less than $2C\cdot\epsilon/4C=\epsilon/2C$, since the least upper bound of f on such a subrectangle is at most C and the greatest lower bound of f there is at least $-C$. Hence $U(f,P)-L(f,P)<\epsilon$, so that f is integrable. □

To apply the theorem, we need examples of sets of zero volume. Let D be a standard region in $\mathbb{R}^n$, so that we have defined ∂D. Since D is compact, it is contained in some rectangle R. Clearly, $\mathbb{R}^n$ is a disjoint union of $\mathrm{int}(D)$, $\mathbb{R}^n-D$, and ∂D, where the first two sets are open. In particular, the characteristic function χ_D is continuous on $\mathrm{int}(D)$ and on $R-D$, and so fails to be continuous just at ∂D. We then have:

Theorem 10.32. *If D is a standard region in $\mathbb{R}^n$, then ∂D has zero volume. Hence D is measurable, and if $f\colon D\to\mathbb{R}$ is continuous, then f is integrable.*

Proof. (Sketch.) First consider the case where X is a bounded subset of an affine hyperplane $H\subseteq\mathbb{R}^n$ parallel to one of the coordinate hyperplanes, defined say by setting $x_i=a$. Then X is contained in a set of the form

$$\{(x_1,\ldots,x_n): x_i=a, |x_j|\le C \text{ if } j\ne i\}.$$

Given $\epsilon>0$, such a set is contained in the rectangle

$$[-C,C]\times\cdots\times[a-\epsilon/3(2C)^{n-1}, a+\epsilon/3(2C)^{n-1}]\times\cdots[-C,C],$$

which has volume

$$(2C)^{n-1}\cdot 2\epsilon/3(2C)^{n-1}=2\epsilon/3<\epsilon.$$

Hence X has zero volume.

Next consider the case where X is the graph of a continuous function $f\colon S\to\mathbb{R}$, where S is a **rectangle** in $\mathbb{R}^{n-1}$. Since f is uniformly continuous on S, given $\epsilon>0$, there exists a $\delta>0$ such that, for all $\mathbf{x},\mathbf{y}\in S$, if $\|\mathbf{x}-\mathbf{y}\|<\delta$, then $|f(\mathbf{x})-f(\mathbf{y})|<\epsilon/\mathrm{vol}(S)$. Choose a partition P of S

such that the diameter of each subrectangle $S_{i_1,\ldots,i_n}$ is less than δ, and let $R_{i_1,\ldots,i_n} = S_{i_1,\ldots,i_n} \times [m_{i_1,\ldots,i_n}, M_{i_1,\ldots,i_n}]$, where $m_{i_1,\ldots,i_n}$ is the minimum value of f in $S_{i_1,\ldots,i_n}$ and $M_{i_1,\ldots,i_n}$ is the maximum value of f there. Thus the graph of f over $S_{i_1,\ldots,i_n}$ is contained in the closed rectangle $R_{i_1,\ldots,i_n}$, and the sum of all the volumes of the closed rectangles $R_{i_1,\ldots,i_n}$ is

$$\sum_{i_1,\ldots,i_n} \operatorname{vol}(S_{i_1,\ldots,i_n})(M_{i_1,\ldots,i_n} - m_{i_1,\ldots,i_n}) < (\epsilon/\operatorname{vol}(S)) \sum_{i_1,\ldots,i_n} \operatorname{vol}(S_{i_1,\ldots,i_n})$$
$$= (\epsilon/\operatorname{vol}(S)) \operatorname{vol}(S) = \epsilon.$$

With a little more work, and induction, we can prove a similar result when X is the graph of a continuous function $f\colon D_{n-1} \to \mathbb{R}$, where D_{n-1} is a standard region in $\mathbb{R}^{n-1}$ The idea is then to show that the boundary in general consists of $2n$ pieces which are either standard regions in an $\mathbb{R}^{n-1}$ or graphs over standard regions, after permuting the variables (as in our discussion of the case $n = 3$) and we will not write out the details. □

Corollary 10.33. *If D is a standard region in $\mathbb{R}^n$ and f is continuous on D, then f is integrable, i.e. the extension of f by 0 to any rectangle containing D is integrable .* □

Remark 10.34. More generally, let D be any compact subset of R^n. We define the *interior* $\operatorname{int}(D)$ the be the set of all $\mathbf{x} \in D$ such that there exists a ball B centered at $\mathbf{x}$ with $B \subseteq D$. Clearly, $\operatorname{int}(D) \subseteq D$, $\operatorname{int}(D)$ is an open subset of $\mathbb{R}^n$, as is $\mathbb{R}^n - D$, and $D - \operatorname{int}(D) = \partial D$, the *boundary* of D, is compact. The arguments before the statement of Theorem 10.32 show ∂D has zero volume $\iff$ D is measurable.

We turn now to a very general statement of Fubini's theorem. For simplicity, we just write out the statement and proof for $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. The generalizations to $\mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$ are straightforward.

Theorem 10.35 (Fubini's theorem)**.** *Let $R = [a, b] \times [c, d]$ be a rectangle in $\mathbb{R}^2$ and let $f\colon R \to \mathbb{R}$ be a bounded integrable function. For each $x \in [a, b]$, let $g_x\colon [c, d] \to \mathbb{R}$ be the function $g_x(y) = f(x, y)$ and let $L(g_x)$ and $U(g_x)$ be the lower and upper integrals of g_x. Hence, if g_x is integrable, then $L(g_x) = U(g_x) = \int_c^d f(x, y)\, dy$. Then $L(g_x)$ and $U(g_x)$ are integrable functions of x and*

$$\int_a^b L(g_x) = \int_a^b U(g_x) = \iint_R f.$$

Proof. Let $P = P_1 \times P_2$ be a partition of R. Let $R_{ij} = [t_{i-1}, t_i] \times [s_{j-1}, s_j]$ be the $(i,j)^{\text{th}}$ rectangle in the partition and let m_{ij} and M_{ij} be the greatest lower bound and least upper bound for f on R_{ij}. Then

$$L(f, P_1 \times P_2) = \sum_{i,j} m_{ij}(t_i - t_{i-1})(s_j - s_{j-1}) = \sum_i \left(\sum_j m_{ij}(s_j - s_{j-1}) \right) (t_i - t_{i-1}).$$

Now clearly, for $x \in [t_{i-1}, t_i]$, m_{ij} is less than or equal the greatest lower bound n_j of $g_x(y) = f(x,y)$ on $[s_{j-1}, s_j]$, since

$$m_{ij} = \inf\{f(u,v) : (u,v) \in [t_{i-1}, t_i] \times [s_{j-1}, s_j]\} \leq \inf\{f(x,v) : v \in [s_{j-1}, s_j]\}$$

(note that by assumption $x \in [t_{i-1}, t_i]$). So

$$\sum_j m_{ij}(s_j - s_{j-1}) \leq \sum_j n_j(s_j - s_{j-1}) = L(g_x, P_2) \leq L(g_x),$$

for every $x \in [t_{i-1}, t_i]$. It follows that

$$\sum_j m_{ij}(s_j - s_{j-1}) \leq \inf\{L(g_x) : x \in [t_{i-1}, t_i]\}$$

and hence, by the definition of $L(L(g_x), P_1)$, that

$$L(f, P_1 \times P_2) = \sum_i \left(\sum_j m_{ij}(s_j - s_{j-1}) \right) (t_i - t_{i-1}) \leq L(L(g_x), P_1).$$

By general properties of upper and lower sums,

$$L(L(g_x), P_1) \leq U(L(g_x), P_1) \leq U(U(g_x), P_1),$$

and an argument symmetric to the one we gave for lower sums shows that

$$U(U(g_x), P_1) \leq U(f, P_1 \times P_2).$$

Combining these inequalities, we see that

$$L(f, P_1 \times P_2) \leq L(L(g_x), P_1) \leq U(L(g_x), P_1) \leq U(U(g_x), P_1) \leq U(f, P_1 \times P_2).$$

Under the assumption that f is integrable, for all $\epsilon > 0$, we can choose a partition $P_1 \times P_2$ such that $U(f, P_1 \times P_2) - L(f, P_1 \times P_2) < \epsilon$. It follows then that

$$U(L(g_x), P_1) - L(L(g_x), P_1) < \epsilon,$$

so that $L(g_x)$ is integrable as well, and since

$$L(f, P_1 \times P_2) \leq L(L(g_x), P_1) \leq \int_a^b L(g_x) \leq U(L(g_x), P_1) \leq U(f, P_1 \times P_2)$$

for every choice of partitions, we see that

$$\iint_R f \leq \int_a^b L(g_x) \leq \iint_R f.$$

Hence $\iint_R f = \int_a^b L(g_x)$. The case of $U(g_x)$ follows by similar arguments. □

Corollary 10.36. *Let D be a Type I region in $\mathbb{R}^2$, defined by $a \leq x \leq b$ and $f_1(x) \leq y \leq f_2(x)$, where f_1 and f_2 are continuous functions from $[a, b]$ to $\mathbb{R}$. Suppose that D is contained in a rectangle $R = [A, B] \times [C, D]$. Let $f\colon D \to \mathbb{R}$ be a continuous function, extended by 0 to give a function on R. Then f is integrable on D. Moreover,*

1. *For every $x \in [a, b]$, the function $g_x = f(x, y)$ is integrable as a function of $y \in [C, D]$ and the integral $\int_C^D g_x$ is equal to $\int_{f_1(x)}^{f_2(x)} f(x, y)\, dy$.*
2. *The function $\int_C^D g_x = \int_{f_1(x)}^{f_2(x)} f(x, y)\, dy$ is an integrable function of x which is zero unless $a \leq x \leq b$ and*

$$\iint_R f = \int_A^B \int_C^D g_x = \int_a^b \int_{f_1(x)}^{f_2(x)} f(x, y)\, dydx.$$

Proof. The function f is integrable by Corollary 10.33. Moreover, for each $x \in [a, b]$, the function $g_x(y)$ is equal to $f(x, y)$ if $f_1(x) \leq y \leq f_2(x)$, and $g_x(y) = 0$ if $C \leq y < f_1(x)$ or $f_2(x) < y \leq D$. Hence g_x is continuous except possibly at the points $f_1(x)$, $f_2(x)$. It follows that g_x is integrable and clearly $\int_C^D g_x = \int_{f_1(x)}^{f_2(x)} f(x, y)\, dy$. The rest of the corollary follows from Fubini's theorem above. □

As an application of Fubini's theorem we note:

Corollary 10.37. *Let $D_1 \subseteq \mathbb{R}^{n_1}$ and $D_2 \subseteq \mathbb{R}^{n_2}$ and suppose that D_1, D_2, and $D_1 \times D_2 \subseteq \mathbb{R}^{n_1+n_2}$ are all measurable. Then*

$$\operatorname{vol}(D_1 \times D_2) = \operatorname{vol}(D_1) \cdot \operatorname{vol}(D_2). \quad \square$$

10.4 Change of variables formula

Next we describe the very important change of variables formula for integrals. We begin with the general formula and then discuss it in $\mathbb{R}^2$ and $\mathbb{R}^3$ for the traditional coordinate changes.

Theorem 10.38. *Let R be a rectangle in $\mathbb{R}^n$, let F be a diffeomorphism from an open set from an open subset U containing R to an open subset of $\mathbb{R}^n$, and let f be a function defined on $\operatorname{Im} F$ such that $\int_{F(R)} f$ exists. Then*

$$\int_{F(R)} f = \int_R (f \circ F)|\det DF|.$$

Here DF is the derivative of F, which is an $n \times n$ matrix $\left(\frac{\partial y_i}{\partial x_j}\right)$, where the y_i's are the coordinates on the range $\mathbb{R}^n$ and the x_i's on the domain. We sometimes call DF the *Jacobian matrix* of F and abbreviate it by

$$DF = \frac{\partial(y_, \ldots, y_n)}{\partial(x_, \ldots, x_n)}.$$

The determinant of DF is then called the *Jacobian determinant* and is written $\det \frac{\partial(y_, \ldots, y_n)}{\partial(x_, \ldots, x_n)}$. With this notation, we can write the change of variables formula in the following form, which is slightly easier to remember:

$$\int_{F(R)} f(y_1, \ldots, y_n)\, dy_1 \cdots dy_n = \int_R f \circ F(x_, \ldots, x_n) \left|\det \frac{\partial(y_, \ldots, y_n)}{\partial(x_, \ldots, x_n)}\right| dx_1 \cdots dx_n.$$

We can then think of the $dx_1 \cdots dx_n|$ as canceling the "denominator" $\partial(x_, \ldots, x_n)$ to leave us with something like $dy_1 \cdots dy_n$. Of course, this formal manipulation does not actually mean anything.

Remark 10.39. 1) This formula looks very similar to the one variable formula

$$\int_a^b f(u)du = \int_r^s f(u(t))u'(t)dt,$$

where u is a C^1 function with $u(r) = a, u(s) = b$. For example, if u is a C^1 diffeomorphism which is increasing, then $u'(t) = |u'(t)|$ and $u(r) = a, u(s) = b$. On the other hand, if u is a C^1 diffeomorphism which is decreasing, then $s < r$ and $u'(t) = -|u'(t)|$. Thus

$$\int_a^b f(u)du = \int_r^s f(u(t))u'(t)dt = -\int_r^s f(u(t))|u'(t)|dt = \int_s^r f(u(t))|u'(t)|dt,$$

so we see that the fact that we have defined an oriented integral has compensated for the sign change in $|u'(t)|$. Further note that, in the one variable case, we don't need u to be injective; the derivative can change sign several times, corresponding to $u(t)$ going back and forth over the same interval. These different parts of the integral will appear with different signs and will in the end cancel each other out. However, this is not allowed for in our formula in higher dimensions. Later, we will discuss the meaning of the sign of $\det DF$ and discuss the analogue of oriented integrals in higher dimensions.

2) In practice, F is often not actually a diffeomorphism. We have seen this for polar and spherical coordinates, where F identifies $\theta = 0$ with $\theta = 2\pi$ and collapses the line segment $r = 0, 0 \leq \theta \leq 2\pi$ to a single point. However, these problems occur along sets of zero volume, and thus do not affect double integrals. A similar remark holds for spherical coordinates. We will use this without comment (and without writing down a formal proof) in what follows.

The most important example in the plane is polar coordinates. For example, if D is the rectangle $[0, a] \times [0, 2\pi]$ and P is the polar coordinates map $P(r, \theta) = (r\cos\theta, r\sin\theta)$, then $P(D)$ is the circle of radius a. Here

$$\det \frac{\partial(x, y)}{\partial(r, \theta)} = r$$

and so we can replace $dxdy$ in a double integral by $rdrd\theta$.

Example 10.40. The area of a circle of radius a is

$$\iint_{x^2+y^2\leq a^2} dxdy = \int_0^{2\pi} \int_0^a rdrd\theta = 2\pi \left.\frac{r^2}{2}\right]_0^a = \pi a^2.$$

Example 10.41. The volume of a sphere of radius a is given the integral

$$\begin{aligned}
\iint_{x^2+y^2\leq a^2} 2\sqrt{a^2 - x^2 - y^2}dxdy &= \int_0^{2\pi} \int_0^a 2\sqrt{a^2 - r^2}rdrd\theta \\
&= \int_0^{2\pi} \int_0^a (a^2 - r^2)^{1/2}(2rdr)d\theta \\
&= -\int_0^{2\pi} \left(\int_0^a (a^2 - r^2)^{1/2} d(a^2 - r^2) \right) d\theta \\
&= 2\pi \left. (-\frac{2}{3}(a^2 - r^2)^{3/2} \right]_0^a = \frac{4\pi a^3}{3}.
\end{aligned}$$

Example 10.42. An integral which appears often in probability theory is the improper integral $I = \int_{-\infty}^{\infty} e^{-x^2}\,dx = \lim_{R\to\infty}\int_{-R}^{R} e^{-x^2}\,dx$. The trick in computing the integral I is to compute its square:

$$\begin{aligned} I^2 &= \lim_{R\to\infty}\left(\int_{-R}^{R} e^{-x^2}\,dx \int_{-R}^{R} e^{-x^2}\,dx\right) \\ &= \lim_{R\to\infty}\left(\int_{-R}^{R} e^{-x^2}\,dx \int_{-R}^{R} e^{-y^2}\,dy\right) = \lim_{R\to\infty}\int_{-R}^{R}\int_{-R}^{R} e^{-x^2}e^{-y^2}\,dxdy \\ &= \lim_{R\to\infty}\int_{-R}^{R}\int_{-R}^{R} e^{-(x^2+y^2)}\,dxdy, \end{aligned}$$

where we have used the fact that the x is just a dummy variable and so could be replaced by another variable such as y. Let D_R be the square $[-R,R]\times[-R,R]$, so that $B_R \subseteq D_R \subseteq B_{\sqrt{2}R}$, where B_R is the disk of radius R centered at 0 and $B_{\sqrt{2}R}$ is the disk of radius $\sqrt{2}R$ centered at 0. Then

$$\iint_{B_R} e^{-(x^2+y^2)}\,dxdy \le \iint_{D_R} e^{-(x^2+y^2)}\,dxdy \le \iint_{B_{\sqrt{2}R}} e^{-(x^2+y^2)}\,dxdy,$$

and all integrals are positive. We can now compute the inner and outer integrals using polar coordinates:

$$\begin{aligned} \iint_{B_R} e^{-(x^2+y^2)}\,dxdy &= \int_0^{2\pi}\int_0^{R} e^{-r^2} r\,drd\theta = 2\pi\left(-\frac{1}{2}\right)e^{-r^2}\Big|_{r=0}^{r=R} \\ &= -\pi(e^{-R^2}-1) = \pi(1-e^{-R^2}). \end{aligned}$$

It follows that

$$\pi(1-e^{-R^2}) \le \iint_{D_R} e^{-(x^2+y^2)}\,dxdy \le \pi(1-e^{-2R^2}),$$

so that taking the limit as $R\to\infty$ gives $I^2=\pi$ and hence $I=\sqrt{\pi}$.

In $\mathbb{R}^3$, we can use either cylindrical coordinates or spherical coordinates. Cylindrical coordinates are not really anything new: the change of variables formula tells us that we can replace $dxdydz$ by $r\,drd\theta dz$. For spherical coordinates we have the calculation

$$\det\frac{\partial(x,y,z)}{\partial(\rho,\theta,\phi)} = -\rho^2\sin\phi.$$

Thus we can replace $dxdydz$ by $\rho^2 \sin\phi \, d\rho d\theta d\phi$. (Note that we are ignoring the problems with spherical coordinates at $\rho = 0$, $\phi = 0$ or π and the identification of $\theta = 0$ with $\theta = 2\pi$ in much the same way as for polar coordinates.)

Example 10.43. To find yet again the volume of a sphere of radius a, we compute:

$$\iiint_{x^2+y^2+z^2 \le a^2} 1 dxdydz = \int_0^{\pi} \int_0^{2\pi} \int_0^a \rho^2 \sin\phi \, d\rho d\theta d\phi =$$
$$= \frac{\rho^3}{3}\Big]_0^a (\theta)]_0^{2\pi} \, (-\cos\phi)]_0^{\pi} = \frac{a^3}{3}(2\pi)(1-(-1)) = \frac{4\pi a^3}{3},$$

as before.

Example 10.44. Consider the solid region D in $\mathbb{R}^3$ defined by

$$\{(x,y,z) : z \ge \sqrt{x^2+y^2} \text{ and } x^2+y^2+z^2 \le 1\}.$$

Geometrically, D consists of all of the points lying inside the top half of the cone $z^2 = x^2 + y^2$ obtained by revolving the line $z = x$ about the z-axis and also in the closed ball of radius one about the origin. Note that the sphere and the cone intersect where

$$z^2 = x^2+y^2, x^2+y^2+z^2 = 1 \text{ so that } 2z^2 = 1, z = \sqrt{2}/2 = x^2+y^2.$$

Pictorially, D looks like an ice cream cone. The volume of D can be computed via polar coordinates:

$$\begin{aligned}
\text{vol}(D) &= \iint_{x^2+y^2 \le \sqrt{2}/2} (\sqrt{1-x^2-y^2} - \sqrt{x^2+y^2}) \, dxdy \\
&= \int_0^{2\pi} \int_0^{\sqrt{2}/2} (\sqrt{1-r^2} - r) \, r \, dr d\theta \\
&= (2\pi)\left[\left(-\frac{1}{2}\right) \frac{2}{3}(1-r^2)^{3/2}\Big|_{r=0}^{r=\sqrt{2}/2} - \frac{r^3}{3}\Big|_{r=0}^{r=\sqrt{2}/2} \right] \\
&= \frac{2\pi}{3}\left(1 - \left(\frac{1}{2}\right)^{3/2} - \left(\frac{1}{2}\right)^{3/2}\right) = \frac{2\pi}{3}\left(1 - \frac{\sqrt{2}}{2}\right).
\end{aligned}$$

If we use spherical coordinates instead, it is easy to see that D is described via the inequalities $\rho \leq 1$ and $0 \leq \phi \leq \pi/4$. Thus

$$\begin{aligned}\mathrm{vol}(D) &= \iiint_D 1\, dxdydz = \int_0^{\pi/4}\int_0^{2\pi}\int_0^{\sqrt{2}/2} \rho^2 \sin\phi\, d\rho d\theta d\phi \\ &= \frac{\rho^3}{3}\Big|_0^{\sqrt{2}/2} \cdot \theta]_0^{2\pi} \cdot (-\cos\phi)]_0^{\pi/4} = \frac{1}{3}\cdot(2\pi)\cdot\left(-1+\frac{\sqrt{2}}{2}\right),\end{aligned}$$

which agrees with the previous computation.

Example 10.45. Using similar ideas, let us work out the volume of the closed n-ball $\overline{B}_r(\mathbf{0}) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq r\}$, for all n. Of course, we would get the same answer if we computed the volume of the open ball, since $\partial\overline{B}_r(\mathbf{0}) = S_r$, the $(n-1)$-sphere of radius r, which has measure zero. Note that, by change of scale,

$$\mathrm{vol}(\overline{B}_r(\mathbf{0})) = r^n \,\mathrm{vol}(\overline{B}_1(\mathbf{0})) = v_n r^n,$$

say, where v_n is the volume of the closed unit ball in $\mathbb{R}^n$. So we will compute v_n, noting that $v_1 = 2$ and $v_2 = \pi$. We have given the analogue of spherical and polar coordinates in the last chapter:

$$\begin{aligned} x_1 &= r\cos\phi_1 \sin\phi_2 \sin\phi_3 \cdots \sin\phi_{n-1}; \\ x_2 &= r\sin\phi_1 \sin\phi_2 \sin\phi_3 \cdots \sin\phi_{n-1}; \\ x_3 &= r\cos\phi_2 \sin\phi_3 \sin\phi_4 \cdots \sin\phi_{n-1}; \\ &\vdots \\ x_{n-1} &= r\cos\phi_{n-2}\sin\phi_{n-1}; \\ x_n &= r\cos\phi_{n-1}, \end{aligned}$$

where $r^2 = \|\mathbf{x}\|^2 = x_1^2 + \cdots + x_n^2$, $0 \leq \phi_1 \leq 2\pi$ and $0 \leq \phi_k \leq \pi$ for $k \geq 2$. In this parametrization, the unit ball corresponds to $0 \leq r \leq 1$, with the ranges for the coordinates ϕ_i as given. This formula has the inductive structure $x_i = g_i \sin\phi_{n-1}$ for $1 \leq i \leq n-1$, and $x_n = r\cos\phi_{n-1}$, where $g_1, \ldots, g_{n-1}$ are the corresponding coordinates on the unit ball in $\mathbb{R}^{n-1}$. By examining the formulas, it is clear that

$$\frac{\partial x_i}{\partial r} = \frac{x_i}{r}, 1 \leq i \leq n, \text{ and } \frac{\partial g_i}{\partial r} = \frac{g_i}{r}, 1 \leq i \leq n-1.$$

We now compute the Jacobian determinant $D_n = \det \dfrac{\partial(x_1,\ldots,x_n)}{\partial(r,\phi_1,\ldots,\phi_{n-1})}$:

Claim 10.46. $D_n = \pm r^{n-1} \sin\phi_2 \sin^2\phi_3 \cdots \sin^{n-2}\phi_{n-1}$.

Proof. The proof is by induction on n. For $n = 1$, $D_1 = 1$, and for $n = 2$, $D_2 = r$. We have also computed $D_3 = \rho^2 \sin\phi$ in the notation of spherical coordinates, which agrees with the answer above. For the inductive step, computing the determinant D_n, we find:

$$\begin{aligned}
\det \frac{\partial(x_1,\ldots,x_n)}{\partial(r,\phi_1,\ldots,\phi_{n-1})} &= \det \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \phi_1} & \cdots & \frac{\partial x_1}{\partial \phi_{n-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial r} & \frac{\partial x_n}{\partial \phi_1} & \cdots & \frac{\partial x_n}{\partial \phi_{n-1}} \end{pmatrix} \\
&= \det \begin{pmatrix} \frac{\partial g_1}{\partial r}\sin\phi_{n-1} & \frac{\partial g_1}{\partial \phi_1}\sin\phi_{n-1} & \ldots & g_1\cos\phi_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ \cos\phi_{n-1} & 0 & \ldots & -r\sin\phi_{n-1} \end{pmatrix} \\
&= \pm \det \begin{pmatrix} \frac{\partial g_1}{\partial \phi_1}\sin\phi_{n-1} & \ldots & \frac{\partial g_1}{\partial r}\sin\phi_{n-1} & g_1\cos\phi_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & \cos\phi_{n-1} & -r\sin\phi_{n-1} \end{pmatrix},
\end{aligned}$$

where we have switched the first and the $(n-1)^{\text{st}}$ columns. Comparing the second-to-last and last columns shows that first $n-1$ entries in the columns are equal up to a scalar multiple, since

$$\frac{\partial g_i}{\partial r}\sin\phi_{n-1} = \frac{g_i}{r}\sin\phi_{n-1} = \left(\frac{\sin\phi_{n-1}}{r\cos\phi_{n-1}}\right) g_i \cos\phi_{n-1},$$

and hence

$$g_i\cos\phi_{n-1} = \left(\frac{r\cos\phi_{n-1}}{\sin\phi_{n-1}}\right)\frac{\partial g_i}{\partial r}\sin\phi_{n-1}.$$

Expanding out about the last row gives

$$\begin{aligned}
\pm D_n &= r D_{n-1}\sin^n\phi_{n-1} + \frac{r\cos\phi_{n-1}}{\sin\phi_{n-1}} D_{n-1}\sin^{n-1}\phi_{n-1}\cos\phi_{n-1} \\
&= r D_{n-1}\sin^{n-2}\phi_{n-1}(\sin^2\phi_{n-1} + \cos^2\phi_{n-1}) = r D_{n-1}\sin^{n-2}\phi_{n-1},
\end{aligned}$$

and so we are done by induction. □

Now the change of variable formula says that

$$v_n = \mathrm{vol}(\overline{B}_1(\mathbf{0})) = \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \int_0^1 r^{n-1}\sin\phi_2\sin^2\phi_3\cdots\sin^{n-2}\phi_{n-1}\, dr d\phi_1 d\pi_2 \cdots d\phi_{n-1}.$$

Note that the n-fold integral is a product of ordinary integrals. Comparing this product with that for v_{n-1}, we see that the difference are: 1) instead of $\int_0^1 r^{n-2}\,dr = \dfrac{1}{n-1}$, we get $\int_0^1 r^{n-1}\,dr = \dfrac{1}{n}$, and 2) there is an extra term $I_n = \int_0^\pi \sin^{n-2}\phi_{n-1}\,d\phi_{n-1}$. Thus,

$$v_n = \left(\frac{n-1}{n}\right) I_n v_{n-1}.$$

So we must evaluate I_n. This is a standard one-variable integral. Integration by parts shows that $I_2 = \pi$, $I_3 = 2$, $I_4 = \pi/2$, $I_5 = 4/3$, and in general I_n satisfies a recursion

$$I_n = \left(\frac{n-3}{n-2}\right) I_{n-2}, n \geq 4.$$

Thus

$$I_{2n} I_{2n+1} = \frac{2\pi}{2n-1};$$
$$I_{2n+1} I_{2n+2} = \frac{\pi}{n}.$$

Then a straightforward induction shows:

$$v_{2n} = \frac{\pi^n}{n!};$$
$$v_{2n+1} = \frac{2^{n+1}\pi^n}{1 \cdot 3 \cdot 5 \cdots (2n+1)}.$$

10.5 Proof of the change of variable formula

We begin with the already very interesting case of an invertible linear map $\mathbb{R}^n \to \mathbb{R}^n$, which we view as given by an $n \times n$ invertible matrix A.

Theorem 10.47. *Let A be an $n \times n$ invertible matrix, and let $D \subseteq \mathbb{R}^n$ be a measurable compact set. Then $A(D)$ is also measurable, and*

$$\operatorname{vol}(A(D)) = |\det A| \operatorname{vol}(D).$$

In particular, let E be the unit cube $[0,1] \times \cdots \times [0,1]$, so that $A(E) = P$ is the parallelpiped spanned by the columns $\mathbf{v}_i$, $1 \leq i \leq n$ of A:

$$P = \{\sum_{i=1}^n t_i \mathbf{v}_i : 0 \leq t_i \leq 1\}.$$

Then $\operatorname{vol}(P) = |\det A|$.

Note that this theorem gives an interpretation of $|\det A|$ in all dimensions: it is the factor by which volume is multiplied. Also, the theorem is meaningful (and easy to check) in case A is not invertible. In this case $\det A = 0$. On the other hand, $A(D)$ is contained in the proper vector subspace $\operatorname{Im} A$, and this easily implies that $\operatorname{vol}(A(D)) = 0$ (compare the proof of Theorem 10.32).

Proof. First, it is easy to see that, if the theorem is true for A_1 and A_2, then it is true for the product A_1A_2. In fact, assuming the theorem for A_1 and A_2, if D is measurable, then $A_2(D)$ is measurable and $\operatorname{vol}(A_2(D)) = |\det A_2| \operatorname{vol}(D)$. Then $A_1(A_2(D)) = A_1A_2(D)$ is measurable, and

$$\begin{aligned}\operatorname{vol}(A_1A_2(D)) &= |\det A_1| \operatorname{vol}(A_2(D)) \\ &= |\det A_1||\det A_2| \operatorname{vol}(D) = |\det(A_1A_2)| \operatorname{vol}(D).\end{aligned}$$

The next remark is that the theorem holds for all invertible matrices $A \iff$ for every invertible matrix A, the volume of the parallelpiped P spanned by the columns of A as in the statement is $|\det A|$. Clearly, if the theorem holds, then $\operatorname{vol}(P) = |\det A| \operatorname{vol}(E) = |\det A| \cdot 1 = |\det A|$. Conversely, suppose that, for every invertible matrix A, the volume of the parallelpiped P spanned by the columns of A $|\det A|$. If A is an invertible matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and R is any rectangle $[0, t_1] \times \cdots \times [0, t_n]$, then clearly $A(R)$ is the parallelpiped spanned by $t_1\mathbf{v}_1, \ldots, t_n\mathbf{v}_n$. Thus,

$$\operatorname{vol}(A(R)) = t_1 \cdots t_n |\det A| = |\det A| \operatorname{vol}(R).$$

By translation invariance, the formula $\operatorname{vol}(A(R)) = |\det A| \operatorname{vol}(R)$ then holds for **every** rectangle, since every rectangle $[a_1, b_1] \times \cdots \times [a_n, b_n]$ is of the form $R + \mathbf{a}$, where $\mathbf{a} = (a_1, \ldots, a_n)$ and $R = [0, b_1 - a_1] \times \cdots \times [0, b_n - a_n]$. Now let D be a measurable compact set. By Remark 10.34, ∂D has measure 0. In fact, there exists a rectangle R containing D, and, for all $\epsilon > 0$, a partition of R into subrectangles, such that the sum of the volumes of all of the subrectangles in the partition meeting ∂D (in other words, not contained in $\operatorname{int}(D)$ but having nonempty intersection with D), is at most $\epsilon / |\det A| \cdot \operatorname{vol}(R)$. Since A is invertible, and hence has a continuous inverse, A is a homeomorphism. It follows easily that $A(\operatorname{int}(D)) = \operatorname{int}(A(D))$ and that $A(\partial D) = \partial A(D)$. Furthermore, if S_1 and S_2 are two different subrectangles in the partition, then $S_1 \cap S_2$ is either empty or contained in a proper vector subspace of $\mathbb{R}^n$, and the same is true for $A(S_1) \cap A(S_2)$. Using this, one shows that $\operatorname{vol}(A(\partial D)) < \epsilon$ and that

$$|\det A| \operatorname{vol}(D) - \epsilon < \operatorname{vol}(A(D)) < |\det A| \operatorname{vol}(D) + \epsilon.$$

It follows that the theorem holds for A. Notice that the proof shows a little more: suppose that, for a fixed invertible matrix A and for every rectangle R of the form $[0, t_1] \times \cdots \times [0, t_n]$ we have that $\mathrm{vol}(A(R)) = |\det A| \,\mathrm{vol}(R)$, then the theorem holds for A.

Now, for a general invertible matrix A, row reduction shows that A can be written as a product $A = A_1 \cdots A_k$ of matrices of the following types:

1. There exists an α, $1 \leq \alpha \leq n$, such that $A_i \mathbf{e}_\alpha = \lambda \mathbf{e}_\alpha$ for some $\lambda \in \mathbb{R}$, $\lambda \neq 0$, and $A_i \mathbf{e}_\beta = \mathbf{e}_\beta$ for all $\beta \neq \alpha$;

2. There exist α, β, $1 \leq \alpha, \beta \leq n$ with $\alpha \neq \beta$ such that $A_i \mathbf{e}_\alpha = \mathbf{e}_\beta$, $A_i \mathbf{e}_\beta = \mathbf{e}_\alpha$, and $A_i \mathbf{e}_\gamma = \mathbf{e}_\gamma$ for all $\gamma \neq \alpha, \beta$;

3. There exist α, β, $1 \leq \alpha, \beta \leq n$ with $\alpha \neq \beta$ such that $A_i \mathbf{e}_\alpha = \mathbf{e}_\alpha + \mathbf{e}_\beta$, and $A_i \mathbf{e}_\gamma = \mathbf{e}_\gamma$ for all $\gamma \neq \alpha$ (including the case $\gamma = \beta$).

By the remarks at the beginning of the proof and the above discussion, it suffices to check that $\mathrm{vol}(A_i(R)) = |\det A_i| \,\mathrm{vol}(R)$ for every matrix A_i of one of the types above and for every rectangle R of the form $[0, t_1] \times \cdots \times [0, t_n]$. Note that, in case (1), after reordering the variables, A_i is of the form $B_i \times \mathrm{Id}$, where $B_i \colon \mathbb{R} \to \mathbb{R}$ is multiplication by λ and Id is the identity matrix on $\mathbb{R}^{n-1}$. Similarly in cases (2) and (3), after reordering the variables, A_i is of the form $B_i \times \mathrm{Id}$, where $B_i \colon \mathbb{R}^2 \to \mathbb{R}^2$ is either given by $B_i(x_1, x_2) = (x_2, x_1)$ or $B_i(x_1, x_2) = (x_1, x_1 + x_2)$, and Id is the identity matrix on $\mathbb{R}^{n-1}$. In all cases, $|\det A_i| = |\det B_i| = |\lambda|$ in case (1) and $= 1$ in cases (2) and (3). So, to check the formula $\mathrm{vol}(A_i(R)) = |\det A_i| \,\mathrm{vol}(R)$ for rectangles R of the form $[0, t_1] \times \cdots \times [0, t_n]$, it is enough to check the formula in case $n = 1$ and A_i is multiplication by λ or $n = 2$ and A_i is one of the 2×2 matrices given above. Case (1) or case (2) are then clear, and case (3) may be verified by a direct elementary argument. This concludes the proof. □

We turn now to the proof of the general change of variable theorem.

Chapter 11

Integration on manifolds

11.1 Submanifolds and submanifolds with boundary

Let X be a smooth k-manifold in $\mathbb{R}^n$, which we shall also call a *submanifold* of $\mathbb{R}^n$. What is the basic geometry of X? What can be measured on X, and how does X "curve," both viewed in $\mathbb{R}^n$ and in some intrinsic sense? We will give a very brief discussion of what can be integrated over X. Mostly we will try to describe what can be done for curves and surfaces. Let us begin with the case of curves, i.e. 1-manifolds in $\mathbb{R}^n$. A *parametrized curve* is a C^∞ function $\mathbf{r}\colon [a,b] \to \mathbb{R}^n$. (We do not in general need that the map is C^∞; usually C^1 or C^2 or C^3 is enough, and we will not worry here about what happens at the endpoints, but just assume that all the higher derivatives in (a,b) have a continuous extension to $[a,b]$.) For geometric purposes we usually assume that $\mathbf{r}$ is injective, or in other words that $\mathbf{r}$ is *simple*, except in case $\mathbf{r}(a) = \mathbf{r}(b)$, in which case we call $\mathbf{r}$ a *simple closed curve*. We can think of $\mathbf{r}$ as the position of a particle at time t. We also have the *velocity vector*

$$\mathbf{v}(t) = \frac{d\mathbf{r}}{dt},$$

and the *acceleration*

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt}.$$

For an example, consider the helix $\mathbf{r}(t) = (\cos t, \sin t, t)$. Its velocity is $\mathbf{v}(t) = (-\sin t, \cos t, 1)$ and its acceleration is $\mathbf{a}(t) = (-\cos t, -\sin t, 0)$. As another example, we consider $\mathbf{r}(t) = (t, t^2)$, $t \in [-1,1]$, which traces out a piece of a parabola in $\mathbb{R}^2$. Here $\mathbf{v}(t) = (1, 2t)$ and $\mathbf{a}(t) = (0,2)$. If

we considered instead (t^3, t^6), we would trace out the same parabola at a different rate, with the velocity $= (3t^2, 6t^5) = 3t^2(1, 2t^3)$ and we see that the two velocity vectors always point in the same direction at corresponding points of the image of $\mathbf{r}$ except at $t = 0$ where the second velocity vector is zero: the particle hesitates instantaneously, then moves on. For a more drastic example, we consider $(t^2, t^4), t \in [-1, 1]$. In this example the particle travels half of the parabola, backwards for $t \in [-1, 0]$, then travels in the forward direction for $t \in [0, 1]$. Necessarily the velocity is zero at $t = 0$. For many reasons, and in particular for studying smooth 1-manifolds, it is natural to assume that the velocity vector is never $\mathbf{0}$, and we shall make this assumption in what follows. In this case, if we have a different way to parametrize $\mathbf{r}$, say $\mathbf{r}(t) = \mathbf{s}(u)$, where $\mathbf{s}\colon [c, d] \to \mathbb{R}^n$ is again a C^∞ function such that the velocity vector $d\mathbf{s}/du$ is never zero, and both $\mathbf{r}$ and $\mathbf{s}$ are injective, then $t = t(u)$ is a function of u by taking $\mathbf{r}^{-1}(\mathbf{s}(u))$. We shall assume for the moment that t is a differentiable function of u and vice versa (as we shall see in more generality, this is usually automatic). By the chain rule,

$$\frac{d\mathbf{s}}{du} = \frac{dt}{du}\frac{d\mathbf{r}}{dt} = \frac{dt}{du}\mathbf{v}(t).$$

Thus the line spanned by $\mathbf{v}(t)$, i.e. the tangent line, depends only on the point of the curve, not the choice of parametrization. Moreover the derivative dt/du is never zero, and in particular t is either a strictly increasing function of u (if $dt/du > 0$) or strictly decreasing function of u (if $dt/du < 0$). In the first case we say that the reparametrization is *orientation-preserving* and in the second that it is *orientation-reversing*. In the first case the curve is traced out in the same direction under the two different parametrizations (but in general at different rates) and the two velocity vectors point in the same direction, whereas in the second case the two parametrizations trace out the curve in opposite directions, and the two velocity vectors point in opposite directions as well. Of course, part of this discussion makes sense for any curve $\mathbf{r}(t)$, not necessarily one-to-one or such that $\mathbf{v}$ never vanishes: if $t = t(u)$ and $\mathbf{s}(u) = \mathbf{r}(t(u))$, then as above

$$\frac{d\mathbf{s}}{du} = \frac{dt}{du}\frac{d\mathbf{r}}{dt} = \frac{dt}{du}\mathbf{v}(t),$$

and in particular the velocity vector for $\mathbf{s}$ is a scalar multiple (possibly $\mathbf{0}$) of $\mathbf{v}$. (The same does not hold for acceleration under reparametrizations; the most you can say is that the acceleration vector for $\mathbf{s}$ is a linear combination of the velocity and acceleration vectors for $\mathbf{r}$.)

Note finally that the curve given by $\mathbf{r}$ has a *boundary*: at the points $\mathbf{r}(a) = \mathbf{p}$ and $\mathbf{r}(b) = \mathbf{q}$, we don't expect that $\mathbf{r}([a,b]) = C$ is a smooth 1-manifold. We denote $\partial C = \{\mathbf{p}, \mathbf{q}\}$. If C is oriented, then we can denote the fact that $\mathbf{q}$ is the endpoint by recording it with a $+$ sign, and that $\mathbf{p}$ is the starting point by recording it with a $-$ sign: the oriented boundary of C is then $\{-\mathbf{p}, +\mathbf{q}\}$. Here however the symbols $+$ and -1 in front of the points are just formal symbols, not to be confused with vector addition or scalar multiplication by -1. Reversing the orientation on C then switches the signs on $\mathbf{p}$ and $\mathbf{q}$.

We turn now to more general k-manifolds X. As we have seen, at least in the C^1 case, k-manifolds in $\mathbb{R}^n$ can be described (locally) as level sets of C^1 functions $F\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ such that DF is onto, as graphs of C^1 functions $G\colon \mathbb{R}^k \to \mathbb{R}^{n-k}$ (possibly after reordering the variables), or as images under one-to-one C^1 maps $H\colon U \to \mathbb{R}^n$, where U is an open set in $\mathbb{R}^k$, such that DH is also one-to-one. We will concentrate on the last approach, which is usually the one best suited to computation, and refer to such a map as a *parametrized k-manifold*. Typically, U is the interior of a standard region D and we also write $\overline{U}$, the *closure* of U, for D in this case. We shall usually assume in general that H and DH extend to continuous functions on $\overline{U}$. The map H is given as

$$H(u_1, \ldots, u_k) = (x_1(u_1, \ldots, u_k), \ldots, x_n(u_1, \ldots, u_k)),$$

where the x_i are C^1 functions of $u_1, \ldots, u_k$. Typically (think of the surface of a sphere) a manifold X cannot be completely described by such a map H, and we shall have to divide X up into various pieces which be so described. To describe how the pieces meet, we shall usually assume that X is contained in a compact set $\overline{X}$, such that $\overline{X} - X$ is a union of submanifolds of smaller dimension. We call $\overline{X} - X = \partial X$ the *boundary* of X. In the parametrized examples above, $\overline{X} = H(D) = H(\overline{U})$ and $\partial X = H(\partial D)$. A typical example is a Type I region in $\mathbb{R}^2$, $\{(x,y) : a \le x \le b, f_1(x) \le y \le f_2(x)\}$, where f_1 and f_2 are assumed to be C^1 functions, not just continuous. Here the boundary usually consists of the four curves (with endpoints), $x = a, f_1(a) \le y \le f_2(a)$, $a \le x \le b, y = f_1(x)$, $x = b, f_1(b) \le y \le f_2(b)$, $a \le x \le b, y = f_2(x)$. Of course two of the curves are straight lines. If say $f_1(a) = f_2(a)$, then there are only three curves, and if in addition $f_1(b) = f_2(b)$, then there are only two. An extreme example is the unit disk B, where ∂B is the single closed curve which is the unit circle. Note that, in general, the boundary of X need not be connected. For example, the boundary of an annulus $\{(x_1, x_2) : 1 < x_1^2 + x_2^2 < 2\}$ consists of two circles; more generally

the boundary of the annular region $\{\mathbf{x} \in \mathbb{R}^n : a < \|\mathbf{x}\| < b\}$ consists of two copies of S^{n-1}. Likewise, the boundary of a cylinder in $\mathbb{R}^3$ is two circles.

In general, we shall define a *smooth k-manifold with boundary* as follows:

Definition 11.1. Let U be an open subset of $\mathbb{R}^n$. A closed subset X of U is a *smooth k-manifold with boundary* in U if for every point $\mathbf{x} \in X$, there exists a ball B contained in U centered at $\mathbf{x}$ and a diffeomorphism $F\colon B \to V$, where V is an open subset of $\mathbb{R}^n$, such that either (i) $F(X \cap B)$ is equal to the intersection $\mathbb{R}^k \cap V$, where $\mathbb{R}^k$ is viewed as a subset of $\mathbb{R}^n$ by setting the last $n-k$ coordinates equal to zero, or (ii) $F(X \cap B)$ is equal to the intersection $(\mathbb{R}^{k-1} \times [0,\infty)) \cap V$. In other words, either X is a k-manifold near $\mathbf{x}$ or we can straighten out X to look like a closed half-space in $\mathbb{R}^k$. We refer to points of type (i) as *manifold points*. A point $\mathbf{x} \in X$ of type (ii) where $F(\mathbf{x})$ corresponds to a point of $(\mathbb{R}^{k-1} \times \{0\}) \cap V$ will be called a *boundary point*. We define ∂X to be the set of all boundary points, and call it the *boundary* of X. It is a closed subset of X. In case X is compact and $\partial X = \emptyset$, we call X a *closed* manifold.

For example, S^{n-1} is a closed manifold, but if H is the "upper hemisphere" $\{\mathbf{x} \in S^{n-1} : x_n \geq 0\}$, then H is a manifold with boundary and $\partial H = S^{n-2}$. In general, the boundary ∂X of a smooth k-manifold with boundary is a smooth $(k-1)$-manifold, and it is easy to see that $\partial(\partial X) = \emptyset$. In particular, if X is compact, then ∂X is always a closed $(k-1)$-manifold.

Remark 11.2. In many applications, the definition of manifold with boundary is not sufficient. For example, we have seen that if $X = D$ is a standard region in $\mathbb{R}^n$, then ∂D is not in general a smooth $(n-1)$-manifold; this fails even if $D = [0,1] \times \cdots \times [0,1]$ is the unit cube. A slight generalization of Definition 11.1 gives the concept of a *manifold with corners*: we require that, in the notation of Definition 11.1, the diffeomeorphism F satisfies: $F(X \cap B)$ is equal to the intersection $(\mathbb{R}^a \times ([0,\infty))^{k-a}) \cap V$ for some a, $0 \leq a \leq k$. In other words, we can locally straighten X out to look like a piece of the closed unit cube. The boundary ∂X is defined as before and (perhaps somewhat surprisingly), in a suitable sense we still have $\partial(\partial X) = \emptyset$.

To emphasize again: although the definitions of manifold with boundary or manifold with corners may look a little daunting, in practice all such manifolds will arise in the following way: let D be a standard region in $\mathbb{R}^k$, where the functions used in the defining inequalities are actually C^1, and let $H\colon U \to \mathbb{R}^n$ be a C^1 map such that H is injective and $DH_{\mathbf{a}}$ is injective for every $\mathbf{a} \in U$. Then, if ∂D is a smooth $(k-1)$-manifold, then $X = H(D)$

is a smooth k-manifold with boundary and $\partial X = H(\partial D)$. In the case of a general standard region D, ∂D locally looks like a piece of the unit cube under a suitable diffeomorphism. In this case $X = H(D)$ is a smooth k-manifold with corners and again $\partial X = H(\partial D)$. Finally, a general compact k-manifold with corners can be cut up into pieces of the above types.

Example 11.3. A standard example of a parametrized surface Σ in $\mathbb{R}^3$ is a graph surface $z = f(x, y)$ over some open set U in the (x, y)-plane. Here the parameters are x, y and $H(x, y) = (x, y, f(x, y))$. The upper hemisphere of the unit sphere is such a graph surface, with $z = \sqrt{1 - x^2 - y^2}$. Note that

$$DH = \begin{pmatrix} 0 & 1 & -\dfrac{x}{\sqrt{1-x^2-y^2}} \\ 0 & 1 & -\dfrac{y}{\sqrt{1-x^2-y^2}} \end{pmatrix},$$

and so this example fails to be differentiable along the equator (the unit circle in the xy-plane). For another example, take X to be the graph $z = xy$ over D, where D is the unit disk. In this case ∂D is the unit circle and $\partial X = \{(x, y, xy) : x^2 + y^2 = 1\}$. Of course, we could also write ∂X as a parametrized closed curve: it is given by $\mathbf{r}(t) = (\cos t, \sin t, \cos t \sin t)$, for $0 \leq t \leq 2\pi$.

Example 11.4. For an example of a parametrized surface which is not a graph parametrization, take the spherical coordinates parametrization of the unit sphere:

$$\begin{aligned} x &= \cos\theta \sin\phi; \\ y &= \sin\theta \sin\phi; \\ z &= \cos\phi, \end{aligned}$$

for $0 \leq \theta \leq 2\pi, 0 \leq \phi \leq \pi$. Note that the parametrization is not exactly 1–1 for the usual reasons with spherical coordinates, and the derivative fails to be one-to-one for $\phi = 0, \pi$. However, it is one-to-one away from a "thin" set, and this will suffice for purposes related to integration such as finding surface area.

Returning to the general situation, with

$$H(u_1, \ldots, u_k) = (x_1(u_1, \ldots, u_k), \ldots, x_n(u_1, \ldots, u_k))$$

defining the parametrized k-manifold X as above, we have the tangent vectors

$$\mathbf{T}_{u_i} = \frac{\partial H}{\partial u_i} = \left(\frac{\partial x_1}{\partial u_i}, \ldots, \frac{\partial x_n}{\partial u_i} \right).$$

For example, in case $k = 2$ and $n = 3$, so that H defines a parametrized surface in $\mathbb{R}^3$, if u and v are the coordinates on U, we have the tangent vectors

$$\mathbf{T}_u = \left(\frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, \frac{\partial z}{\partial u}\right), \mathbf{T}_v = \left(\frac{\partial x}{\partial v}, \frac{\partial y}{\partial v}, \frac{\partial z}{\partial v}\right).$$

Recall that the meaning of the vectors $\mathbf{T}_{u_i}$ is as follows: If we hold all of the variables fixed except for u_i, say $u_j = a_j$ for $i \neq j$, then $U \cap \{(a_1, \ldots, u_i, \ldots, a_k) : u_i \in \mathbb{R}\}$ is an open subset of a line with coordinate u_i, and it is sent via H to a C^1 curve in $\mathbb{R}^n$. The vector $\mathbf{T}_{u_i}$ is the velocity vector of this curve, with u_i as parameter, and in particular it is tangent to the curve.

By definition, the $\mathbf{T}_{u_i}(\mathbf{a})$ are the column vectors of the derivative $DH_{\mathbf{a}}$ and they span the tangent space $T_{\mathbf{p}}X$, where $\mathbf{p} = H(\mathbf{a})$.

In case X is a parametrized surface in $\mathbb{R}^3$ we have another way to describe the tangent plane. Recall that in $\mathbb{R}^3$ there is the cross product of two vectors $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$, which is symbolically given by taking the determinant:

$$\mathbf{v} \times \mathbf{w} = \det \begin{pmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}.$$

Thus $\mathbf{v} \times \mathbf{w} = (v_2 w_3 - v_3 w_2, v_3 w_1 - v_1 w_3, v_1 w_2 - v_2 w_1)$. The most important feature of $\mathbf{v} \times \mathbf{w}$ is that $\mathbf{v} \times \mathbf{w} = 0$ if and only if $\mathbf{v}$ and $\mathbf{w}$ are linearly dependent, and otherise $\mathbf{v} \times \mathbf{w}$ is a nonzero vector orthogonal to both $\mathbf{v}$ and $\mathbf{w}$. Among all such vectors, it is uniquely specified by saying that its length is the area of the parallelogram spanned by $\mathbf{v}$ and $\mathbf{w}$, and that it points in the direction given by the right hand rule. Thus, for a surface Σ in $\mathbb{R}^3$, we can define a *unit normal* $\mathbf{N}$ by the rule

$$\mathbf{N} = \frac{\mathbf{T}_u \times \mathbf{T}_v}{\|\mathbf{T}_u \times \mathbf{T}_v\|}.$$

The normal $\mathbf{N}$ is then orthogonal to the tangent plane. For example, if Σ is also described as a level set $f(x, y, z) = 0$, then

$$\mathbf{N} = \pm \frac{\nabla f}{\|\nabla f\|}.$$

For surfaces in $\mathbb{R}^n, n > 3$, there is no analogous construction. However, we can still define a "product" of $n-1$ vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ in a similar way, and this will be a vector perpendicular to all of them. Hence the construction also works for $(n-1)$-manifolds in $\mathbb{R}^n$—these are called *hypersurfaces*.

Example 11.5. If Σ is a graph surface $(x, y, f(x, y))$, then using x and y as the natural parameters on Σ we find:

$$\begin{aligned}\mathbf{T}_x &= \left(1, 0, \frac{\partial f}{\partial x}\right);\\ \mathbf{T}_y &= \left(0, 1, \frac{\partial f}{\partial y}\right);\\ \mathbf{T}_x \times \mathbf{T}_y &= \left(-\frac{\partial f}{\partial x}, -\frac{\partial f}{\partial y}, 1\right).\end{aligned}$$

For another example, using the spherical coordinates parametrization of the unit sphere, a little calculation gives:

$$\begin{aligned}\mathbf{T}_\theta &= (-\sin\theta\sin\phi, \cos\theta\sin\phi, 0);\\ \mathbf{T}_\phi &= (\cos\theta\cos\phi, \sin\theta\cos\phi, -\sin\phi);\\ \mathbf{T}_\theta \times \mathbf{T}_\phi &= (-\cos\theta\sin^2\phi, -\sin\theta\sin^2\phi, -\sin\phi\cos\phi) = -\sin\phi(x, y, z).\end{aligned}$$

Here $\mathbf{T}_\theta \times \mathbf{T}_\phi$ is zero for $\sin\phi = 0$ because the parmetrization sends the entire line segments $\phi = 0, \pi$ and $\theta \in [0, 2\pi]$ to one point.

If we compare the above calculation with the parametrization of the upper hemisphere using the graph parametrization with $f(x, y) = \sqrt{1 - x^2 - y^2}$, we get instead

$$\mathbf{T}_x \times \mathbf{T}_y = \left(\frac{x}{\sqrt{1 - x^2 - y^2}}, \frac{x}{\sqrt{1 - x^2 - y^2}}, 1)\right) = \frac{1}{z}(x, y, z).$$

Note that $\mathbf{T}_x \times \mathbf{T}_y$ fails to be defined where $z = 0$. The normal vectors $\mathbf{T}_x \times \mathbf{T}_y$ and $\mathbf{T}_\theta \times \mathbf{T}_\phi$ differ by a negative scalar function: $\mathbf{T}_x \times \mathbf{T}_y$ points outward while $\mathbf{T}_\theta \times \mathbf{T}_\phi$ points inward. Of course, we can also calculate the normal vector via the gradient of the function $f(x, y, z) = x^2 + y^2 + z^2$, with gradient $2(x, y, z)$. This again differs from our answer in the other cases by a scalar function.

We turn now to the problem of how to compare two different parametrizations. Suppose that $H_1 \colon U_1 \to X$ and $H_2 \colon U_2 \to X$ are two different parametrizations of the same manifold. Here U_1 and U_2 are two possibly different open subsets of $\mathbb{R}^k$, and both H_1 and H_2 are C^1 functions such that $H_1 \colon U_1 \to X$ is a homeomorphism, and similarly for H_2. Let $\mathbf{u} = (u_1, \ldots, u_k)$ be the coordinates on U_1 and $\mathbf{s} = (s_1, \ldots, s_k)$ be coordinates on U_2. Then $T(\mathbf{s}) = H_1^{-1} \circ H_2(\mathbf{s})$ is well-defined, and by definition

$$H_1 \circ T(\mathbf{s}) = H_1 \circ (H_1^{-1} \circ H_2)(\mathbf{s}) = H_2(\mathbf{s}).$$

Thus, using T to write $\mathbf{u} = T(\mathbf{s}) = \mathbf{u}(\mathbf{s})$, we see that H_2 is a reparametrization of H_1, and vice versa.

Lemma 11.6. *The function T is a C^1 diffeomorphism from U_2 to U_1.*

Proof. It is easy to see that T is one-to-one and onto. We sketch the argument that it is differentiable. It is enough to check this at every point $\mathbf{x} \in U_2$. After reordering the coordinates and possibly shrinking U_1, we know by the proof of the second version of the implicit function theorem that $p \circ H_1$ is a local C^1 diffeomorphism onto an open subset of $\mathbb{R}^k$, where $p(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$ is projection onto the first k coordinates. A similar statement holds for $p \circ H_2$. It then follows for $T = H_1^{-1} \circ H_2$. □

Suppose that H_1 and H_2 are two different parametrizations as above, with $\mathbf{q} \in U_2$ and $\mathbf{q}' = T(\mathbf{q})$ the corresponding point of U_1. The chain rule says that $(DH_2)_{\mathbf{q}} = (DH_1)_{\mathbf{q}'} \cdot DT_{\mathbf{q}}$, where $DT_{\mathbf{q}}$ is an invertible $k \times k$ matrix. It follows that the columns of $(DH_1)_{\mathbf{q}'}$ and $(DH_2)_{\mathbf{q}}$ span the same k-dimensional subspace of $\mathbb{R}^k$, and the matrix multiplication $(DH_2)_{\mathbf{q}} = (DH_1)_{\mathbf{q}'} \cdot DT_{\mathbf{q}}$ tells us how to write the columns of DH_2, i.e. the vectors $\mathbf{T}_{s_j}$, as linear combinations of the $\mathbf{T}_{u_i}$.

For a concrete example, let Σ be a surface in $\mathbb{R}^3$. A reparametrization of Σ consists in writing u and v as functions of two variables $(r, s) \in E$, where the function $T(r, s) = (u(r, s), v(r, s))$ is a diffeomorphism of E onto D. Thus in the new coordinates we see that

$$T_r = \frac{\partial u}{\partial r}\mathbf{T}_u + \frac{\partial v}{\partial r}\mathbf{T}_v;$$
$$T_s = \frac{\partial u}{\partial s}\mathbf{T}_u + \frac{\partial v}{\partial s}\mathbf{T}_v.$$

Since T is a diffeomorphism, the matrix $\dfrac{\partial(u, v)}{\partial(r, s)}$ is invertible, and hence the span of $\mathbf{T}_u$ and $\mathbf{T}_v$ at every point is the same as the span of $\mathbf{T}_r$ and $\mathbf{T}_s$. In particular the tangent plane does not depend on the parametrization.

The calculations of the different normal vectors for the sphere are a special case of a general fact about reparametrizations which follows from the formula above for $\mathbf{T}_r$ and $\mathbf{T}_s$ in terms of $\mathbf{T}_u$ and $\mathbf{T}_v$:

$$\mathbf{T}_r \times \mathbf{T}_s = \left(\det \frac{\partial(u, v)}{\partial(r, s)}\right) \mathbf{T}_u \times \mathbf{T}_v.$$

Example 11.7. In the case of the sphere, we can write the coordinates x and y in terms of θ and ϕ: $x = \cos\theta\sin\phi, y = \sin\theta\sin\phi$. Then

$$\frac{\partial(x,y)}{\partial(\theta,\phi)} = \begin{pmatrix} -\sin\theta\sin\phi & \cos\theta\cos\phi \\ \cos\theta\cos\phi & \sin\theta\cos\phi \end{pmatrix},$$

so that

$$\det\frac{\partial(x,y)}{\partial(\theta,\phi)} = -\sin\phi\cos\phi.$$

Since $z = \cos\phi$, we can check directly that $\mathbf{T}_\theta \times \mathbf{T}_\phi = -(\sin\phi\cos\phi)\mathbf{T}_x \times \mathbf{T}_y$.

Returning to the case of a general surface Σ, if $\det\dfrac{\partial(u,v)}{\partial(r,s)} > 0$ then the corresponding normal vectors point in the same direction, whereas if $\det\dfrac{\partial(u,v)}{\partial(r,s)} < 0$ they point in opposite directions. In case the normal vectors point in the same direction, we say that the reparametrization is *orientation preserving*, and in the other case it is *orientation reversing*. A choice of orientation thus amounts to a choice of a side for Σ, the side away from which $\mathbf{N}$ points, and reversing the orientation is the same as changing sides. Some surfaces (the Möbius band, for example) do not have a consistent choice of side (outward normal, in this case) and are therefore said to be *non-orientable*. Orientation is similarly defined for k-manifolds X with $k \geq 2$, by agreeing that if H_1 and H_2 are two parametrizations of X with $H_2 = H_1 \circ T$, then they define the same orientation if $\det T > 0$ and opposite orientations if $\det T < 0$. If $H\colon U \to X \subseteq \mathbb{R}^n$ is a parametrized manifold, then H defines an orientation on X. Likewise, if $X \subseteq \mathbb{R}^3$ is a smooth surface defined as a level set $f^{-1}(0)$ for some C^1 function f, then X is orientable and an orientation is given by taking the normal to be $\nabla f/\|\nabla f\|$. In fact, one can show that every **compact** surface in $\mathbb{R}^3$ can be oriented. But there exist manifolds such as the Möbius band in $\mathbb{R}^3$ or the Klein bottle, which is a compact 2-manifold in $\mathbb{R}^4$, which cannot be described by a single parametrization and which furthermore cannot be oriented.

11.2 Cross product and wedge product

As it stands, cross product is an operation which only exists for two vector in $\mathbb{R}^3$. To put this in a more general context, we introduce wedge product, which generalizes both the determinant in $\mathbb{R}^n$ for any n and cross product in $\mathbb{R}^3$. However, for the most part we shall just state its basic properties

without proof. In particular, we shall not discuss how it behaves under change of basis.

Begin with $\mathbb{R}^n$ with the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$. We make a new vector space $\bigwedge^k \mathbb{R}^n$ as follows: it has a basis given by formal symbols $\mathbf{e}_I$, where I is a subset of $\{1, \ldots, n\}$ with k elements. If $I = \{i_1, i_2, \ldots, i_k\}$ where $i_1 < i_2 < \cdots < i_k$ then we also write $\mathbf{e}_I = \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_k}$, where the reason for this notation will become clearer in a minute. Thus $\bigwedge^k \mathbb{R}^n$ has a basis with $\binom{n}{k}$ elements. For example, $\bigwedge^1 \mathbb{R}^n$ is the same as $\mathbb{R}^n$ (identifying $\mathbf{e}_{\{i\}}$ with $\mathbf{e}_i$). For $n = 2$, $\bigwedge^2 \mathbb{R}^n$ has a basis with the $n(n-1)/2$ elements $\mathbf{e}_i \wedge \mathbf{e}_j$ for $i < j$. In general $\bigwedge^k \mathbb{R}^n$ and $\bigwedge^{n-k} \mathbb{R}^n$ have a basis with the same number of elements and as we shall see can in fact be identified. We can think of $\bigwedge^0 \mathbb{R}^n$ is as a one-dimensional vector space with basis vector $\mathbf{e}_\emptyset$ and we identify $a\mathbf{e}_\emptyset$ with the real number $a \in \mathbb{R}$. The vector space $\bigwedge^n \mathbb{R}^n$ has a basis with one element, $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \cdots \wedge \mathbf{e}_n$, and $\bigwedge^k \mathbb{R}^n = \{0\}$ for $k > n$.

Now we can define an associative bilinear product from $\bigwedge^k \mathbb{R}^n \times \bigwedge^\ell \mathbb{R}^n$ to $\bigwedge^{k+\ell} \mathbb{R}^n$ as follows: define $\mathbf{e}_i \wedge \mathbf{e}_i = 0$, and $\mathbf{e}_j \wedge \mathbf{e}_i = -\mathbf{e}_i \wedge \mathbf{e}_j$. More generally, suppose that $I = \{i_1, i_2, \cdots, i_k\}$ with $i_1 < i_2 < \cdots < i_k$, and abbreviate $\mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_k}$ by $\mathbf{e}_I$. Let $\mathbf{e}_J = \mathbf{e}_{j_1} \wedge \mathbf{e}_{j_2} \wedge \cdots \wedge \mathbf{e}_{j_\ell}$, say, with $J = \{j_1, j_2, \cdots, j_\ell\}$, where $j_1 < j_2 < \cdots < j_\ell$. Then

$$\mathbf{e}_I \wedge \mathbf{e}_J = \begin{cases} \mathbf{0}, & \text{if } I \cap J \neq \emptyset; \\ \pm \mathbf{e}_{I \cup J}, & \text{if } I \cap J = \emptyset. \end{cases}$$

Here the sign is determined by the number of times we have to switch the order until the factors are multiplied in increasing order. For example,

$$(\mathbf{e}_2 \wedge \mathbf{e}_4) \wedge (\mathbf{e}_1 \wedge \mathbf{e}_3) = -\mathbf{e}_2 \wedge \mathbf{e}_1 \wedge \mathbf{e}_4 \wedge \mathbf{e}_3 = \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_4 \wedge \mathbf{e}_3 = -\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_4.$$

To multiply linear combinations $\sum_I a_I \mathbf{e}_I$ and $\sum_J b_J \mathbf{e}_J$, the rule is to expand out and get $\sum_{I,J} a_I b_J \mathbf{e}_I \wedge \mathbf{e}_J$. In this way, we get an associative bilinear operation from $\bigwedge^k \mathbb{R}^n \times \bigwedge^\ell \mathbb{R}^n$ to $\bigwedge^{k+\ell} \mathbb{R}^n$. It is not however commutative. In fact, we have the rule: if $\mathbf{v} \in \bigwedge^k \mathbb{R}^n$ and $\mathbf{w} \in \bigwedge^\ell \mathbb{R}^n$, then

$$\mathbf{w} \wedge \mathbf{v} = (-1)^{k\ell} \mathbf{v} \wedge \mathbf{w}.$$

Thus, if one of k or ℓ is even, then $\mathbf{w} \wedge \mathbf{v} = \mathbf{v} \wedge \mathbf{w}$, but if both k and ℓ are odd then $\mathbf{w} \wedge \mathbf{v} = -\mathbf{v} \wedge \mathbf{w}$. Such a product is sometimes called *anti-commutative* or *skew-commutative*. Note that by the definitions we always have

$$\mathbf{e}_\emptyset \wedge \mathbf{e}_I = \mathbf{e}_I \wedge \mathbf{e}_\emptyset = \mathbf{e}_I,$$

so that $\mathbf{e}_\emptyset$ acts like a multiplicative identity (which is one reason why we agree to denote it by 1).

For example, given $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3$ and $\mathbf{w} = w_1\mathbf{e}_1 + w_2\mathbf{e}_2 + w_3\mathbf{e}_3$ in $\mathbb{R}^3$, then a calculation gives:

$$\mathbf{v} \wedge \mathbf{w} = (v_2w_3 - v_3w_2)\mathbf{e}_2 \wedge \mathbf{e}_3 + (v_3w_1 - v_1w_3)\mathbf{e}_3 \wedge \mathbf{e}_1 + (v_1w_2 - v_2w_1)\mathbf{e}_1 \wedge \mathbf{e}_2.$$

If we identify $\bigwedge^2 \mathbb{R}^3$ with $\mathbb{R}^3$ by letting $\mathbf{e}_2 \wedge \mathbf{e}_3$ correspond to $\mathbf{e}_1$, $\mathbf{e}_3 \wedge \mathbf{e}_1$ to $\mathbf{e}_2$, and $\mathbf{e}_1 \wedge \mathbf{e}_2$ to $\mathbf{e}_3$, then we see that $\mathbf{v} \wedge \mathbf{w}$ corresponds to the cross product $\mathbf{v} \times \mathbf{w}$. Thus cross product is a special case of wedge product. More generally, given $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + \cdots + v_n\mathbf{e}_n$ and $\mathbf{w} = w_1\mathbf{e}_1 + w_2\mathbf{e}_2 + \cdots + w_n\mathbf{e}_n$ in $\mathbb{R}^n$, then

$$\mathbf{v} \wedge \mathbf{w} = \sum_{i<j} (v_iw_j - v_jw_i)(\mathbf{e}_i \wedge \mathbf{e}_j).$$

Thus $\mathbf{v} \wedge \mathbf{w}$ records all of the determinants of the 2×2 matrices we can form by choosing 2 columns of the matrix $\begin{pmatrix} v_1 & v_2 & \dots & v_n \\ w_1 & w_2 & \dots & w_n \end{pmatrix}$. Note that there are $n(n-1)/2$ of these, and this number is strictly bigger than n for $n > 3$.

For another example, take vectors

$$\mathbf{v}_1 = (a_{11}, \dots, a_{n1}), \dots, \mathbf{v}_n = (a_{1n}, \dots, a_{nn}) \in \mathbb{R}^n.$$

Then the wedge product $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n$ is just

$$\det(a_{ij})\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n.$$

In general, if we take k vectors $\mathbf{v}_1 = (a_{11}, \dots, a_{n1}), \dots, \mathbf{v}_k = (a_{1k}, \dots, a_{nk}) \in \mathbb{R}^n$, then the wedge product $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k$ will be a sum of the basis vectors $\mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_k} = \mathbf{e}_I$ and the coefficient of $\mathbf{e}_I$ is the $k \times k$ determinant $\det(a_{i_r,s})$, where $1 \le r, s \le k$. Thus the coefficients of $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k$ are just the determinants of all the possible square $k \times k$ matrices we can find inside of the $n \times k$ matrix (a_{ij}). Thus wedge product is a generalization of determinant, and you can think of it as a fancy way to keep track of all of the $\binom{n}{k}$ determinants that we can build out of an $n \times k$ matrix (a_{ij}) (for $n \ge k$).

Let us make precise the way in which $\bigwedge^k \mathbb{R}^n$ and $\bigwedge^{n-k} \mathbb{R}^n$ can be identified. Given a subset $I \subseteq \{1, \dots, n\}$ with k elements, it naturally determines (and is determined by) a subset of $\{1, \dots, n\}$ with $n - k$ elements, namely $\{1, \dots, n\} - I$. In other words the function $F(I) = \{1, \dots, n\} - I$ defines a

bijection from the set of all subsets $\{1,\dots,n\}$ with k elements to the set of all subsets $\{1,\dots,n\}$ with $n-k$ elements, whose inverse is defined by the same formula. This bijection is the reason for the combinatorial identity $\binom{n}{k} = \binom{n}{n-k}$. We then define the *Hodge $*$-operator* from $\bigwedge^k \mathbb{R}^n$ to $\bigwedge^{n-k} \mathbb{R}^n$as follows: for a basis vector $\mathbf{e}_I$, define

$$*\mathbf{e}_I = \pm \mathbf{e}_{\{1,\dots,n\}-I},$$

where the sign is chosen so that

$$\mathbf{e}_I \wedge *\mathbf{e}_I = \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n,$$

noting that in any case $\mathbf{e}_I \wedge \mathbf{e}_{\{1,\dots,n\}-I} = \pm \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n$ by the definition of wedge product. Thus for example, in $\mathbb{R}^3$, $*(\mathbf{e}_1 \wedge \mathbf{e}_2) = \mathbf{e}_3$. But $*(\mathbf{e}_1 \wedge \mathbf{e}_3) = -\mathbf{e}_2$, since

$$(\mathbf{e}_1 \wedge \mathbf{e}_3) \wedge (-\mathbf{e}_2) = -(-\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3) = \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3.$$

By definition, identifying $1 \in \mathbb{R}$ with $\mathbf{e}_\emptyset \in \bigwedge^0 \mathbb{R}^n$,

$$*1 = \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \cdots \wedge \mathbf{e}_n \text{ and } * \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \cdots \wedge \mathbf{e}_n = 1.$$

Having defined $*$ on basis vectors, we can extend it to a linear map $\bigwedge^k \mathbb{R}^n \to \bigwedge^{n-k} \mathbb{R}^n$ via the formula

$$* \left(\sum_I a_I \mathbf{e}_I \right) = \sum_I a_I (*\mathbf{e}_I).$$

We leave it as homework to check that $*$ is a bijection, and hence an isomorphism; in fact, $*^{-1} = \pm *$, where the second $*$-operator is the one from $\bigwedge^{n-k} \mathbb{R}^n$ to $\bigwedge^{n-(n-k)} \mathbb{R}^n = \bigwedge^k \mathbb{R}^n$.

Using the $*$-operator, we see that the cross product of two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$ is just given by the formula

$$\mathbf{v} \times \mathbf{w} = *(\mathbf{v} \wedge \mathbf{w}).$$

This leads to a generalization of cross product which involves $n-1$ vectors in $\mathbb{R}^n$: if $\mathbf{v}_1, \dots, \mathbf{v}_{n-1} \in \mathbb{R}^n$, define

$$\mathbf{v}_1 \times \cdots \times \mathbf{v}_{n-1} = *(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_{n-1}).$$

It turns out that the generalized cross product $\mathbf{v}_1 \times \cdots \times \mathbf{v}_{n-1}$ produces a vector in $\mathbb{R}^n$ which is orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$. As we shall see, its length $\|\mathbf{v}_1 \times \cdots \times \mathbf{v}_{n-1}\|$ also has meaning.

A second application of the $*$-operator is the following: There is an inner product on $\bigwedge^k \mathbb{R}^n$ such that the $\mathbf{e}_I$ are an orthonormal basis. In fact, we can define, for $v, w \in \bigwedge^k \mathbb{R}^n$,

$$\langle v, w \rangle = *(v \wedge *w) \in \bigwedge^0 \mathbb{R}^n = \mathbb{R}.$$

Note that, with this definition, if I and J both have k elements, then

$$\langle \mathbf{e}_I, \mathbf{e}_J \rangle = *(\mathbf{e}_I \wedge *\mathbf{e}_J) = \begin{cases} *(e_1 \wedge \cdots \wedge e_n) = 1, & \text{if } I = J; \\ *0 = 0, & \text{if } I \neq J. \end{cases}$$

Hence the $\mathbf{e}_I$ are an orthonormal basis as claimed. It is not hard to check that, for $k = 1$, this inner product is the usual inner product on $\mathbb{R}^n$. More generally, we have the formula

$$\langle \mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k, \mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_k \rangle = \det(\langle \mathbf{v}_i, \mathbf{w}_j \rangle),$$

the determinant of the $k \times k$ matrix whose $(i,j)^{\text{th}}$ entry is $\langle \mathbf{v}_i, \mathbf{w}_j \rangle$. We denote the corresponding norm on $\bigwedge^k \mathbb{R}^n$ by $\|\cdot\|$, so that $\|v\|^2 = \langle v, v \rangle$. It is given by:

$$\left\| \sum_I a_I \mathbf{e}_I \right\|^2 = \sum_I a_I^2.$$

A basic fact which we shall not prove is the following: let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an arbitrary orthonormal basis of $\mathbb{R}^n$. Then we can write any element $v \in \bigwedge^k \mathbb{R}^n$ in terms of the orthonormal basis:

$$v = \sum_I a_I \mathbf{e}_I = \sum_I b_I \mathbf{u}_I,$$

where are before, if $I = \{i_1, \ldots, i_k\}$ with $i_1 < \cdots < i_k$, we define $\mathbf{u}_I = \mathbf{u}_{i_1} \wedge \cdots \wedge \mathbf{u}_{i_k}$. Then the $\{\mathbf{u}_I\}$ are again an orthonormal basis of $\bigwedge^k \mathbb{R}^n$, so that

$$\|v\| = \sum_I a_I^2 = \sum_I b_I^2.$$

Why is wedge product important? One reason has to do with the study of k-dimensional vector subspaces of $\mathbb{R}^n$. If V is such a subspace, then it has a basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$; in fact it has many such. If $\mathbf{w}_1, \ldots, \mathbf{w}_k$ is another such basis, then there exists an invertible $k \times k$ matrix $B = (b_{ij})$ such that $\mathbf{v}_i = \sum_{j=1}^k b_{ij}\mathbf{w}_k$. Now we can define the element $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k$ of $\bigwedge^k \mathbb{R}^n$. If instead we consider $\mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_k$, then one can show that

$$\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k = (\det B)\mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_k.$$

Thus the k-dimensional subspace V in $\mathbb{R}^n$ determines a line in the vector space $\bigwedge^k \mathbb{R}^n$. One can show that conversely the line determines V. In this way we have replaced complicated geometric objects, k-dimensional vector subspaces of $\mathbb{R}^n$, with simpler geometric objects—lines—in a more complicated vector space, namely $\bigwedge^k \mathbb{R}^n$. One problem is that not every line in $\bigwedge^k \mathbb{R}^n$ arises from a k-dimensional subspace of $\mathbb{R}^n$, and it is an interesting problem to write down the conditions on a line in $\bigwedge^k \mathbb{R}^n$ which say that it is of the form $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k$.

Another application of wedge product has to do with generalized k-volume. First consider the area of the parallelogram P defined by two linearly independent vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, i.e. $P = \{t\mathbf{v} + s\mathbf{w} : 0 \leq t \leq 1, 0 \leq s \leq 1\}$. We claim that the area of P is

$$\sqrt{\|\mathbf{v}\|^2\|\mathbf{w}\|^2 - \langle \mathbf{v}, \mathbf{w}\rangle^2}.$$

To see this, note that $\langle \mathbf{v}, \mathbf{w}\rangle = \|\mathbf{v}\|\|\mathbf{w}\| \cos\theta$, where θ is the angle between the two vectors, and thus

$$\|\mathbf{v}\|^2\|\mathbf{w}\|^2 - \langle \mathbf{v}, \mathbf{w}\rangle^2 = \|\mathbf{v}\|^2\|\mathbf{w}\|^2(1 - \cos^2\theta) = \|\mathbf{v}\|^2\|\mathbf{w}\|^2 \sin^2\theta,$$

which is the base times the height, thinking of $\mathbf{v}$ as the base, and so is the area of P. Another way to think of this formula is to see directly that the height is given by the length of the difference between $\mathbf{w}$ and its projection to $\mathbf{v}$, so that the area is

$$\|\mathbf{v}\| \cdot \left\| \mathbf{w} - \frac{\langle \mathbf{v}, \mathbf{w}\rangle}{\|\mathbf{v}\|^2}\mathbf{v} \right\|.$$

Thus the square of the area is

$$\|\mathbf{v}\|^2 \left(\|\mathbf{w}\|^2 - 2\frac{\langle \mathbf{v}, \mathbf{w}\rangle\langle \mathbf{v}, \mathbf{w}\rangle}{\|\mathbf{v}\|^2} + \frac{\langle \mathbf{v}, \mathbf{w}\rangle^2}{\|\mathbf{v}\|^4}\|\mathbf{v}\|^2 \right) = \|\mathbf{v}\|^2\|\mathbf{w}\|^2 - \langle \mathbf{v}, \mathbf{w}\rangle^2,$$

which is the first formula. Finally, note that, if $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$ in $\mathbb{R}^n$, then we have the identity

$$\|\mathbf{v}\|^2\|\mathbf{w}\|^2 - \langle \mathbf{v}, \mathbf{w}\rangle^2 = \sum_{i<j}(v_i w_j - v_j w_i)^2,$$

which you may (or may not) enjoy checking. This just says that

$$\text{area}(P) = \|\mathbf{v} \wedge \mathbf{w}\|,$$

using the norm on $\bigwedge^2 \mathbb{R}^n$ defined above.

In general, k linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in $\mathbb{R}^n$ span a "parallelpiped"

$$P = \left\{ \sum_{i=1}^{k} t_i \mathbf{v}_i : 0 \leq t_i \leq 1 \text{ for all } i \right\}.$$

We would like to assign a meaning to the k-volume of P. A general principle is that, if we have the usual inclusion of $\mathbb{R}^k$ in $\mathbb{R}^n$ as the span of $\mathbf{e}_1, \ldots, \mathbf{e}_k$, and P is contained in $\mathbb{R}^k$, then the k-volume of P in $\mathbb{R}^n$ should agree with its k-volume as previously defined in $\mathbb{R}^k$. For example, if P is the parallelpiped spanned by $\mathbf{e}_1, \ldots, \mathbf{e}_k$, then we would expect the k-volume of P, viewed as a subset of $\mathbb{R}^n$, to still be 1. More generally, if P is the parallelpiped spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$, where the $\mathbf{v}_i$ are linear combinations of just the first k standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_k$, then we would expect the k-volume of P as a subset of $\mathbb{R}^n$ to be its k-volume in $\mathbb{R}^k$, namely $|\det(\mathbf{v}_1, \ldots, \mathbf{v}_k)|$. A similar statement should hold if we replace the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ by **any** orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$, since volume should be unchanged by isometries. Given any collection $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of k linearly independent vectors in $\mathbb{R}^n$, the Gram-Schmidt procedure tells us that there exists an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ such that $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$. In this case, if $\mathbf{v}_i = \sum_{i=1}^{k} a_{ij}\mathbf{u}_j$, then we expect that the k-volume of P is given by $|\det(a_{ij})|$. But $|\det(a_{ij})| = \|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k\|$ since $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k = \pm(\det(a_{ij})\mathbf{u}_1 \wedge \cdots \wedge \mathbf{u}_k$. Thus we are essentially led to the following definition:

Definition 11.8. Suppose that P is the k-dimensional parallelpiped in $\mathbb{R}^n$ spanned by the linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Then we define:

$$\text{vol}(P) = \|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k\|.$$

Of course, we could write down a formula for $\|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k\|$, and so for $\text{vol}(P)$, which just involves $\binom{n}{k}$ $k \times k$ determinants, without ever mentioning wedge product. This would be a more elementary but notationally more cumbersome way to define k-volume.

11.3 Arc length and curvature

Next we discuss arc length, surface area, and more general notions of the volume of a submanifold. As usual, we begin with curves. By definition the *speed* of $\mathbf{r}$ is $\|\mathbf{v}(t)\|$, the magnitude of velocity. In case the speed is always one, for example in case $\mathbf{r}(t) = \mathbf{p}_0 + t\mathbf{u}$ where $\mathbf{u}$ is a unit vector, or in case $\mathbf{r}(t) = (\cos t, \sin t)$ is the unit circle, then we say that the parametrization is a *unit speed parametrization*. A *constant speed parametrization* is defined similarly. For example, a straight line $\mathbf{r}(t) = \mathbf{p}_0 + t\mathbf{v}$ where $\mathbf{v}$ is an arbitrary nonzero vector, is a constant speed parametrization. The helix $\mathbf{r}(t) = (\cos t, \sin t, t)$ gives another example: here $\mathbf{v}(t) = (-\sin t, \cos t, 1)$ and hence $\|\mathbf{v}(t)\| = \sqrt{2}$.

One geometric way in which speed comes up is in trying to define arc length. Suppose that $\mathbf{r}\colon [a,b] \to \mathbb{R}^n$ is a curve. To define the arc length of $\mathbf{r}$ we could first try to approximate $\mathbf{r}$ by straight lines. Subdivide the interval $[a,b]$ via a partition P, via $a = t_0 < t_1 < \cdots < t_N = b$. For a partition of small size, we expect that the curve is well-approximated by the line segments from $\mathbf{r}(t_{i-1})$ to $\mathbf{r}(t_i)$. Now if $\mathbf{r}(t) = (x_1(t), \ldots, x_n(t))$, then the length of the line segment from $\mathbf{r}(t_{i-1})$ to $\mathbf{r}(t_i)$ is

$$\|\mathbf{r}(t_i) - \mathbf{r}(t_{i-1})\| = \sqrt{(x_1(t_i) - x_1(t_{i-1}))^2 + \cdots + (x_n(t_i) - x_n(t_{i-1}))^2}.$$

Lemma 11.9. *Suppose that $\mathbf{v}(t)$ extends to a continuous function on $[a,b]$. Then, for all $\epsilon > 0$, there exists a δ such that, if the size of the partition P is less than δ, i.e. if $t_i - t_{i-1} < \delta$ for all i, then*

$$\left| \sum_{i=1}^{N} \sqrt{(x_1(t_i) - x_1(t_{i-1}))^2 + \cdots + (x_n(t_i) - x_n(t_{i-1}))^2} - \int_a^b \|\mathbf{v}(t)\| dt \right| < \epsilon.$$

Proof. By the mean value theorem, for each i, $1 \le i \le N$, and each k, $1 \le k \le n$, there exists a $c_{k,i} \in [t_{i-1}, t_i]$ such that

$$x_k(t_i) - x_k(t_{i-1}) = x_k'(c_{k,i})(t_i - t_{i-1}).$$

Summing these terms up, we see that

$$\begin{aligned} &\sum_{i=1}^{N} \sqrt{(x_1(t_i) - x_1(t_{i-1}))^2 + \cdots + (x_n(t_i) - x_n(t_{i-1}))^2} \\ &\quad = \sum_{i=1}^{n-1} \sqrt{(x_1'(c_{1,i}))^2 + \cdots + (x_n'(c_{n,i}))^2}(t_i - t_{i-1}). \end{aligned}$$

Since each of the components x_i' of $\mathbf{v}$ is continuous, the function

$$F(s_1, \ldots, s_n) = \sqrt{(x_1'(s_1))^2 + \cdots + (x_n'(s_n))^2}$$

is continuous on $[a,b] \times \cdots \times [a,b]$, and hence is uniformly continuous there. Working with the $\| \cdot \|_1$ norm on $\mathbb{R}^n$, this means that, given $\epsilon > 0$, there exists a $\delta > 0$ such that, for all $\mathbf{s}_1, \mathbf{s}_2 \in [a,b] \times \cdots \times [a,b]$, if $|s_{1,k} - s_{2,k}| < \delta$ for all k, then

$$\left| \sqrt{(x_1'(s_{1,1}))^2 + \cdots + (x_n'(s_{1,n}))^2} - \sqrt{(x_1'(s_{2,1}))^2 + \cdots + (x_n'(s_{2,n}))^2} \right| < \epsilon/2(b-a).$$

Thus, if the size of the partition P is less than δ, then, for every i, $1 \leq i \leq N$,

$$|\sqrt{(x_1'(c_{1,i}))^2 + \cdots + (x_n'(c_{n,i}))^2} - \sqrt{(x_1'(t_i))^2 + \cdots + (x_n'(t_i))^2}| < \epsilon/2(b-a).$$

Summing over all i gives

$$\left| \sum_{i=1}^{N} \sqrt{(x_1'(c_{1,i}))^2 + \cdots + (x_n'(c_{n,i}))^2}(t_i - t_{i-1}) - \sum_{i=1}^{N} \|\mathbf{v}(t_i)\|(t_i - t_{i-1}) \right|$$
$$< \sum_{i=1}^{N} (\epsilon/2(b-a))(t_i - t_{i-1}) < \epsilon/2.$$

If the size of P is small enough, then $U(\|\mathbf{v}\|, P) - L((\|\mathbf{v}\|, P)$ is small, say less than $\epsilon/4$, so that $\left| \int_a^b \|\mathbf{v}(t)\| dt - L(\|\mathbf{v}\|, P) \right| < \epsilon/4$ as well. As

$$L((\|\mathbf{v}\|, P) \leq \sum_{i=1}^{N} \|\mathbf{v}(t_i)\|(t_i - t_{i-1}) \leq U(\|\mathbf{v}\|, P),$$

for such P we can assume that

$$\left| \sum_{i=1}^{N} \|\mathbf{v}(t_i)\|(t_i - t_{i-1}) - \int_a^b \|\mathbf{v}(t)\| dt \right| \leq \left| \sum_{i=1}^{N} \|\mathbf{v}(t_i)\|(t_i - t_{i-1}) - L(\|\mathbf{v}\|, P) \right| + \left| L(\|\mathbf{v}\|, P) - \int_a^b \|\mathbf{v}(t)\| dt \right|$$
$$< \epsilon/4 + \epsilon/4 = \epsilon/2.$$

Combining these gives

$$\left| \sum_{i=1}^{N} \sqrt{(x_1(t_i) - x_1(t_{i-1}))^2 + \cdots + (x_n(t_i) - x_n(t_{i-1}))^2} - \int_a^b \|\mathbf{v}(t)\| dt \right| < \epsilon.$$

□

In physical terms, we see that the integral of speed equals the total distance traveled (which is true whether or not $\mathbf{r}$ is injective). Note that this integral is unchanged under reparametrization, since for a different parametrization with $t = t(u)$, $u \in [c, d]$, we have

$$\int_c^d \left\| \frac{d\mathbf{r}}{du} \right\| du = \int_c^d \left| \frac{dt}{du} \right| \left\| \frac{d\mathbf{r}}{dt} \right\| du = \int_c^d \left| \frac{dt}{du} \right| \|\mathbf{v}(t)\| du = \int_a^b \|\mathbf{v}(t)\| dt.$$

Here one has to be careful using the change of variables formula to make sure that the signs come out right (all quantities must be positive).

In case of the arc length of a graph $y = f(x), a \leq x \leq b$, we get the familiar calculus integral

$$\int_a^b \sqrt{1 + (f'(x))^2} dx.$$

For another example, the arc length of a piece of the unit circle $(\cos\theta, \sin\theta)$, where $\theta \in [\theta_0, \theta_1]$, is just

$$\int_{\theta_0}^{\theta_1} \sqrt{(-\sin\theta)^2 + \cos^2\theta}\, d\theta = \int_{\theta_0}^{\theta_1} 1\, d\theta = \theta_1 - \theta_0.$$

In fact, this is the whole point of "radian measure," in other words that the angle defined by a segment of the circle should be the same as arc length on the circle. More generally, for any unit speed parametrization $\mathbf{r}\colon [a, b] \to \mathbb{R}^n$, the arc length is just $b - a$, and a similar result holds for constant speed parametrizations. For example, for one turn around the helix, we have

$$\int_0^{2\pi} \|\mathbf{v}(t)\| dt = 2\pi\sqrt{2}.$$

In general, however, arc length integrals are notoriously difficult to do.

One natural attempt to parametrize a curve in a natural way is to use arc length itself as the parameter. To do so, we define a function $s = s(t)$ by an integral:

$$s(t) = \int_a^t \|\mathbf{v}\|.$$

By definition, $s(t)$ is the arc length from a to t, and

$$\frac{ds}{dt} = \|\mathbf{v}(t)\| > 0.$$

Thus s is a strictly increasing function C^∞ function of t, and so there is a well-defined inverse $t = t(s)$. By the usual calculus formula,

$$\frac{dt}{ds} = \frac{1}{\|\mathbf{v}(t)\|},$$

and thus if we try to find the derivative of $\mathbf{r} = \mathbf{r}(t(s))$ as a function of s, we find that

$$\frac{d\mathbf{r}}{ds} = \frac{dt}{ds}\frac{d\mathbf{r}}{dt} = \frac{1}{\|\mathbf{v}(t)\|}\mathbf{v}(t).$$

Thus we obtain a unit speed parametrization, as we must since the arc length is just s again. Note that the function t as a function of s is usually the inverse function to an unknown integral, and so can almost never be written down explicitly. However we do have a formula for the derivative, and that is enough for most purposes.

The arc length parametrization of a curve is unique up to the choice of an origin and a direction, in other words an orientation of the curve. It tells us that there is no intrinsic way we can tell whether we are on a straight line or a twisted curve in $\mathbb{R}^n$, except for the issue of whether the curve is open or closed, or finite or infinite in arc length. On the other hand, curves in $\mathbb{R}^n$ do really twist, and we would like to be able to measure this somehow. One way to do so is to use the arc length parametrization as intrinsic and define new quantities from it. For example, given a curve $\mathbf{r}$, we define the unit tangent vector

$$\mathbf{t} = \frac{d\mathbf{r}}{ds} = \frac{1}{\|\mathbf{v}(t)\|}\mathbf{v}(t)$$

and then define the *curvature* κ by the formula

$$\kappa = \left\|\frac{d\mathbf{t}}{ds}\right\| = \left\|\frac{dt}{ds}\frac{d\mathbf{t}}{dt}\right\| = \frac{1}{\|\mathbf{v}(t)\|}\left\|\frac{d\mathbf{t}}{dt}\right\|.$$

Using the formula for $\mathbf{t}$ in terms of $\mathbf{v}$ and a little patience with the chain rule and quotient rule, one can also work out κ in terms of t. For example, a straight line has $\kappa = 0$, the unit circle has $\kappa = 1$, and for a circle of radius r around the origin, $\kappa = 1/r$. (Note that the straight line is the limiting case $r = \infty$.) Conversely, it is not difficult to show that a curve in $\mathbb{R}^2$ with $\kappa = 0$ is a straight line, and a curve with κ a nonzero positive constant is a circle of radius $1/\kappa$. More generally, for a curve in $\mathbb{R}^2$, κ viewed as a function turns out in a suitable sense to be a complete invariant of the congruence class of the curve. The helix satisfies $\kappa(t) = 1/2$. For curves in $\mathbb{R}^3$, there is

the following useful formula:

$$\kappa(t) = \frac{\|\mathbf{v} \times \mathbf{a}\|}{\|\mathbf{v}\|^3}.$$

To see this, note that $\frac{d}{dt}\langle \mathbf{v}, \mathbf{v} \rangle = 2\langle \mathbf{a}, \mathbf{v} \rangle$, and thus

$$\frac{d}{dt}\frac{1}{\|\mathbf{v}\|} = \frac{d}{dt}(\langle \mathbf{v}, \mathbf{v} \rangle)^{-1/2} = -\frac{\langle \mathbf{a}, \mathbf{v} \rangle}{\|\mathbf{v}\|^3}.$$

Using this, we have

$$\begin{aligned}
\kappa &= \frac{1}{\|\mathbf{v}\|}\left\|\frac{d\mathbf{t}}{dt}\right\| = \frac{1}{\|\mathbf{v}\|}\left\|\frac{d}{dt}\frac{1}{\|\mathbf{v}\|}\mathbf{v}\right\| \\
&= \frac{1}{\|\mathbf{v}\|}\left\|\frac{1}{\|\mathbf{v}\|}\mathbf{a} - \frac{\langle \mathbf{a}, \mathbf{v} \rangle}{\|\mathbf{v}\|^3}\mathbf{v}\right\| \\
&= \frac{1}{\|\mathbf{v}\|^3}\|\mathbf{v}\|\left\|\mathbf{a} - \frac{\langle \mathbf{a}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}\mathbf{v}\right\| = \frac{1}{\|\mathbf{v}\|^3}\|\mathbf{v}\| \cdot \|\mathbf{a} - p_{\mathbf{v}}(\mathbf{a})\|.
\end{aligned}$$

But $\|\mathbf{v}\| \cdot \|\mathbf{a} - p_{\mathbf{v}}(\mathbf{a})\|$ is the area of the parallelogram defined by $\mathbf{v}$ and $\mathbf{a}$ and is thus $\|\mathbf{v} \times \mathbf{a}\|$. Putting this together, we get the formula, which is usually easy to work out. For example, the graph of the function $y = f(x)$ has

$$\kappa(x) = \frac{|f''(x)|}{(1 + (f'(x))^2)^{3/2}}.$$

One special property of taking the derivative of the unit tangent vector is that it is always orthogonal to the velocity vector. This is a consequence of the following more general fact, applied to the curve $\mathbf{t}$:

Proposition 11.10. *Let $\mathbf{s}(t)$ be a curve such that $\|\mathbf{s}(t)\|$ is constant, in other words $\mathbf{s}(t)$ is contained in a sphere about the origin. Then $\mathbf{s}'(t)$ is always orthogonal to $\mathbf{s}(t)$.*

Proof. Since $\|\mathbf{s}(t)\|$ is constant, $\|\mathbf{s}(t)\|^2$ is constant as well, and so its derivative is zero. But a calculation (from homework or otherwise) gives

$$\frac{d}{dt}\|\mathbf{s}(t)\|^2 = 2\left\langle \frac{d\mathbf{s}}{dt}, \mathbf{s}(t) \right\rangle,$$

and thus $\mathbf{s}'(t)$ is orthogonal to $\mathbf{s}(t)$. □

If $d\mathbf{t}/ds$ is identically zero, then it is easy to see that $\mathbf{r}$ is a line. In general, we make the assumption that $d\mathbf{t}/ds$ is never zero. Thus, in the case of $\mathbb{R}^2$, the vectors $\mathbf{t}$ and $\mathbf{n} = \frac{1}{\kappa}\frac{d\mathbf{t}}{ds}$ are an orthonormal basis for $\mathbb{R}^2$ at every point of $\mathbf{r}$. Such a basis is called a *moving frame* along $\mathbf{r}$. For $\mathbb{R}^2$, the function κ also describes the congruence class of the curve. For $\mathbb{R}^3$, we only get two orthonormal vectors this way, and need to find a third which is orthogonal to both. We can always find such a third using the cross product by taking $\mathbf{b} = \mathbf{t} \times \mathbf{n}$. The curves $\mathbf{t}$, $\mathbf{n}$, and $\mathbf{b}$ are called the *unit tangent vector*, the *normal*, and the *binormal* respectively. For example, for the helix, $\mathbf{t} = \frac{1}{\sqrt{2}}(-\sin t, \cos t, 1)$, $\mathbf{n} = (-\cos t, -\sin t, 0)$, and so $\mathbf{b} = \frac{1}{\sqrt{2}}(\sin t, -\cos t, 1)$. Aside from the curvature $\kappa(s)$, there is a second function $\tau = \tau(s)$, and the functions κ and τ determine the curve $\mathbf{r}$ up so congruence.

11.4 Surface area

We turn next to surface area for a surface Σ in $\mathbb{R}^n$. First we have the following formula for the area of a parallelogram spanned by two linearly independent vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$: the area is variously given by

$$\sqrt{\|\mathbf{v}\|^2\|\mathbf{w}\|^2 - \langle \mathbf{v}, \mathbf{w}\rangle^2} = \|\mathbf{v} \wedge \mathbf{w}\|$$

and in case $n = 3$ it is also given by $\|\mathbf{v} \times \mathbf{w}\|$.

Now suppose that we have a parametrized surface Σ given by $F\colon D \to \mathbb{R}^n$. Let R be some rectangle containing D and let $P_1 \times P_2$ be a partition of R. We will not worry about what happens near ∂D since this is a set of zero area. The image of a small rectangle R_{ij} contained in D with side lengths Δu and Δv will be approximated by the parallelogram in $\mathbb{R}^n$ spanned by the vectors $\Delta u \mathbf{T}_u$ and $\Delta v \mathbf{T}_v$ (where the tangent vectors $\mathbf{T}_u, \mathbf{T}_v$ are evaluated at the lower left hand corner of R_{ij}). Thus the area of this paralleogram is

$$\sqrt{\|\mathbf{T}_u\|^2\|\mathbf{T}_v\|^2 - \langle \mathbf{T}_u, \mathbf{T}_v\rangle^2}\Delta u \Delta v = \|\mathbf{T}_u \wedge \mathbf{T}_v\|\Delta u \Delta v,$$

so that it is reasonable to define the surface area as

$$\iint_D \sqrt{\|\mathbf{T}_u\|^2\|\mathbf{T}_v\|^2 - \langle \mathbf{T}_u, \mathbf{T}_v\rangle^2}\, du dv = \iint_D \|\mathbf{T}_u \wedge \mathbf{T}_v\|\, du dv.$$

For example, in case Σ lies in $\mathbb{R}^3$, the surface area is defined to be

$$\iint_D \|\mathbf{T}_u \times \mathbf{T}_v\|\, du dv.$$

This quantity is independent of the choice of parametrization, since if u and v are functions of r and s, then

$$\|\mathbf{T}_r \times \mathbf{T}_s\| = \left|\det \frac{\partial(u,v)}{\partial(r,s)}\right| \|\mathbf{T}_u \times \mathbf{T}_v\|.$$

and we can then apply the change of variables formula to the double integral. A similar calculation works in the case of a parametrized surface in $\mathbb{R}^n$.

Example 11.11. The surface area of the unit sphere using the (θ, ϕ) parametrization is given by

$$\iint_{[0,2\pi]\times[0,\pi]} \sin\phi \, d\theta d\phi = \int_0^{2\pi} d\theta \int_0^{\pi} \sin\phi \, d\phi = (2\pi)(2) = 4\pi.$$

If instead we work with the graph parametrization of the upper hemisphere we get

$$\iint_{x^2+y^2\le 1} \frac{1}{z}\, dxdy = \iint_{x^2+y^2\le 1} \frac{dxdy}{\sqrt{1-x^2-y^2}}$$
$$= \int_0^{2\pi}\int_0^1 \frac{r\, drd\theta}{\sqrt{1-r^2}} = 2\pi\left(-\frac{1}{2}\right)\cdot 2(1-r^2)^{1/2}\Big]_0^1 = 2\pi,$$

which is half of the previous answer. Note that the integral is an improper integral, but we still get the correct answer (as we must).

Similar methods can define the k-volume of a parametrized k-submanifold. If $H\colon U \to \mathbb{R}^n$ is a parametrized k-manifold X, where U is an open subset of $\mathbb{R}^k$ with boundary ∂U, we have the function $\|\mathbf{T}_{u_1} \wedge \cdots \wedge \mathbf{T}_{u_k}\|$. If this function extends to a continuous function on $\overline{U}$ (and in some other circumstances as well), we can define the k-volume of X to be

$$\int_U \|\mathbf{T}_{u_1} \wedge \cdots \wedge \mathbf{T}_{u_k}\|\, du_1 \cdots du_k.$$

Just as in the case of curves and surfaces, the general change of variables formula for multiple integrals says that this number is independent of the parametrization, and thus only depends on the manifold X.

Example 11.12. Let us compute the $(n-1)$-volume of the unit sphere

S^{n-1}. We can parametrize S^{n-1} via the analogue of spherical coordinates:

$$\begin{aligned} x_1 &= \cos\phi_1 \sin\phi_2 \sin\phi_3 \cdots \sin\phi_{n-1}; \\ x_2 &= \sin\phi_1 \sin\phi_2 \sin\phi_3 \cdots \sin\phi_{n-1}; \\ x_3 &= \cos\phi_2 \sin\phi_3 \sin\phi_4 \cdots \sin\phi_{n-1}; \\ &\vdots \\ x_{n-1} &= \cos\phi_{n-2} \sin\phi_{n-1}; \\ x_n &= \cos\phi_{n-1}, \end{aligned}$$

for $0 \le \phi_1 \le 2\pi$ and $0 \le \phi_k \le \pi$ for $k \ge 2$. Let $D_n = \det \dfrac{\partial(x_1, \ldots, x_n)}{\partial(r, \phi_1, \ldots, \phi_{n-1})}$, so that as we have seen $D_n = \pm r^{n-1} \sin\phi_2 \sin^2\phi_3 \cdots \sin^{n-2}\phi_{n-1}$. Write $D_n = D_n(r)$ to note the dependence on r. We claim:

$$\|\mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}\| = |D_n(1)| = \sin\phi_2 \sin^2\phi_3 \cdots \sin^{n-2}\phi_{n-1}.$$

It then follows that, if σ_{n-1} is the $(n-1)$-volume of S^{n-1}, then the n-volume v_n of the unit ball in $\mathbb{R}^n$ is

$$v_n = \sigma_n \int_{r=0}^{r=1} r^{n-1}\, dr,$$

and so $\sigma_n = n v_n$. Thus for example the length of S^1 is $2\pi = 2(\pi \cdot 1^2)$ and the surface area of S^2 is $4\pi = 3 \cdot \frac{4}{3}\pi(1)^3$. Note that, by a straightforward scaling argument, the $(n-1)$-volume of an $(n-1)$-sphere of radius r is $\sigma_n r^{n-1}$, so the n-volume of the unit n-ball is the integral of the cross-sectional volumes:

$$v_n = \int_{r=0}^{r=1} r^{n-1} \sigma_n r^{n-1}\, dr = \int_{r=0}^{r=1} r^{n-1} n v_n r^{n-1}\, dr = v_n.$$

However, this principle must be applied with some care.

We shall sketch a proof of the formula that $\|\mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}\| = |D_n(1)|$ that avoids direct computation. Note that

$$\|\mathbf{T}_r \wedge \mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}\| = |\det(\mathbf{T}_r, \mathbf{T}_{\phi_1}, \ldots, \mathbf{T}_{\phi_{n-1}})| = |D_n(r)|.$$

On the other hand,

$$*(\mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}) = \mathbf{T}_{\phi_1} \times \cdots \times \mathbf{T}_{\phi_{n-1}}$$

is the cross product of the $n-1$ vectors $\mathbf{T}_{\phi_1}, \ldots, \mathbf{T}_{\phi_{n-1}}$ and hence is orthogonal to all of them and hence to the tangent space of S^{n-1} at every point

$\mathbf{x} = (x_1, \ldots, x_n)$. Since the tangent space to S^{n-1} at $\mathbf{x}$ is $\{\mathbf{x}\}^\perp$, it follows that

$$\mathbf{T}_{\phi_1} \times \cdots \times \mathbf{T}_{\phi_{n-1}} = f(\mathbf{x})\mathbf{x}$$

for some function f, and it is easy to see that

$$\|\mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}\| = \|\mathbf{T}_{\phi_1} \times \cdots \times \mathbf{T}_{\phi_{n-1}}\| = |f(\mathbf{x})|.$$

Moreover, by properties of the $*$-operator,

$$\mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}} = \pm * (\mathbf{T}_{\phi_1} \times \cdots \times \mathbf{T}_{\phi_{n-1}}) = \pm f(\mathbf{x}) * \mathbf{x}.$$

On the other hand, for $\mathbf{x} \in S^{n-1}$, it follows by inspection that $\mathbf{T}_r = \mathbf{x}$. Thus, for $\mathbf{x} \in S^{n-1}$,

$$|D_n(1)| = |D_n(r)| = \|\mathbf{T}_r \wedge \mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}\| = \|\mathbf{x} \wedge (\pm f(\mathbf{x}) * \mathbf{x})\| = |f(\mathbf{x})| \|\mathbf{x} \wedge * \mathbf{x}\|.$$

Since $\mathbf{x}$ has unit length, $\|\mathbf{x} \wedge * \mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = 1$, so that

$$|D_n(1)| = |f(\mathbf{x})| = \|\mathbf{T}_{\phi_1} \wedge \cdots \wedge \mathbf{T}_{\phi_{n-1}}\|,$$

which finishes the proof of the formula.

Lastly, we want to say a few words about the problem of describing how a surface curves. This problem has two aspects. First, given a surface Σ in, say, $\mathbb{R}^3$, we could try to describe how the surface is curved in $\mathbb{R}^3$ by analogy with the definitions of curvature which we gave for curves. There is also the problem of intrinsic curvature. Intuitively, given a parametrization $H \colon U \to \Sigma$, we want to describe how much distortion H introduces in terms of the measurement of length, angles, and area. In more practical terms, if Σ is a sphere, we would like to understand what kind of maps can be made of the globe. Of course, since a 2-sphere is compact, there is no homeomorphism from a 2-sphere to $\mathbb{R}^2$ or to an open subset of $\mathbb{R}^2$. But we could look at small pieces of the 2-sphere and ask if it is possible to parametrize these via **some** function H defined on an open subset of $\mathbb{R}^2$ such that measurements of length, angles, and area correspond under H. In this case, H is called an *isometry*. In other words, can we make a perfect map of a small piece of the globe? Gauss showed in 1827 that the answer was no; he may have been motivated by his experience in surveying northern Germany. Gauss showed that this was not possible by defining an intrinsic curvature K on every surface Σ in $\mathbb{R}^3$ and showing i) this curvature is equal to $1/R^2$ on the 2-sphere of radius R and ii) it is identically zero on every

parametrized surface $H\colon U \to \Sigma$, where H is an isometry from the open subset U of $\mathbb{R}^2$ to $\Sigma \subset \mathbb{R}^3$. This result (and generalizations of it) are called the *Theorema Egregium*, from the Latin words which Gauss used to describe it (remarkable theorem).

Let us explain some of the ideas behind this argument. For a manifold X, we can define the analogue of line segments, namely curves C lying on X with endpoints $\mathbf{p}$, $\mathbf{q}$ which are the paths of shortest distance with endpoints $\mathbf{p}$ and $\mathbf{q}$. Such curves are called *geodesics*. For example, if $X = S^2$, the geodesics are segments of great circles (circles on S^2 whose center in $\mathbb{R}^3$ is the origin). One can show (but this is not easy) that, if $\mathbf{p}$ and $\mathbf{q}$ are close, then there is a unique geodesic with endpoints $\mathbf{p}$ and $\mathbf{q}$. Instead of taking this approach, one can try a simple approach: given a parametrization $H\colon U \to \Sigma$ and a curve C in U, we can try to compare the length of C, viewed as a curve in U, with the length of the image curve $H(C)$ in Σ, and write down the condition on H for these two curves to have the same length. In U, C is given as a parametrized curve, say, by $\mathbf{r}(t) = (u(t), v(t))$, $a \leq t \leq b$, and its length is just

$$\int_a^b \sqrt{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}\, dt.$$

We abbreviate this by saying the length is $\displaystyle\int_a^b ds$, where

$$\left(\frac{ds}{dt}\right)^2 = \left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2.$$

Turning to the image curve $H \circ \mathbf{r}$ on Σ, it is given by

$$H \circ \mathbf{r} = (x(u(t), v(t)), y(u(t), v(t)), z(u(t), v(t))).$$

Then the length of $H(C)$ is given by

$$\int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt.$$

By the chain rule,

$$\frac{d}{dt} H \circ \mathbf{r} = \frac{du}{dt}\mathbf{T}_u + \frac{dv}{dt}\mathbf{T}_v.$$

Hence

$$\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 = \left\langle \frac{du}{dt}\mathbf{T}_u + \frac{dv}{dt}\mathbf{T}_v, \frac{du}{dt}\mathbf{T}_u + \frac{dv}{dt}\mathbf{T}_v \right\rangle$$
$$= \|\mathbf{T}_u\|^2 \left(\frac{du}{dt}\right)^2 + 2\left\langle \mathbf{T}_u, \mathbf{T}_v \right\rangle \left(\frac{du}{dt}\right)\left(\frac{dv}{dt}\right) + \|\mathbf{T}_v\|^2 \left(\frac{dv}{dt}\right)^2,$$
$$= E\left(\frac{du}{dt}\right)^2 + 2F\left(\frac{du}{dt}\right)\left(\frac{dv}{dt}\right) + G\left(\frac{dv}{dt}\right)^2,$$

say, where $E = \|\mathbf{T}_u\|^2$, $F = \langle \mathbf{T}_u, \mathbf{T}_v \rangle$, and $G = \|\mathbf{T}_v\|^2$. One way to think of this is as follows: for every point of U, we have a positive definite bilinear form $B(x_1, x_2) = Ex_1^2 + 2Fx_1x_2 + Gx_2^2$, whose coefficients are functions of u, v. To compute arc length for the image curve $H(C)$, we look at

$$\int_a^b \sqrt{B(\mathbf{v}, \mathbf{v})}\, dt,$$

where $\mathbf{v}$ is the tangent vector to $\mathbf{r}(t)$ in U. The condition that H is an isometry is then the condition that $E = G = 1$ and $F = 0$, i.e. that $\mathbf{T}_u$ and $\mathbf{T}_v$ are an orthonormal basis at every point, so that the length of C and $H(C)$ are the same for every curve C. For example, it is clear (by cutting it open) that one can make a map of the cylinder $x^2 + y^2 = 1$, z arbitrary, which is parametrized by the coordinates θ and z ($x = \cos\theta$, $y = \sin\theta$, $z = z$). Direct computation shows in this example that $\mathbf{T}_\theta$ and $\mathbf{T}_z$ are in fact an orthonormal basis at every point.

For any given parametrization, the quantities E, F, G are straightforward (if messy) to compute. However, just because we can find one parametrization H such that H is not an isometry does not answer the question of whether there exists **some** parametrization which is an isometry. What Gauss showed is that, if Σ is a surface in $\mathbb{R}^3$, then there exists a function $K\colon \Sigma \to R$ such that K is identically $0 \iff$ for every point $\mathbf{x} \in \Sigma$, there exists an isometry from an open subset U of $\mathbb{R}^2$ to $B \cap \Sigma$, where B is a small ball in $\mathbb{R}^3$ centered at $\mathbf{x}$, and that, if Σ is a sphere of radius R (or a piece of a sphere), then K is constant and equal to $1/R^2$. Hence there is no isometry from an open subset of $\mathbb{R}^2$ to any piece of a sphere.

The study of measurement and curvature on a submanifold of $\mathbb{R}^n$, and more generally, is called *differential geometry*. Aside from its natural intuitive appeal in answering questions like those above, differential geometry is the natural language for general relativity, where light travels along geodesic curves, and gravity is caused by the curvature of space-time due to matter.

11.5 Line integrals

Next we want to describe more general kinds of objects that can be integrated over manifolds. As usual, we begin with the case of curves. If $\mathbf{r}\colon [a,b] \to \mathbb{R}^n$ is a parametrized curve, which we will also denote by the letter γ, and f is a (continuous) scalar function on $\mathbb{R}^n$, then we can define $\displaystyle\int_\gamma f$ by taking $\displaystyle\int_a^b f(\mathbf{r}(t))\|\mathbf{v}(t)\|dt$. Just as in the case of arc length, it is easy to see that this is independent of parametrization. Also, we only need f to be defined and continuous on the image γ of $\mathbf{r}$. Sometimes this integral is written $\displaystyle\int_\gamma f\,ds$, to emphasize that we are integrating with respect to arc length. These integrals have an interpretation in terms of "adding up" the values of f along γ; for example, f might be the density function of γ viewed as a thin wire and then the integral would be the total density.

However, these kind of integrals will not be our main interest, and instead we shall define how to integrate a *vector field* $\mathbf{F}$ along a curve $\mathbf{r}$. The resulting integral $\displaystyle\int_{\mathbf{r}} \mathbf{F}$ is called the *line integral* or sometimes the *path integral* of the vector field $\mathbf{F}$ along $\mathbf{r}$. It has the following physical interpretation: suppose that $\mathbf{r}$ is the path of a particle moving in $\mathbb{R}^n$ and that $\mathbf{F}$ is a force field in $\mathbb{R}^n$. We want to measure the amount of work the force does in moving the particle from $\mathbf{r}(a)$ to $\mathbf{r}(b)$. Clearly only the component along the curve is relevant, and so we should take

$$\int_{\mathbf{r}} \mathbf{F} = \int_a^b \langle \mathbf{F}, \mathbf{t}\rangle ds,$$

where $\mathbf{t}$ is the unit tangent vector. In physics, this integral is defined as the *work*. At first sight this integral looks quite complicated, since $ds = \|\mathbf{v}(t)\|dt$ and $\mathbf{t} = \dfrac{1}{\|\mathbf{v}(t)\|}\mathbf{v}(t)$. However, plugging in gives

$$\int_a^b \langle \mathbf{F}, \mathbf{t}\rangle ds = \int_a^b \left\langle \mathbf{F}, \frac{1}{\|\mathbf{v}(t)\|}\mathbf{v}(t)\right\rangle \|\mathbf{v}(t)\|dt = \int_a^b \langle \mathbf{F}, \mathbf{v}(t)\rangle dt,$$

and this is usually fairly easy to compute, and can always be computed if the components of $\mathbf{F}$ and $\mathbf{r}$ are polynomial functions. In particular, if $\mathbf{F} = (F_1, \ldots, F_n)$ and $\mathbf{r}(t) = (x_1(t), \ldots, x_n(t))$, $0 \le t \le a$, then

$$\int_{\mathbf{r}} \mathbf{F} = \int_a^b \left(F_1 \frac{dx_1}{dt} + \cdots + F_n \frac{dx_n}{dt}\right) dt.$$

For example, if $\mathbf{F} = (x^2y, yz, xz)$ and $\mathbf{r}(t) = (1, t^2, t)$, $t \in [0,1]$, then $\mathbf{v}(t) = (0, 2t, 1)$ and

$$\int_{\mathbf{r}} \mathbf{F} = \int_0^1 (0 + (t^2)(t)(2t) + (1)(t)(1))dt = \int_0^1 (2t^4 + t)dt = \frac{2}{5} + \frac{1}{2} = \frac{9}{10}.$$

The line integral is almost, but not quite, independent of the parametrization. Let $t = t(u)$ be a reparametrization of the curve $\mathbf{r}$, where $u \in [c,d]$. If t is an increasing function of u, then $t(c) = a$ and $t(d) = b$, and working out the line integral in the u parametrization gives

$$\int_c^d \left\langle \mathbf{F}, \frac{d\mathbf{r}}{du} \right\rangle du = \int_c^d \left\langle \mathbf{F}, \frac{d\mathbf{r}}{dt} \right\rangle \frac{dt}{du} du = \int_{t(c)}^{t(d)} \left\langle \mathbf{F}, \frac{d\mathbf{r}}{dt} \right\rangle dt = \int_a^b \left\langle \mathbf{F}, \frac{d\mathbf{r}}{dt} \right\rangle dt,$$

which is the line integral computed in the t-parametrization. But if t is a decreasing function of u, then $t(c) = b$ and $t(d) = a$, and the same calculation gives

$$\int_c^d \left\langle \mathbf{F}, \frac{d\mathbf{r}}{du} \right\rangle du = \int_b^a \left\langle \mathbf{F}, \frac{d\mathbf{r}}{dt} \right\rangle dt = -\int_a^b \left\langle \mathbf{F}, \frac{d\mathbf{r}}{dt} \right\rangle dt.$$

In other words the integral has changed sign, which is plausible when we consider the physical meaning of the line integral. We say that the line integral is a *oriented integral*, meaning that it depends not just on the geometric curve $\mathbf{r}$ but also on the choice of a direction. If we symbolically write $-\mathbf{r}$ for the oriented curve which is $\mathbf{r}$ taken in the opposite direction, then we have the formula

$$\int_{-\mathbf{r}} \mathbf{F} = -\int_{\mathbf{r}} \mathbf{F}.$$

We can also define the line integral on a curve $\mathbf{r}$ which is not necessarily C^∞, or C^1, but can be written as a union of curves $\mathbf{r}_1 \cup \mathbf{r}_2 \cup \cdots \cup \mathbf{r}_k$. Here each $\mathbf{r}_i$ is a C^∞ or C^1 curve defined on an interval $[a_i, b_i]$, and $\mathbf{r}_i(b_i) = \mathbf{r}_{i+1}(a_i)$, so that the endpoint of one curve is the starting point of the next. We call such a curve a *piecewise* C^∞ curve. We also write $\mathbf{r} = \mathbf{r}_1 + \mathbf{r}_2 + \cdots + \mathbf{r}_k$ and set

$$\int_{\mathbf{r}} \mathbf{F} = \int_{\mathbf{r}_1} \mathbf{F} + \cdots + \int_{\mathbf{r}_k} \mathbf{F}.$$

It is easy to see by basic properties of the integral that this formula is also true if $\mathbf{R}$ is a C^∞ or C^1 curve and we split it into a union of curves by taking a partition of the domain $[a,b]$.

In physics, if $\mathbf{r}(t)$ is the trajectory of a particle moving under a force $\mathbf{F}$, then $\mathbf{r}$ and $\mathbf{F}$ are related by Newton's second law: $\mathbf{F} = m\mathbf{a}$, where m is the

mass of the particle and $\mathbf{a}$ is the acceleration. In this case, the work, i.e. the line integral, becomes

$$\int_{\mathbf{r}} \mathbf{F} = \int_a^b m\langle \mathbf{a}, \mathbf{v}\rangle dt.$$

Now $\mathbf{a} = d\mathbf{v}/dt$, and thus (by a previous formula)

$$2\langle \mathbf{a}, \mathbf{v}\rangle = \frac{d}{dt}\langle \mathbf{v}, \mathbf{v}\rangle = \frac{d}{dt}\|\mathbf{v}\|^2.$$

Applying the fundamental theorem of calculus, we see that the work is equal to

$$\frac{1}{2}m\|\mathbf{v}(b)\|^2 - \frac{1}{2}m\|\mathbf{v}(a)\|^2.$$

In physics, the quantity $\frac{1}{2}m\|\mathbf{v}\|^2$ is called the *kinetic energy*, and the above formula says that work (for a particle moving under Newton's second law) is equal to the change in kinetic energy.

Among all vector fields, there are special ones, namely those which are gradient vector fields. Another kind of special vector field is a *conservative* vector field:

Definition 11.13. A vector field $\mathbf{F}$ is *conservative* if $\int_{\mathbf{r}} \mathbf{F}$ only depends on the endpoints of $\mathbf{r}$. In other words, $\mathbf{F}$ is conservative if, for every two piecewise C^1 curves $\mathbf{r}_1$ and $\mathbf{r}_2$ with the same endpoints (as oriented curves), $\int_{\mathbf{r}_1} \mathbf{F} = \int_{\mathbf{r}_2} \mathbf{F}$.

Proposition 11.14 (Fundamental theorem for line integrals)**.** *If $\mathbf{F}$ is a gradient vector field ∇f, then $\mathbf{F}$ is conservative. In fact, for every piecewise C^1 curve $\mathbf{r}$,*

$$\int_{\mathbf{r}} \nabla f = f(\mathbf{r}(b)) - f(\mathbf{r}(a)).$$

Proof. First assume that $\mathbf{r}$ is actually C^1. In this case, the proof is just the fundamental theorem of calculus and the chain rule: if

$$\mathbf{F} = \nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right),$$

and $\mathbf{r}(t) = (x_1(t), \ldots, x_n(t)$, then by definition

$$\int_{\mathbf{r}} \nabla f = \int_a^b \left(\frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt}\right) dt.$$

But, by the chain rule, the last integrand is just $\frac{d}{dt}f(\mathbf{r}(t))$, and so by the fundamental theorem of calculus the integral is

$$\int_a^b \frac{d}{dt}f(\mathbf{r}(t))dt = f(\mathbf{r}(b)) - f(\mathbf{r}(a)),$$

as claimed.

If $\mathbf{r}$ is only piecewise C^1, the we write it as a sum of C^1 curves, apply the above, and note that the endpoints cancel off against each other except the last and (with a minus sign) the first. □

As a consequence of the above, we have another proof of the following:

Corollary 11.15. *Let U be a connected open set. If $\nabla f = \nabla g$, then $f = g + c$ for some constant c.*

Proof. Let $h = f - g$, so that $\nabla h = \mathbf{0}$. We must show that $h = c$ for some constant c. But a connected open set in $\mathbb{R}^n$ is path connected—in fact, every two points can be connected by a piecewise smooth curve which is a union of lines. Given $\mathbf{x}, \mathbf{y} \in U$, let $\mathbf{r}$ be a piecewise smooth curve connecting them. By the proposition, $0 = \int_{\mathbf{r}} \nabla h = h(\mathbf{y}) - h(\mathbf{x})$. Thus $h(\mathbf{y}) = h(\mathbf{x})$ for every two points $\mathbf{x}, \mathbf{y} \in U$, and so h is constant. □

In physics, a function V such that $\nabla V = -\mathbf{F}$ is called a *potential* for $\mathbf{F}$ (note the minus sign!). The formula then reads: if $\mathbf{F}$ has a potential, then the work is equal to minus the difference in potential energy. If in addition $\mathbf{F} = m\mathbf{a}$ as in our previous discussion, then the change in kinetic energy is equal to minus the change in potential energy. Hence, if we define the total energy to be the sum of kinetic energy plus potential energy, then total energy is constant (under our assumptions). Thus energy is *conserved*, which is the reason for the name conservative.

Returning to the study of conservative vector fields, we have:

Proposition 11.16. *Let $\mathbf{F} = (F_1, \ldots, F_n)$ be a vector field on $\mathbb{R}^n$. The following are equivalent:*

(i) $\mathbf{F}$ *is a gradient vector field.*

(ii) $\mathbf{F}$ *is conservative.*

(iii) *For every closed curve $\mathbf{r}$, $\int_{\mathbf{r}} \mathbf{F} = 0$.*

Proof. First we show that (ii) and (iii) are equivalent. If $\mathbf{F}$ is conservative, then $\displaystyle\int_{\mathbf{r}} \mathbf{F} = 0$ for every closed curve $\mathbf{r}$. Conversely, suppose that $\displaystyle\int_{\mathbf{r}} \mathbf{F} = 0$ for every closed curve $\mathbf{r}$. If $\mathbf{r}_1$ and $\mathbf{r}_2$ are two curves with the same endpoints (as oriented curves), then $\mathbf{r}_1 + (-\mathbf{r}_2)$ is a piecewise smooth closed curve, and thus $\displaystyle\int_{\mathbf{r}_1+(-\mathbf{r}_2)} \mathbf{F} = 0$. But then

$$0 = \int_{\mathbf{r}_1+(-\mathbf{r}_2)} \mathbf{F} = \int_{\mathbf{r}_1} \mathbf{F} - \int_{\mathbf{r}_2} \mathbf{F},$$

so that

$$\int_{\mathbf{r}_1} \mathbf{F} = \int_{\mathbf{r}_2} \mathbf{F}$$

as claimed.

Now we have already seen that (i) $\implies$ (ii). So we will be done if we show that (ii) $\implies$ (i). To see this, we need to show that, given a conservative vector field $\mathbf{F}$, we can find a function f with $\nabla f = \mathbf{F}$. The natural way to try to find such a function is via integration: fix a base point $\mathbf{p}_0$, and define

$$f(\mathbf{x}) = \int_{\mathbf{p}_0}^{\mathbf{x}} \mathbf{F}.$$

Here the notation means that we should choose some oriented piecewise C^∞ curve $\mathbf{r}$ starting at $\mathbf{p}_0$ and ending at $\mathbf{x}$, and define $\displaystyle\int_{\mathbf{p}_0}^{\mathbf{x}} \mathbf{F} = \int_{\mathbf{r}} \mathbf{F}$. By hypothesis the answer is independent of the choice of $\mathbf{r}$. To finish the proof we have to show that $\nabla f = \mathbf{F}$. For example, let us show that $\dfrac{\partial f}{\partial x_1} = \mathbf{F}_1$, with the arguments for the other components similar. It is enough to compute

$$\lim_{h\to 0} \frac{1}{h} \left(\int_{\mathbf{p}_0}^{\mathbf{x}+h\mathbf{e}_1} \mathbf{F} - \int_{\mathbf{p}_0}^{\mathbf{x}} \mathbf{F} \right) = \lim_{h\to 0} \frac{1}{h} \left(\int_{\mathbf{x}}^{\mathbf{x}+h\mathbf{e}_1} \mathbf{F} \right).$$

Here the last equality follows since we can always assume that the path from $\mathbf{p}_0$ to $\mathbf{x} + h\mathbf{e}_1$ is of the form $C_1 + C_2$, where C_1 is a path from $\mathbf{p}_0$ to $\mathbf{x}$ and C_2 is a path from $\mathbf{x}$ to $\mathbf{x} + h\mathbf{e}_1$, and then use

$$\int_{C_1+C_2} \mathbf{F} = \int_{C_1} \mathbf{F} + \int_{C_1} \mathbf{F}.$$

Now a natural choice of path from $\mathbf{x}$ to $\mathbf{x}+h\mathbf{e}_1$ is a straight line parametrized via $\mathbf{r}(t) = \mathbf{x} + t\mathbf{e}_1$, $0 \le t \le h$. For this choice the velocity vector is $\mathbf{e}_1$ and

the line integral is given by

$$\int_{\mathbf{x}}^{\mathbf{x}+h\mathbf{e}_1} \mathbf{F} = \int_0^h \langle \mathbf{F}(\mathbf{x}+t\mathbf{e}_1), \mathbf{e}_1 \rangle \, dt = \int_0^h F_1(\mathbf{x}+t\mathbf{e}_1) \, dt.$$

By the fundamental theorem of calculus, if g is any continuous function defined on an interval $[0, a]$, and $G(u) = \int_0^u g(t)\,dt$, then $G(0) = 0$ and

$$\lim_{h\to 0} \frac{1}{h} \left(\int_0^h g(t)\,dt \right) = \lim_{h\to 0} \frac{G(h) - G(0)}{h - 0} = G'(0) = g(0).$$

Thus, in our situation,

$$\lim_{h\to 0} \frac{1}{h} \left(\int_0^h F_1(\mathbf{x}+t\mathbf{e}_1) dt \right) = F_1(\mathbf{x}+t\mathbf{e}_1)\Big|_{t=0} = F_1(\mathbf{x}),$$

and so $\dfrac{\partial f}{\partial x_1} = \mathbf{F}_1$ as claimed. □

Note that the conditions of the proposition on when a vector field is a gradient vector field are not easily checkable, since it is not possible to integrate over all possible curves or all possible closed curves. What can be checked is the *mixed partials condition*: as we have seen, if $\mathbf{F}$ is a C^1 vector field and if $\mathbf{F} = \nabla f$ for some function f, then by the equality of mixed partials,

$$\frac{\partial F_i}{\partial x_j} = \frac{\partial F_j}{\partial x_i}, \qquad \text{for all } i \neq j.$$

By the symmetry of the conditions, we may assume that $i < j$, so that the mixed partials conditions involves a total of $n(n-1)/2$ different equations. We know that the mixed partials condition is a necessary condition for $\mathbf{F}$ to be a gradient vector field. In other words, if the mixed partials condition is not satisfied, then $\mathbf{F}$ **cannot** be a gradient vector field. When is the mixed partials condition a sufficient condition? The answer is rather subtle. First, let us note that the mixed partials condition is necessary and sufficient in case the vector field is defined in a **convex** open set:

Theorem 11.17. *Let $\mathbf{F} = (F_1, \ldots, F_n)$ be a C^1 vector field defined on a convex open set $U \subseteq \mathbb{R}^n$. Then $\mathbf{F}$ is a gradient vector field $\iff$ $\mathbf{F}$ satisfies the mixed partials condition.*

Proof. We have seen that, if $\mathbf{F}$ is a gradient vector field, then $\mathbf{F}$ satisfies the mixed partials condition (where $\mathbf{F}$ can be defined in an arbitrary open subset of $\mathbb{R}^n$). Conversely, suppose that $\mathbf{F}$ is defined on the convex open set U and that $\mathbf{F}$ satisfies the mixed partials condition. We seek a function f such that $\nabla f = \mathbf{F}$, or equivalently such that $\dfrac{\partial f}{\partial x_i} = F_i$ for every i. We may assume, after a translation (which does not affect convexity or the mixed partials condition) that U is a convex set containing $\mathbf{0}$. Hence, for all $\mathbf{x} \in U$, $t\mathbf{x} \in U$ for $0 \leq t \leq 1$. Define

$$f(\mathbf{x}) = \sum_{j=1}^{n} x_j \int_0^1 F_j(t\mathbf{x})\, dt.$$

We compute $\partial f/\partial x_i$:

$$\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} x_j \int_0^1 F_j(t\mathbf{x})\, dt\right) = \sum_{j=1}^{n} x_j \frac{\partial}{\partial x_i}\int_0^1 F_j(t\mathbf{x})\, dt + \int_0^1 F_i(t\mathbf{x})\, dt,$$

by the product rule. Now we will use the fact that we can "differentiate under the integral sign" to conclude that

$$\frac{\partial}{\partial x_i}\int_0^1 F_j(t\mathbf{x})\, dt = \int_0^1 \frac{\partial}{\partial x_i}[F_j(t\mathbf{x})]\, dt = \int_0^1 t\frac{\partial F_j}{\partial x_i}(t\mathbf{x})\, dt,$$

by an application of the (one-variable) chain rule, since

$$\frac{\partial}{\partial x_i}[F_j(t\mathbf{x})] = \frac{\partial F_j(t\mathbf{x})}{\partial x_i}\frac{\partial (tx_i)}{\partial x_i} = t\frac{\partial F_j}{\partial x_i}(t\mathbf{x}).$$

By the mixed partials condition, $\partial F_j/\partial x_i = \partial F_i/\partial x_j$, and so since x_j is a constant as far as the t-integration is concerned, we can rewrite the sum on the right hand side of the above equation as

$$\sum_{j=1}^{n} x_j \frac{\partial}{\partial x_i}\int_0^1 F_j(t\mathbf{x})\, dt = \int_0^1 \sum_{j=1}^{n} tx_j \frac{\partial F_j}{\partial x_i}(t\mathbf{x})\, dt = \int_0^1 \sum_{j=1}^{n} tx_j \frac{\partial F_i}{\partial x_j}(t\mathbf{x})\, dt.$$

By the chain rule,

$$\sum_{j=1}^{n} x_j \frac{\partial F_i}{\partial x_j}(t\mathbf{x}) = \frac{d}{dt}F_i(t\mathbf{x}).$$

Thus,

$$\int_0^1 \sum_{j=1}^{n} tx_j \frac{\partial F_i}{\partial x_j}(t\mathbf{x})\, dt = \int_0^1 t\sum_{j=1}^{n} x_j \frac{\partial F_i}{\partial x_j}(t\mathbf{x})\, dt = \int_0^1 t\frac{d}{dt}F_i(t\mathbf{x})\, dt.$$

We can now apply integration by parts:

$$\int_0^1 t\frac{d}{dt}F_i(t\mathbf{x})\,dt = tF_i(t\mathbf{x})]_{t=0}^{t=1} - \int_0^1 F_i(t\mathbf{x})\,dt = F_i(\mathbf{x}) - \int_0^1 F_i(t\mathbf{x})\,dt.$$

Plugging this all back into the formula for $\partial f/\partial x_i$, we see that

$$\frac{\partial f}{\partial x_i} = F_i(\mathbf{x}) - \int_0^1 F_i(t\mathbf{x})\,dt + \int_0^1 F_i(t\mathbf{x})\,dt = F_i(\mathbf{x}),$$

and hence $\nabla f = \mathbf{F}$ as desired. □

Example 11.18. Let $\mathbf{F}$ be the vector field in $\mathbb{R}^3$ defined by

$$\mathbf{F} = (e^y + 2xz^2, xe^y + 3y^2z, 2x^2z + y^3 + z^4) = (F_1, F_2, F_3),$$

say. Checking the mixed partials condition, we find that

$$\begin{aligned}\frac{\partial F_1}{\partial y} &= e^y = \frac{\partial F_2}{\partial x};\\ \frac{\partial F_1}{\partial z} &= 4xz = \frac{\partial F_3}{\partial x};\\ \frac{\partial F_2}{\partial z} &= 3y^2 = \frac{\partial F_3}{\partial y}.\end{aligned}$$

Hence $\mathbf{F}$ satisfies the mixed partials condition, and since $\mathbb{R}^3$ is convex, we are guaranteed that there is a function f such that $\nabla f = \mathbf{F}$. To find f, since $\partial f/\partial x = F_1$, take the x-antiderivative of F_1 to see that

$$f(x, y, z) = xe^y + x^2z^2 + h_1(y, z),$$

where $h_1(y, z)$ is some function of y and z alone, not involving x. Taking the partial with respect to y, we find that

$$xe^y + 3y^2z = F_2 = \frac{\partial f}{\partial y} = xe^y + \frac{\partial h_1}{\partial y},$$

so that $\partial h_1/\partial y = 3y^2z$. Taking the y-antiderivative of the function $3y^2z$, we find that $h_1(y, z) = y^3z + h_2(z)$, where $h_2(z)$ is some function of z alone, not involving x or y. Thus $f(x, y, z) = xe^y + x^2z^2 + y^3z + h_2(z)$. Taking the partial with respect to z gives

$$2x^2z + y^3 + z^4 = F_3 = \frac{\partial f}{\partial z} = 2x^2z + y^3 + \frac{dh_2}{dz},$$

so that $dh_2/dz = z^4$ and $h_2 = z^5/5 + C$, where C is a constant. It follows that all possible f such that $\nabla f = \mathbf{F}$ are given by

$$f(x, y, z) = xe^y + x^2z^2 + y^3z + z^5/5 + C,$$

where C is an arbitrary constant, and if we only care about finding one particular f we can take $C = 0$.

Example 11.19. To see that some hypothesis such as convexity is necessary, consider the vector field $\mathbf{F}$ defined in $\mathbb{R}^2 - \{\mathbf{0}\}$ by:

$$\mathbf{F} = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right).$$

A calculation shows that

$$\frac{\partial}{\partial x}\left(\frac{x}{x^2+y^2}\right) = \frac{x^2+y^2-2x^2}{(x^2+y^2)^2} = \frac{-x^2+y^2}{(x^2+y^2)^2};$$
$$\frac{\partial}{\partial y}\left(\frac{-y}{x^2+y^2}\right) = \frac{-(x^2+y^2)+2y^2}{(x^2+y^2)^2} = \frac{-x^2+y^2}{(x^2+y^2)^2}.$$

Hence $\mathbf{F}$ satisfies the mixed partials condition. But, if C is the closed curve defined by the unit circle, i.e. $\mathbf{r}(t) = (\cos t, \sin t)$, then

$$\int_C \mathbf{F} = \int_0^{2\pi} \langle \mathbf{F}(\cos t, \sin t), (-\sin t, \cos t)\rangle \, dt = \int_0^{2\pi} \langle (-\sin t, \cos t), (-\sin t, \cos t)\rangle \, dt$$
$$= \int_0^{2\pi} (\sin^2 t + \cos^2 t)\, dt = \int_0^{2\pi} 1 \, dt = 2\pi \neq 0.$$

Thus, by Part (iii) of Proposition 11.16, $\mathbf{F}$ cannot be a gradient vector field.

It is instructive to try and apply the procedure of the preceding example here. With a little patience, you will find that taking x and y antiderivatives leads to the formula

$$\mathbf{F} = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right) = \nabla \tan^{-1}(y/x).$$

The problem with this formula is that $\tan^{-1}(y/x)$ fails to be defined along the entire y-axis, not just at the origin. So $\mathbf{F}$ is a gradient vector field in $\mathbb{R}^2-$ the y-axis, but as we have seen it is not a gradient vector field in $\mathbb{R}^2 - \{\mathbf{0}\}$. (Note that the closed curve $C = S^1$ does not live entirely in $\mathbb{R}^2-$ the y-axis.) It is perhaps more enlightening to write $\mathbf{F} = \nabla\theta$, where θ is the coordinate in polar coordinates; here θ cannot be defined as a continuous (much less

C^1) function on $\mathbb{R}^2 - \{\mathbf{0}\}$, because of the problems at 0 and 2π, although it can be defined in the open set $\mathbb{R}^2 - ([0,\infty) \times \{0\})$, the complement of the nonnegative x-axis. Since the ambiguity in the values of θ is just a constant (an integer multiple of 2π), it makes perfect sense to take the gradient $\nabla\theta$ of θ which is now well-defined. We can then interpret the fundamental theorem for line integrals as saying that

$$\int_{S^1} \mathbf{F} = \int_{S^1} \nabla\theta = \theta_+(1,0) - \theta_-(1,0) = 2\pi - 0 = 2\pi,$$

where $\theta_+(1,0) = 2\pi$ is the limit of the angle as we approach $(1,0)$ from below the x-axis and $\theta_-(1,0) = 0$ is the limit of the angle as we approach $(1,0)$ from above the x-axis. Pursuing this idea, one can show that, if C is any piecewise smooth closed curve, then $\displaystyle\int_C \mathbf{F}$ is equal to $2n\pi$, where n is always an integer (positive or negative), and n counts how many times C wraps around the origin.

A final comment about this example. The problem arises because $\mathbb{R}^2 - \{\mathbf{0}\}$ is not convex, but more seriously because $\mathbb{R}^2 - \{\mathbf{0}\}$ has a "hole" where we removed the origin. It turns out, though, that the problem is connected with the **type** of hole that we introduced. For example, if we make the same construction in higher dimensions, by considering $\mathbb{R}^n - \{\mathbf{0}\}$, then, for $n \geq 3$, every C^1 vector field $\mathbf{F}$ in $\mathbb{R}^n - \{\mathbf{0}\}$ which satisfies the mixed partials condition is in fact a gradient vector field.

We conclude this section with *differential forms notation* for vector fields, which will be important in the next section. Instead of writing $\mathbf{F}$ in components, we write it as $F_1 dx_1 + \cdots + F_n dx_n$, and call the resulting object a 1-*form*. Of course, this notation contains exactly the same information as the vector field $\mathbf{F}$. Often, 1-forms will be denoted by ω (the Greek letter omega), or some other Greek letter. To define the integral $\displaystyle\int_\gamma \omega$, write $\omega = F_1 dx_1 + \cdots + F_n dx_n$ and represent γ by a parametrized curve $\mathbf{r}(t) = (x_1(t), \ldots, x_n(t))$. In order to convert ω into something we can integrate, we evaluate F_i along $\mathbf{r}(t)$, and replace dx_i by $x_i'(t)dt$. Collecting terms, we write

$$\int_\gamma \omega = \int_a^b \left(F_1 \frac{dx_1}{dt} + \cdots + F_n \frac{dx_n}{dt} \right) dt,$$

so that this last integral has meaning and can be evaluated. Sometimes we write

$$\left(F_1 \frac{dx_1}{dt} + \cdots + F_n \frac{dx_n}{dt} \right) dt = \mathbf{r}^* \omega,$$

and think that we have used the parametrized curve $\mathbf{r}$ to convert the formal symbol ω into something of the form $g(t)dt$, which we know how to integrate. Of course, the final answer is the same as the line integral of $\mathbf{F} = (F_1, \ldots, F_n)$ over $\mathbf{r}$. In this notation, we write the gradient in the form

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n,$$

which looks like the standard linear approximation with the calculus convention that "in the limit Δx becomes dx." Then the fundamental theorem for line integrals becomes

$$\int_{\mathbf{r}} df = f(\mathbf{r}(b)) - f(\mathbf{r}(a)).$$

11.6 Surface integrals

We turn now to surface integrals. Just as for curves, it is possible to integrate scalar functions on surfaces. Let $f(x, y, x)$ be a continuous function defined on $\mathbb{R}^3$, or on a parametrized surface Σ, and let $H\colon D \to \mathbb{R}^3$ be the parametrization. We set $\iint_\Sigma f = \int_D f \cdot \|\mathbf{T}_u \times \mathbf{T}_v\| du dv$ and think of this as summing up the values of f over the surface. As before, this integral is independent of the parametrization and can be defined for surfaces in $\mathbb{R}^n$ as well, by replacing $\|\mathbf{T}_u \times \mathbf{T}_v\|$ with $\|\mathbf{T}_u \wedge \mathbf{T}_v\|$. Sometimes we write $\iint_\Sigma f dA$ to emphasize that we are integrating with respect to surface area. One way such an integral might arise is as follows: we could think of Σ as a thin surface and f as a density function on Σ, and then the integral would be the total density of Σ.

Instead of integrals of scalar functions, we will again be mainly concerned with defining the integral of a *vector field* $\mathbf{F}$ on a surface in $\mathbb{R}^3$. At first, this will only work in $\mathbb{R}^3$, because the objects that can be integrated over surfaces in $\mathbb{R}^n$, $n > 3$, are **not** vector fields. The physical motivation is as follows. Suppose that we have a vector field $\mathbf{F}$ in $\mathbb{R}^3$, which we think of as the vector field associated to a fluid flowing through space, by taking the velocity vector to a curve traced by a particle in the fluid at time t_0. If Σ is a surface in $\mathbb{R}^3$, we want to measure the total amount of fluid leaving the surface at time t_0. Only the component of $\mathbf{F}$ in the normal direction to Σ matters, and so we can measure this amount by the *surface integral*

$$\iint_\Sigma \mathbf{F} = \iint_D \langle \mathbf{F}, \mathbf{N} \rangle \, dA,$$

where $\mathbf{N} = \mathbf{T}_u \times \mathbf{T}_v / \|\mathbf{T}_u \times \mathbf{T}_v\|$ is the unit outward normal and dA is the "area element" $\|\mathbf{T}_u \times \mathbf{T}_v\|\, dudv$. In physics, this integral is called the *flux* of the vector field $\mathbf{F}$ through the surface Σ. As with line integrals, this potentially complicated expression simplifies and is easier to compute than a surface area:

$$\begin{aligned}\iint_D \langle \mathbf{F}, \mathbf{N} \rangle \, dA &= \iint_D \left\langle \mathbf{F}, \frac{\mathbf{T}_u \times \mathbf{T}_v}{\|\mathbf{T}_u \times \mathbf{T}_v\|} \right\rangle \|\mathbf{T}_u \times \mathbf{T}_v\| \, dudv \\ &= \iint_D \langle \mathbf{F}, \mathbf{T}_u \times \mathbf{T}_v \rangle \, dudv,\end{aligned}$$

and this expression is usually computable. In physics this quantity is called the *flux* of $\mathbf{F}$ through Σ. A little calculation gives

$$\mathbf{T}_u \times \mathbf{T}_v = \left(\frac{\partial(y,z)}{\partial(u,v)}, \frac{\partial(z,x)}{\partial(u,v)}, \frac{\partial(x,y)}{\partial(u,v)} \right).$$

Thus, for a vector field $\mathbf{F} = (F_1, F_2, F_3)$,

$$\langle \mathbf{F}, \mathbf{T}_u \times \mathbf{T}_v \rangle = F_1 \frac{\partial(y,z)}{\partial(u,v)} + F_2 \frac{\partial(z,x)}{\partial(u,v)} + F_3 \frac{\partial(x,y)}{\partial(u,v)}.$$

and this is the integrand in a surface integral.

A standard argument with the change of variables formula shows that $\iint_\Sigma \mathbf{F}$ is independent of the parametrization, as long as we use an orientation preserving reparametrization. However if we take an orientation reversing reparametrization, then we replace $\iint_\Sigma \mathbf{F}$ by $-\iint_\Sigma \mathbf{F}$. Physically this means that when we change what we mean by the outward side of Σ, then we replace fluid flowing outward by fluid flowing inward and so expect that the sign of the surface integral will change.

Let us compute some examples.

Example 11.20. 1) Take $\mathbf{F} = (x^2+z^2, xy, x^2z+y)$ and let Σ be the graph of $z = f(x,y) = x^2+y^2$, for $1 \le x \le 2, 0 \le y \le 1$. Then $\mathbf{T}_x \times \mathbf{T}_y = (-\nabla f, 1) = (-2x, -2y, 1)$. Thus (using $z = x^2 + y^2$ on Σ) the surface integral is equal to

$$\int_0^1 \int_1^2 \langle (x^2 + (x^2+y^2)^2, xy, x^2(x^2+y^2) + y), (-2x, -2y, 1) \rangle dxdy.$$

Without working out the full calculation, we get the double integral over $[1,2] \times [0,1]$ of some complicated polynomial in x and y, but this integral is

certainly feasible. Of course, this discussion is incomplete because we have not said how we are going to orient Σ. In this case, the graph parametrization orients Σ by the normal $\mathbf{N} = \mathbf{T}_x \times \mathbf{T}_y / \|\mathbf{T}_x \times \mathbf{T}_y\|$, which points radially inward (towards the z-axis) because the first two components of $\mathbf{T}_x \times \mathbf{T}_y$ are $(-2x, -2y)$ and up, because the z-component is 1.

2) Take Σ to be the unit sphere, oriented by the *outward* normal, and let $\mathbf{F}$ be the radial vector field. In this case, we expect the flux to be positive. In fact, choose the spherical coordinates parametrization on Σ. Then $\mathbf{T}_\theta \times \mathbf{T}_\phi = -\sin\phi(x, y, z) = -(\sin\phi)\mathbf{r}$, which has the **wrong** orientation, and so we should multiply by -1 to get the right answer. Thus (after adjusting the sign) $\mathbf{N} = \mathbf{r} = (x, y, z)$, and so

$$\iint_\Sigma \mathbf{r} = -\iint_\Sigma -\sin\phi\langle \mathbf{r}, \mathbf{r}\rangle = \int_0^\pi \int_0^{2\pi} \sin\phi \, d\theta d\phi.$$

In this case, the flux is equal to 4π, which is also the surface area. This is to be expected, because on Σ, $\mathbf{r} = \mathbf{N}$ is the unit outward normal, and hence $\langle \mathbf{r}, -\mathbf{T}_\theta \times \mathbf{T}_\phi\rangle = \|\mathbf{T}_\theta \times \mathbf{T}_\phi\|$, so the integrand is the one which computes the surface area.

Remark 11.21. It is sometimes hard to describe the orientation exactly. If we have computed $\mathbf{T}_u \times \mathbf{T}_v$, then evaluating at a sample point may help. Another result worth mentioning in this context is the following:

Theorem 11.22 (Jordan-Brouwer separation theorem in dimension three)**.** *Let X be a compact connected smooth surface in $\mathbb{R}^3$. Then the open set $\mathbb{R}^3 - X$ is equal to $U_1 \cup U_2$, where U_1 and U_2 are open and connected, $U_1 \cap U_2 = \emptyset$, U_1 is bounded and U_2 is unbounded.*

Here, you should think of the bounded open set U_1 as the "inside" and the unbounded open set U_2 as the "outside." For example, if $X = S^2$, then U_1 is the open unit ball and U_2 is the complement of the closed unit ball. This is the analogue of the Jordan curve theorem, which says that a simple closed curve in $\mathbb{R}^2$ divides the plane into two sets U_1 and U_2 as in the theorem above. A similar result holds for compact connected $(n-1)$-manifolds in $\mathbb{R}^n$. Returning to the case of a compact connected surface X in $\mathbb{R}^3$, we see that X is always orientable, because we can always consistently choose the orientation $\mathbf{N}$ which points **outward**, into U_2, or **inward**, into U_1. However, the convention is that, unless otherwise stated, we orient a compact connected surface by taking the **outward** pointing normal.

So far, we have only defined the surface integral of a vector field in $\mathbb{R}^3$. It is natural to ask what can be done in $\mathbb{R}^n$ for $n > 3$. The answer is that we can define surface integrals, but not of vector fields. Instead we integrate a more complicated object called a 2-*form*. To see this in $\mathbb{R}^3$, given the vector field $\mathbf{F} = (F_1, F_2, F_3)$, we make it correspond to the formal symbol

$$\omega = F_1 dy \wedge dz + F_2 dz \wedge dx + F_3 dx \wedge dy.$$

Here the symbols dx, dy, dz can be multiplied according to the rules,

$$dx \wedge dy = -dy \wedge dx, dx \wedge dz = -dz \wedge dx, dy \wedge dz = -dz \wedge dy,$$
$$dx \wedge dx = dy \wedge dy = dz \wedge dz = 0.$$

Now suppose that x, y, z are functions of u, v on some open set D in $\mathbb{R}^2$ via the parametrization H of Σ. We use the usual convention

$$dx = \frac{\partial x}{\partial u} du + \frac{\partial x}{\partial v} dv,$$

and similarly for y and z, together with the analogous rules for multiplication of du and dv:

$$dv \wedge du = -du \wedge dv, du \wedge du = dv \wedge dv = 0.$$

With these rules, a calculation shows that

$$dy \wedge dz = \frac{\partial(y,z)}{\partial(u,v)} du \wedge dv, dz \wedge dx = \frac{\partial(z,x)}{\partial(u,v)} du \wedge dv, dx \wedge dy = \frac{\partial(x,y)}{\partial(u,v)} du \wedge dv.$$

If we plug in these formulas in the expression

$$\iint_\Sigma F_1 dy \wedge dz + F_2 dz \wedge dx + F_3 dx \wedge dy,$$

then we see that the "integrand" becomes

$$\left(F_1 \frac{\partial(y,z)}{\partial(u,v)} + F_2 \frac{\partial(z,x)}{\partial(u,v)} + F_3 \frac{\partial(x,y)}{\partial(u,v)} \right) du \wedge dv,$$

and as we have already seen, this is the same as $\langle \mathbf{F}, \mathbf{T}_u \times \mathbf{T}_v \rangle \, dudv$ if we agree to turn the odd-looking expression $du \wedge dv$ into the more familiar (but equally meaningless) expression $dudv$. Thus

$$\iint_\Sigma \omega = \iint_D \langle \mathbf{F}, \mathbf{T}_u \times \mathbf{T}_v \rangle \, dudv = \iint_\Sigma \mathbf{F}$$

as previously defined. As for the case of line integrals, given the parametrization H, we write $H^*\omega = \langle \mathbf{F}, \mathbf{T}_u \times \mathbf{T}_v \rangle \, du \wedge dv$ and think that the parametrization H has converted the mysterious symbol ω into an expression of the form $f(u,v)\, dudv$ which is then something we can integrate on D.

Now we can generalize to $\mathbb{R}^n$: define a 2-*form* ω on $\mathbb{R}^n$ to be an expression of the form

$$\sum_{i<j} F_{ij} dx_i \wedge dx_j.$$

Here the F_{ij}, the components of the form, are just real-valued C^∞ functions on $\mathbb{R}^n$. Note that there are in general $n(n-1)/2$ such components, so we are no longer trying to integrate a vector field if $n > 3$. We multiply by the same rules as before:

$$dx_i \wedge dx_j = -dx_j \wedge dx_i, \qquad i \neq j; \qquad dx_i \wedge dx_i = 0.$$

Now on a parametrized surface Σ, given by a C^∞ parametrization $H\colon D \to \mathbb{R}^n$, we have the same rules for plugging in for u and v as before and write $F_{ij} dx_i \wedge dx_j$ in the form $f(u,v)\, du \wedge dv$, which as before we denote by $H^*\omega$. Once we agree to replace $du \wedge dv$ by $dudv$, we can then integrate $H^*\omega$ over D. Just as in the surface case, we have to choose an **orientation** on Σ. Here, roughly, two parametrizations H_1 and H_2 define the same orientation $\iff$ $\det D(H_1 \circ H_2^{-1})$ is always positive. Not all surfaces Σ have an orientation, but a choice is necessary in order to define the integral of ω.

Although the definition looks unmotivated, 2-forms and their integrals play a very important role both in mathematics and in physics. For example, in mathematics they arise in trying to measure the curvature of a manifold. They also play a role in the modern formulation of Maxwell's equations and their generalizations, to what are called *Yang-Mills fields*. For example, in Maxwell's equations, if $\mathbf{E} = (E_1, E_2, E_3)$ is the electric field and $\mathbf{B} = (B_1, B_2, B_3)$ is the magnetic field, both depending on x, y, z, t (i.e. time-dependent), then let E be the 1-form $E_1 dx + E_2 dy + E_3 dz$ corresponding to $\mathbf{E}$ and let $B = B_1\, dy \wedge dz + B_2\, dz \wedge dx + B_3\, dx \wedge dy$ be the 2-form corresponding to $\mathbf{B}$, and define the 2-form ω via

$$\omega = E \wedge dt + B.$$

Then Maxwell's equations become the two equations $d\omega = 0$ and $d(*\omega) = 0$, where $*$ is an analogue of the Hodge $*$-operator, but for the (indefinite) Lorentz metric on $\mathbb{R}^4$.

11.7 Differential forms

We want to generalize the discussion of the previous section to higher dimensional submanifolds. Doing so leads to the definition of differential forms. We will begin by just treating them as formal objects, very similar to wedge product. In the next section, we will describe how one might go about defining them algebraically (although only the formal properties will be important to us).

More generally we can define a C^∞ *k-form* in $\mathbb{R}^n$ to be a sum of expression of the type

$$f(x_1, \ldots, x_n)\, dx_{i_1} \wedge \cdots \wedge dx_{i_k},$$

where f is a C^∞ function on $\mathbb{R}^n$ and $i_1 < i_2 < \cdots < i_k$. If $I = \{i_1, i_2, \ldots, i_k\}$ with $i_1 < i_2 < \cdots < i_k$, then we abbreviate $dx_{i_1} \wedge \cdots \wedge dx_{i_k}$ by dx_I and write a general k-form on $\mathbb{R}^n$ as

$$\sum_I f_I(x_1, \ldots, x_n)\, dx_I,$$

where the coefficients f_I are C^∞ functions on $\mathbb{R}^n$. More generally, if U is an open subset of $\mathbb{R}^n$, then we define a C^∞ *k-form on* U similarly, where we just assume that the f_I are C^∞ functions on U. In general a k-form on U has $\binom{n}{k}$ terms. For example, an n-form on $U \subseteq \mathbb{R}^n$ has the form $f(x_1, \ldots, x_n)\, dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$, and so is the same thing as a function. We also define a 0-*form* on U to be a function $f(x_1, \ldots, x_n)$ on U (i.e. there are no dx_i terms). We denote the set of all k-forms on an open set U of $\mathbb{R}^n$ by $\Omega^k(U)$. It is a vector space (infinite dimensional) under addition and scalar multiplication. If $\omega \in \Omega^k(U)$, we call k the *degree* of the k-form ω.

There are three important operations on k-forms: first we can multiply a k-form times an ℓ-form according to rules similar to those we have defined above. In mathematics this is usually called *wedge product* and is often denoted by $\wedge$. The product is commutative up to a sign, but is not commutative in general. As with wedge product in $\mathbb{R}^n$, wedge product of forms is specified by the rules

$$dx_i \wedge dx_j = -dx_j \wedge dx_i,$$

from which it follows that $dx_i \wedge dx_i = -dx_i \wedge dx_i$ and so $dx_i \wedge dx_i = 0$. In general, the requirement that wedge product be an associative bilinear operation gives the rules

$$\left(\sum_I f_I(x_1, \ldots, x_n)\, dx_I\right) \wedge \left(\sum_J g_J(x_1, \ldots, x_n)\, dx_J\right) = \sum_{I,J} f_I g_J\, dx_I \wedge dx_J,$$

together with the requirement that

$$dx_I \wedge dx_J = \begin{cases} \mathbf{0}, & \text{if } I \cap J \neq \emptyset; \\ \pm dx_{I \cup J}, & \text{if } I \cap J = \emptyset, \end{cases}$$

where, as with wedge product in $\mathbb{R}^n$, the sign is determined by the number of times we have to switch the order until the factors are multiplied in increasing order.

Wedge product on forms is anticommutative: if $\omega \in \Omega^k(U)$ and $\psi \in \Omega^\ell(U)$, then

$$\omega \wedge \psi = (-1)^{k\ell} \psi \wedge \omega.$$

Thus if k and ℓ are both odd, then $\omega \wedge \psi = -\psi \wedge \omega$, and in particular if ω has odd degree then $\omega \wedge \omega = 0$. This does not hold true necessarily for even degree forms: for example, if $\omega = dx_1 \wedge dx_2 + dx_3 \wedge dx_4$, then

$$\omega \wedge \omega = (dx_1 \wedge dx_2 + dx_3 \wedge dx_4) \wedge (dx_1 \wedge dx_2 + dx_3 \wedge dx_4) = 2dx_1 \wedge dx_2 \wedge dx_3 \wedge dx_4 \neq 0.$$

There is also a $*$-operator from $\Omega^k(U)$ to $\Omega^{n-k}(U)$, but we shall not discuss it here.

The second operation we can do with forms is to take a derivative. First, if f is a function, then we have already defined df to be the 1-form which corresponds to the gradient:

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n.$$

In particular, with this definition, $d(x_i) = dx_i$ as previously defined. For a general k-form on U, we begin by defining

$$d(f\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}) = df \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

with the usual rules. Then we extend d by forcing it to be linear:

$$d\left(\sum_I f_I\, dx_I \right) = \sum_I d(f_I\, dx_I) = \sum_I df_I \wedge dx_I.$$

This defines a linear function $d\colon \Omega^k(U) \to \Omega^{k+1}(U)$, called the *exterior derivative*.

For example, if $\mathbf{F}$ is a vector field (F_1, F_2, F_3) in $\mathbb{R}^3$, and we make it correspond to the 1-form $\omega = F_1 dx + F_2 dy + F_3 dz$, then

$$d\omega = dF_1 \wedge dx + dF_2 \wedge dy + dF_3 \wedge dz.$$

The first term for example is

$$\left(\frac{\partial F_1}{\partial x}dx + \frac{\partial F_1}{\partial y}dy + \frac{\partial F_1}{\partial z}dz\right) \wedge dx = \frac{\partial F_1}{\partial y}dx \wedge dy + \frac{\partial F_1}{\partial z}dz \wedge dx$$
$$= -\frac{\partial F_1}{\partial y}dy \wedge dx + \frac{\partial F_1}{\partial z}dz \wedge dx,$$

where we have used for example the rules $dx \wedge dx = 0$ and $dy \wedge dx = -dx \wedge dy$. Collecting terms gives

$$d\omega = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right) dy \wedge dz + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right) dz \wedge dx + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dx \wedge dy,$$

which corresponds to the curl of $\mathbf{F}$, usually written as $\nabla \times \mathbf{F}$, and given formally as a determinant:

$$\nabla \times \mathbf{F} = \det \begin{pmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{pmatrix}.$$

The case of a 1-form in $\mathbb{R}^2$ is really a special case of this calculation: if $\omega = F_1\,dx + F_2\,dy$, then

$$d\omega = d(F_1\,dx + F_2\,dy) = \left(\frac{\partial F_2}{dx} - \frac{\partial F_1}{dy}\right) dx \wedge dy.$$

Likewise, given the 2-form $\psi = F_1 dy \wedge dz + F_2 dz \wedge dx + F_3 dx \wedge dy$, we have, by a similar calculation,

$$d\psi = \left(\frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}\right) dx \wedge dy \wedge dz,$$

which corresponds to the divergence of $\mathbf{F}$, usually written symbolically as $\nabla \cdot \mathbf{F}$. The operators grad, curl, div are the standard operators of vector calculus in $\mathbb{R}^3$.

One basic identity in vector calculus involves the following sequence of operations

$$\nabla \times (\nabla f) = \mathbf{0}; \qquad \nabla \cdot (\nabla \times \mathbf{G}) = 0.$$

In other words, curl of the gradient is zero, and divergence of the curl is zero. (The other possible combinations $\nabla \cdot (\nabla f)$ and $\nabla(\nabla \cdot \mathbf{G})$ are **not** usually zero.) In general, one has the following identity for forms:

Theorem 11.23. *The function $d \circ d = d^2$ from $\Omega^k(U)$ to $\Omega^{k+2}(U)$ is identically zero.*

Proof. We shall just check this for d^2 from 0-forms to 2-forms. In other words, we shall show that, if f is a C^∞ function, then $d(df) = 0$. Now

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n.$$

Hence

$$d(df) = d\left(\frac{\partial f}{\partial x_1}\right) dx_1 + \cdots + d\left(\frac{\partial f}{\partial x_n}\right) dx_n.$$

In this expression, the term $dx_i \wedge dx_j$, $i < j$, appears twice: first from the corresponding term in

$$d\left(\frac{\partial f}{\partial x_j}\right) dx_j$$

and again from

$$d\left(\frac{\partial f}{\partial x_i}\right) dx_i.$$

The first term contributes

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) dx_i \wedge dx_j = \frac{\partial^2 f}{\partial x_i \partial x_j} dx_i \wedge dx_j$$

and the second gives

$$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) dx_j \wedge dx_i = -\frac{\partial^2 f}{\partial x_j \partial x_i} dx_i \wedge dx_j.$$

By equality of mixed partials, the coefficient of $dx_i \wedge dx_j$ is zero for all $i < j$, and thus $d(df) = 0$. □

Definition 11.24. A k-form ω is *exact* if $\omega = d\varphi$ for some $(k-1)$-form φ.

Since $d^2 = 0$, if ω is exact, then $d\omega = 0$. A form ω such that $d\omega = 0$ is called *closed*. We see that every exact form is closed. If U is a connected open subset of $\mathbb{R}^n$, then a 0-form (i.e. function) f is closed $\iff$ it is constant, and it is never exact since there are no forms of degree -1. An n-form $f\, dx_1 \wedge \cdots \wedge dx_n$ is always closed, since $\Omega^{n+1}(U) = \{0\}$. It turns out (but it is much harder to prove) that an n-form on an open subset of $\mathbb{R}^n$ is always exact.

For 1-forms ω defined in a connected open set, if $\omega = df$ for some function f, then f is unique up to adding a constant. In case ω is a k-form with $k > 1$

and $\omega = d\psi$, the most we can say is that ψ is unique up to adding a $(k-1)$-form α with $d\alpha = 0$, in other words α is a closed $(k-1)$-form, and there are in general many such forms α. For example, if $\alpha = d\beta$ itself is exact, then $d\alpha = 0$ since $d^2 = 0$ and so $d\psi = d(\psi + \alpha)$.

A result usually called the Poincaré lemma is the following, which generalizes the statement of Theorem 11.17, that a vector field defined on a convex open set is a gradient vector field if and only if it satisfies the mixed partials condition (and is proved in a similar but more complicated way):

Theorem 11.25 (Poincaré lemma)**.** *Let U be a convex open subset of $\mathbb{R}^n$. If ω is a C^∞ k-form on U such that $d\omega = 0$, then $\omega = d\varphi$ for some $(k-1)$-form φ on U. In other words, a C^∞ closed form on U is exact.*

In terms of vector calculus, this says that if $\mathbf{F}$ is a vector field defined on a convex open subset U of $\mathbb{R}^3$, then (i) if $\nabla \times \mathbf{F} = \mathbf{0}$, then $\mathbf{F} = \nabla f$ for some function f, and (ii) if $\nabla \cdot \mathbf{F} = 0$, then $\mathbf{F} = \nabla \times \mathbf{G}$ for some vector field $\mathbf{G}$.

There is also a formula for how d behaves with respect to wedge product, which is a straightforward calculation from the definitions:

Proposition 11.26. *If ω is a C^∞ k-form and ψ is a C^∞ ℓ-form in U, then*

$$d(\omega \wedge \psi) = d\omega \wedge \psi + (-1)^k \omega \wedge d\psi.$$

This is the generalization of various vector calculus identities such as:

$$\nabla \times (f\mathbf{F}) = (\nabla f) \times \mathbf{F} + f(\nabla \times \mathbf{F});$$
$$\nabla \cdot (\mathbf{F} \times \mathbf{G}) = (\nabla \times \mathbf{F}) \cdot \mathbf{G} - \mathbf{F} \cdot (\nabla \times \mathbf{G}).$$

The third operation on differential forms is *pullback*, and it does not exactly have an analogue in vector calculus (although it is certainly used, as in our discussion of line and surface integrals). Suppose that $\omega \in \Omega^k(U)$, where U is an open subset of $\mathbb{R}^n$. Let V be an open subset of $\mathbb{R}^m$ and let $H \colon V \to U$ be a C^∞ map. In other words, $H \colon V \to \mathbb{R}^n$ is a C^∞ function whose image is contained in U. We don't make any additional requirements on H or on DH as we did in the discussion of parametrizations. For example, H could be constant. Writing $H = (h_1, \ldots, h_n$, a vector-valued function of $u_1, \ldots, u_m$, we can also think of H as defined by $x_i = h_i(u_1, \ldots, u_m)$, or even by writing $x_i = x_i(u_1, \ldots, u_m)$. Then, for $\omega \in \Omega^k(U)$ we define the pullback $H^*\omega \in \Omega^k(V)$, by the following rules: for a function (i.e. 0-form) f, $H^*(f)$ is the pullback of functions: $H^*(f) = f \circ H$. For the 1-form dx_i, $H^*(dx_i)$ is the 1-form we would expect by writing $x_i = h_i(u_1, \ldots, u_m)$:

$$H^*(dx_i) = \sum_{j=1}^{m} \frac{\partial h_i}{\partial u_j} du_j.$$

More generally, for a k-form of the form $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_k}$, we set

$$H^*(dx_I) = H^*(dx_{i_1} \wedge \cdots \wedge dx_{i_k}) = H^*(dx_{i_1}) \wedge \cdots \wedge H^*(dx_{i_k}).$$

For a general k-form $\omega = \sum_I f_I dx_I$, we define

$$H^*(\omega) = \sum_I H^*(f_I) H^*(dx_I).$$

The key properties of pullback are as follows (proofs by direct calculation):

Proposition 11.27. *Let U be an open subset of $\mathbb{R}^n$ and let $H\colon V \to U$ be a C^∞ map, where V is an open subset of $\mathbb{R}^m$. Then:*

1. *For all $\omega \in \Omega^k(U)$ and $\psi \in \Omega^\ell(U)$,*

$$H^*(\omega \wedge \psi) = H^*(\omega) \wedge H^*(\psi).$$

2. *For all $\omega \in \Omega^k(U)$,*

$$H^*(d\omega) = d(H^*(\omega)).$$

It is a straightforward calculation that, in case $\omega = dx_1 \wedge \cdots \wedge dx_n$ and H is a diffeomorphism from an open set U in $\mathbb{R}^n$ to an open set V in $\mathbb{R}^n$, then, writing $H = (h_1, \ldots, h_n)$ with $h_i = h_i(u_1, \ldots, u_n)$, we have

$$H^*(\omega) = H^*(dx_1 \wedge \cdots \wedge dx_n) = \det \frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)}\, du_1 \wedge \cdots \wedge du_n.$$

In other words, under a coordinate change, a differential form of degree n in $\mathbb{R}^n$ transforms via multiplication by the determinant of the Jacobian matrix. This is the basic reason why differential forms can be integrated on submanifolds.

11.8 Algebraic interpretation of differential forms

We presented differential forms in much the same way as we did wedge product, as a formal set of symbols which could be manipulated according to a formal set of rules, with no attempt to explain why these symbols should exist and why these rules should hold. Here, we want to make a connection between differential forms and alternating multilinear functions, and explain how one might go about constructing the objects and operations we have been using. However, only the formal properties will matter to us.

First, we recall the following definition:

Definition 11.28. Let V be a vector space. A function

$$F\colon V^k = \underbrace{V \times \cdots \times V}_{k \text{ times}} \to \mathbb{R}$$

is *k-multilinear* or a *k-tensor* (or simply *multilinear* or a *tensor* if k is clear from the context) if it is linear in each variable, in other words, if for all $\mathbf{v}_j, \mathbf{w}_1, \mathbf{w}_2 \in V$, $t \in R$,

$$\begin{gathered} F(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}_1 + \mathbf{w}_2, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k) = \\ = F(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}_1, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k) + F(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{w}_2, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k); \\ F(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, t\mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k) = tF(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k). \end{gathered}$$

(Strictly speaking there are more general kinds of tensors as well.) The multilinear function F is *symmetric* if, for all permutations $f \in S_k$,

$$F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}) = F(\mathbf{v}_1, \ldots, \mathbf{v}_k),$$

in other words its value remains unchanged if we permute the arguments. The multilinear function F is *alternating* if, for all permutations $f \in S_k$,

$$F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}) = (\operatorname{sign} f) F(\mathbf{v}_1, \ldots, \mathbf{v}_k),$$

and as we have seen it is enough to assume that, for all i,

$$F(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k) = -F(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \mathbf{v}_i, \ldots, \mathbf{v}_k).$$

We let $T^k V^*$ be the vector space of all k-multilinear functions from V to $\mathbb{R}$, $S^k V^*$ the vector subspace of all symmetric k-multilinear functions, and $\bigwedge^k V^*$ the vector subspace of all alternating k-multilinear functions. For consistency (or by logic), we set $T^0 V^* = S^0 V^* = \bigwedge^0 V^* = \mathbb{R}$, at least if $V \neq \{\mathbf{0}\}$. If $k = 1$, then, since $S_1 = \{\mathrm{Id}\}$, $T^1 V^* = S^1 V^* = \bigwedge^1 V^* = V^*$, the dual space of V. But if $k > 1$, $S^k V^*$ and $\bigwedge^k V^*$ are proper subspaces of $T^k V^*$ and in fact $S^k V^* \cap \bigwedge^k V^* = \{\mathbf{0}\}$ if $k > 1$. For $k = 2$, we shall see in a moment that $S^2 V^* \oplus \bigwedge^2 V^* = T^2 V^*$, i.e. that every bilinear function from V to $\mathbb{R}$ can be uniquely written as a sum of a symmetric and an alternating bilinear function. This no longer holds for $k \geq 3$. In fact, taking $V = \mathbb{R}^n$ (the only case we shall be interested in), it is easy to see the following: a multilinear function $F\colon (\mathbb{R}^n)^k \to \mathbb{R}$ is specified by the n^k numbers

$$a_{i_1, \ldots, i_k} = F(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}),$$

where $i_1, \dots, i_k$ is any sequence of integers with $1 \le i_j \le n$ for all j, and that conversely any collection of real numbers $a_{i_1,\dots,i_k}$ uniquely defines a k-multilinear function F. In this sense, we see that k-tensors are a generalization of matrices. In particular,

$$\dim T^k(\mathbb{R}^n)^* = n^k.$$

The condition that the function F defined by the coefficients $a_{i_1,\dots,i_k}$ is symmetric is just that, for all $f \in S_k$, $a_{i_{f(1)},\dots,i_{f(1)}} = a_{i_1,\dots,i_k}$. Since we can always find a permutation f such that $f(1) \le \cdots \le f(k)$, it is easy to see that F is specified by the numbers $a_{i_1,\dots,i_k}$ with $i_1 \le \cdots \le i_k$. The collection of all non-decreasing (i.e. increasing, but not necessarily strictly increasing) sequences $i_1 \le \cdots \le i_k$, with $1 \le i_j \le n$, can be counted: the number of such sequences is $\binom{n+k-1}{k}$. Hence

$$\dim S^k(\mathbb{R}^n)^* = \binom{n+k-1}{k}.$$

Finally, a multilinear function F specified by $a_{i_1,\dots,i_k}$ is alternating $\iff$ for all $f \in S_k$, $a_{i_{f(1)},\dots,i_{f(1)}} = (\operatorname{sign} f) a_{i_1,\dots,i_k}$. As usual this forces $a_{i_1,\dots,i_k} = 0$ if there are any repeated indices, and again by suitably permuting we can arrange that $i_1 < \cdots < i_k$, in other words that $i_1, \dots, i_k$ is a (strictly) increasing sequence with $1 \le i_j \le n$. But the number of strictly increasing sequences of $\{1, \dots, n\}$ of length k is the same as the number of subsets of $\{1, \dots, n\}$ with k elements, since given a subset I of $\{1, \dots, n\}$, there is a unique way to list its elements in increasing order. Hence

$$\dim \bigwedge^k (\mathbb{R}^n)^* = \binom{n}{k}.$$

In particular, for $k > n$, $\bigwedge^k (\mathbb{R}^n)^* = \{0\}$. Note that, for $k = 1$ we have $\binom{n+k-1}{k} = \binom{n}{1} = n$. For $n = 2$,

$$\dim S^2(\mathbb{R}^n)^* + \dim \bigwedge^2 (\mathbb{R}^n)^* = \binom{n+1}{2} + \binom{n}{2} = \frac{(n+1)n}{2} + \frac{n(n-1)}{2} = n^2.$$

But for $k > 2$ and $n > 1$, a calculation shows that $\dim S^k(\mathbb{R}^n)^* + \dim \bigwedge^k (\mathbb{R}^n)^*$ is always strictly less than $n^k = \dim T^k(\mathbb{R}^n)^*$. (The attempt to describe which multilinear functions are not a sum of a symmetric and an alternating function leads to studying representations of the symmetric group S_k.)

Given an arbitrary k-multilinear function $F \in T^k V^*$, one can try to associate to it a symmetric or alternating multilinear function. It is easier to see how to do this to get a symmetric function: we can just average the values of F over all permutations of the variables. Thus, given $F \in T^k V^*$, define

$$\pi_s(F)(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \frac{1}{k!} \sum_{f \in S_k} F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}).$$

Proposition 11.29. *With notation as above,*

1. *If $F \in T^k V^*$, then $\pi_s(F)$ is a symmetric multilinear function, i.e. $\pi_s(F) \in S^k V^*$.*
2. *The function $\pi_s \colon T^k V^* \to S^k V^*$ is linear.*
3. *If $F \in S^k V^*$, then $\pi_s(F) = F$.*

Proof. We shall just outline the argument. Since $\pi_s(F)$ is a sum of multilinear terms, it is multilinear. Given $g \in S_k$,

$$\pi_s(F)(\mathbf{v}_{g(1)}, \ldots, \mathbf{v}_{g(1)}) = \frac{1}{k!} \sum_{f \in S_k} F(\mathbf{v}_{f \circ g(1)}, \ldots, \mathbf{v}_{f \circ g(k)}) = \frac{1}{k!} \sum_{f \in S_k} F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}),$$

since summing over all permutations of the form $f \circ g$, where g is a fixed element of S_k, is the same as summing over all elements of S_k. (In other words, for a fixed element g of S_k, the function $h \colon S_k \to S_k$ defined by $h(f) = f \circ g$ is a bijection, with inverse $h^{-1}(f) = f \circ g^{-1}$.) This proves (1); (2) is an immediate calculation. Finally, if $F \in S^k V^*$ to begin with, then $F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)})$ for all $f \in S_k$, and so

$$\pi_s(F)(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \frac{1}{k!} \sum_{f \in S_k} F(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \frac{1}{k!} k! F(\mathbf{v}_1, \ldots, \mathbf{v}_k) = F(\mathbf{v}_1, \ldots, \mathbf{v}_k).$$

□

Note that it is possible for $\pi_s(F)$ to be zero even if F is nonzero. For example, if F is alternating and $k > 1$ then a calculation shows that $\pi_s(F) = 0$. The key point is that, if $k > 1$, then $\sum_{f \in S_k} \operatorname{sign} f = 0$ (because the number of even permutations, i.e. the number of elements of A_k, the subset of S_k consisting of even permutations, is the same as the number of odd permutations).

We can make a very similar construction which associates to a k-multilinear function an alternating function: define, for $F \in T^k V^*$,

$$\pi_a(F)(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \frac{1}{k!} \sum_{f \in S_k} (\operatorname{sign} f) F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}).$$

Proposition 11.30. *With notation as above,*

1. *If $F \in T^k V^*$, then $\pi_a(F)$ is an alternating multilinear function, i.e. $\pi_a(F) \in \bigwedge^k V^*$.*
2. *The function $\pi_a \colon T^k V^* \to \bigwedge^k V^*$ is linear.*
3. *If $F \in \bigwedge^k V^*$, then $\pi_a(F) = F$.*

Proof. The arguments are very similar to those of the previous proposition using $\operatorname{sign}(f \circ g) = (\operatorname{sign} f)(\operatorname{sign} g)$ and $(\operatorname{sign} f)^2 = 1$. □

Again, it is possible for $\pi_a(F)$ to be zero even if F is nonzero. For example, if F is symmetric and $k > 1$ then a calculation as above shows that $\pi_a(F) = 0$.

Example 11.31. For $k = 2$ and $F \colon V \times V \to \mathbb{R}$ a bilinear form on V,

$$\pi_s(F)(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{2}(F(\mathbf{v}_1, \mathbf{v}_2) + F(\mathbf{v}_2, \mathbf{v}_1));$$
$$\pi_a(F)(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{2}(F(\mathbf{v}_1, \mathbf{v}_2) - F(\mathbf{v}_2, \mathbf{v}_1)).$$

Hence, we see that, for all $F \in T^2 V^*$,

$$F = \pi_s(F) + \pi_a(F),$$

and in particular F is (in fact uniquely) a sum of a symmetric and an alternating bilinear form.

We turn now to various products involving the vector spaces we have defined. The confusing point is that there are a lot of different, but related, products. For example, as we mentioned briefly last semester, an element F of $S^k(\mathbb{R}^n)^*$ is the same thing as a homogeneous polynomial of degree k in n variables $x_1, \ldots, x_n$. We describe this correspondence in one direction: if $F \in S^k(\mathbb{R}^n)^*$, then, setting $\mathbf{x} = (x_1, \ldots, x_n)$, it is easy to see that

$$P(\mathbf{x}) = F(\mathbf{x}, \ldots, \mathbf{x})$$

is a homogeneous polynomial of degree k in $x_1, \ldots, x_n$. (This part works even if F is not necessarily symmetric, just multilinear.) Conversely, if $P(\mathbf{x})$ is a homogeneous polynomial of degree k in $x_1, \ldots, x_n$, then there is a polarization identity which associates to P a symmetric multilinear function, and these two constructions are inverse to each other. Now we can multiply two polynomials; in fact, this is a commutative, associative operation. So there should be a corresponding way to multiply an element of $S^k(\mathbb{R}^n)^*$ times an element of $S^\ell(\mathbb{R}^n)^*$ to obtain an element of $S^{k+\ell}(\mathbb{R}^n)^*$, and similarly for $T^k(\mathbb{R}^n)^*$ and $\bigwedge^k(\mathbb{R}^n)^*$.

We begin with how to multiply $F \in T^k(\mathbb{R}^n)^*$, or more generally $F \in T^kV^*$, with $G \in T^\ell V^*$; this operation is called *tensor product*, and is denoted by the symbol $\otimes$. Despite the strange symbol, it is defined in a straightforward way: $F \otimes G \in T^{k+\ell}V^*$ is the multilinear function given by

$$(F \otimes G)(\mathbf{v}_1, \ldots, \mathbf{v}_{k+\ell}) = F(\mathbf{v}_1, \ldots, \mathbf{v}_k)G(\mathbf{v}_{k+1}, \ldots, \mathbf{v}_{k+\ell}).$$

For example, in case F and G are **linear** functions, then $F \otimes G$ is the **bilinear** function defined by

$$(F \otimes G)(\mathbf{v}_1, \mathbf{v}_2) = F(\mathbf{v}_1)G(\mathbf{v}_2).$$

It is easy to see that $F \otimes G$ is multilinear; this amounts to the fact that multiplication of real numbers distributes over addition and is associative. It is also easy to see from the associativity of multiplication or real numbers that $\otimes$ is associative: for all $F \in T^kV^*$, $G \in T^\ell V^*$, and $H \in T^mV^*$,

$$(F \otimes G) \otimes H = F \otimes (G \otimes H)$$

as elements of $T^{k+\ell+m}V^*$. However, there is **no** sense in which $\otimes$ is commutative, as can be seen even in the case $k = \ell = 1$.

Having defined a multiplication for $F \in T^kV^*$ and $G \in T^\ell V^*$, we now try to define similar multiplications in case $F \in S^kV^*$ and $G \in S^\ell V^*$, or $F \in \bigwedge^k V^*$ and $G \in \bigwedge^\ell V^*$. Of course, given $F \in S^kV^*$ and $G \in S^\ell V^*$, we could just try to define a product by taking $F \otimes G \in T^{k+\ell}V^*$. However, even if F and G are both symmetric, $F \otimes G$ will almost never be symmetric. To remedy this, we can "symmetrize" the product by defining instead $F \cdot G = \pi_s(F \otimes G)$. And in fact one can check that this operation corresponds to multiplication of polynomials.

The alternating case is a little trickier. Our first guess is to define the "alternating product" of $F \in \bigwedge^k V^*$ and $G \in \bigwedge^\ell V^*$ by the formula

$$\pi_a(F \otimes G) = \frac{1}{(k+\ell)!} \sum_{f \in S_{k+\ell}} (\operatorname{sign} f) F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}) G(\mathbf{v}_{f(k+1)}, \ldots, \mathbf{v}_{f(k+\ell)}).$$

For example, in case $k = \ell = 1$, we would define the alternating product of two linear functions F and G to be

$$\frac{1}{2}(F(\mathbf{v}_1)G(\mathbf{v}_2) - F(\mathbf{v}_2)G(\mathbf{v}_1)).$$

But for various reasons the coefficient $\frac{1}{2}$ is not ideal. If instead we define

$$F \wedge G = \frac{(k+\ell)!}{k!\ell!}\pi_a(F \otimes G) = \frac{1}{k!\ell!}\sum_{f \in S_{k+\ell}} (\operatorname{sign} f) F(\mathbf{v}_{f(1)}, \ldots, \mathbf{v}_{f(k)}) G(\mathbf{v}_{f(k+1)}, \ldots, \mathbf{v}_{f(k+\ell)}),$$

then it turns out that $\wedge$ has the following properties:

Theorem 11.32. *With $\wedge$ as defined above,*

1. *Wedge product is associative, i.e. for all $\omega_1 \in \bigwedge^{k_1} V^*$, $\omega_2 \in \bigwedge^{k_2} V^*$, and $\omega_3 \in \bigwedge^{k_3} V^*$,*

   $$(\omega_1 \wedge \omega_2) \wedge \omega_3 = (\omega_1 \wedge \omega_2 \wedge \omega_3).$$

   In particular, for all $t \in \mathbb{R}$ and $\omega_1 \in \bigwedge^{k_1} V^*$, $\omega_2 \in \bigwedge^{k_2} V^*$,

   $$(t\omega_1) \wedge \omega_2 = \omega_1 \wedge (t\omega_2) = t(\omega_1 \wedge \omega_2).$$

2. *Wedge product distributes over addition: for all $\omega_1, \omega_2 \in \bigwedge^{k_1} V^*$ and $\omega_3 \in \bigwedge^{k_3} V^*$,*

   $$(\omega_1+\omega_2)\wedge\omega_3 = (\omega_1\wedge\omega_3)+(\omega_2\wedge\omega_3) \text{ and } \omega_3\wedge(\omega_1+\omega_2) = (\omega_3\wedge\omega_1)+(\omega_3\wedge\omega_2).$$

3. *Wedge product is anti-commutative: for all $\omega_1 \in \bigwedge^{k_1} V^*$, $\omega_2 \in \bigwedge^{k_2} V^*$,*

   $$\omega_2 \wedge \omega_1 = (-1)^{k_1 k_2}\omega_1 \wedge \omega_2.$$

4. *If $\ell_1, \ldots, \ell_k \in V^*$ and $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$, then*

   $$(\ell_1 \wedge \cdots \wedge \ell_k)(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \det(\ell_i(\mathbf{v}_j)).$$

   In particular, if $V = \mathbb{R}^n$, and $\mathbf{e}_i^*$ is the dual basis to the standard basis in $\mathbb{R}^n$, in other words $\mathbf{e}_i^*$ is the unique linear function $\mathbb{R}^n \to \mathbb{R}$ such that $\mathbf{e}_i^*(\mathbf{e}_i) = 1$ and $\mathbf{e}_i^*(\mathbf{e}_j) = 0$ for $i \neq j$, then, for $I = \{i_1, \ldots, i_k\}$ a subset of $\{1, \ldots, n\}$ with k elements,

   $$(\mathbf{e}_{i_1}^* \wedge \cdots \wedge \mathbf{e}_{i_k}^*)(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) = \begin{cases} 1, & \text{if } i_\ell = j_\ell \text{ for all } \ell; \\ 0, & \text{if } \{j_1, \ldots, j_k\} \neq I. \end{cases}$$

In the last statement above, if $\{j_1, \ldots, j_k\} = I$ but the j_ℓ are just a permutation of the i_ℓ, then $(\mathbf{e}^*_{i_1} \wedge \cdots \wedge \mathbf{e}^*_{i_k})(\mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_k}) = \pm 1$, depending on the sign of the permutation. This is the main reason for the combinatorial factor we introduced above.

In light of (4) above, we think of $\bigwedge^k(\mathbb{R}^n)^*$ as the dual vector space to $\bigwedge^k \mathbb{R}^n$, with the dual basis $\mathbf{e}^*_{i_1} \wedge \cdots \wedge \mathbf{e}^*_{i_k} = \mathbf{e}^*_I$ to the basis $\mathbf{e}_I$ introduced earlier. It is customary however (and is consistent with various definitions of the exterior derivative d) to use the symbol dx_i for the dual basis element $\mathbf{e}^*_i$, so that $\bigwedge^k(\mathbb{R}^n)^*$ has the basis dx_I and is then just a vector space of dimension $\binom{n}{k}$.

We then, finally, define a *C^∞ differential form ω of degree k* on the open subset U of $\mathbb{R}^n$ to be a C^∞ function from U to $\bigwedge^k(\mathbb{R}^n)^*$. Using the basis dx_I, we can write ω in terms of its components: $\omega = \sum_I f_I dx_I$, where the sum is taken over all subsets I of $\{1, \ldots, n\}$ with k elements. As such, a differential form is just a collection of $\binom{n}{k}$ C^∞ functions f_I. But we also have the operation of wedge product, taken pointwise.

11.9 Stokes' theorem

Let X be a k-dimensional smooth submanifold of $\mathbb{R}^n$, or more generally of an open subset U of $\mathbb{R}^n$. If we want to integrate over X, we will need to assume that X is compact. One possibility is that X is a compact manifold, in other words that $\partial X = \emptyset$. More generally, we will assume that X is a compact manifold with boundary ∂X, which is itself either a smooth conmpact submanifold of $\mathbb{R}^n$ (with boundary $\partial\partial X = \emptyset$), or perhaps more generally that X is a manifold with corners, for example the unit cube. Often X will come with a parametrization $H\colon D \to X$, where D is a standard region in $\mathbb{R}^k$, and thus is also a compact manifold with corners. In the parametrized case, if ω is a C^∞ k-form on U, then we can define

$$\int_X \omega = \int_D H^*(\omega).$$

This quantity depends on the parametrization, but the change of variables formula implies that, if $H_1\colon D_1 \to X$ and $H_2\colon D_2 \to X$ are two different parametrizations which define the same orientation, then

$$\int_{D_1} H_1^*(\omega) = \int_{D_2} H_2^*(\omega).$$

In other words, once X has been oriented, then $\int_X \omega$ is well-defined. Note that, if we change the orientation, then we replace $\int_X \omega$ by $-\int_X \omega$. In the general case, we might not be able to find a single parametrization which covers X, but if we can cut X up into reasonable pieces which can be parametrized, and for which the orientations agree in a suitable sense, then we can still define $\int_X \omega$ and it is independent of the choices made, as long as we are consistent about the orientation.

We now state the main theorems of vector calculus, in the language of vector analysis and of differential forms.

Theorem 11.33 (Green's Theorem, 1828)**.** *Let $D \subseteq \mathbb{R}^2$ be a region, in other words a compact smooth 2-manifold with piecewise smooth boundary in $\mathbb{R}^2$. Orient ∂D so that, as you travel along ∂D, standing on the side of the surface given by the choice of the normal, the region D is to the left. Let $\mathbf{F} = (F_1, F_2)$ be a C^∞ vector field on D. Then*

$$\int_{\partial D} \mathbf{F} = \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx \wedge dy.$$

In differential forms notation, if $\mathbf{F}$ corresponds to the 1-form $\omega = F_1 dx + F_2 dy$, then

$$\int_{\partial D} \omega = \iint_D d\omega.$$

The following result is universally known as Stokes' theorem, but it was first stated by Lord Kelvin (William Thomson) in a letter to Stokes in 1850, and was later posed by Stokes on a competitive examination for students at Cambridge:

Theorem 11.34 (Stokes' Theorem)**.** *Let $\Sigma \subseteq \mathbb{R}^3$ be a compact surface with boundary, in other words a 2-manifold with boundary in $\mathbb{R}^3$. Suppose that Σ has been oriented by the continuous choice of a normal. Orient $\partial\Sigma$ so that, as you travel along $\partial\Sigma$, the surface Σ is to the left. Let $\mathbf{F} = (F_1, F_2, F_3)$ be a C^∞ vector field on $\mathbb{R}^3$. Then*

$$\int_{\partial\Sigma} \mathbf{F} = \iint_\Sigma \nabla \times \mathbf{F}.$$

In differential forms notation, if $\mathbf{F}$ corresponds to the 1-form $\omega = F_1 dx + F_2 dy + F_3 dz$, then

$$\int_{\partial\Sigma} \omega = \iint_\Sigma d\omega.$$

It is easy to check that Green's theorem is a special case of Stokes' theorem: view $\mathbb{R}^2$ as a surface in $\mathbb{R}^3$, oriented by taking the upward pointing normal $\mathbf{N} = \mathbf{e}_3$, and extend $\mathbf{F} = (F_1, F_2)$ to a vector field on $\mathbb{R}^3$ by taking $\mathbf{F}(x, y, x) = (F_1(x, y), F_2(x, y), 0)$. The formula in Stokes' theorem then gives the formula in Green's theorem. Conversely, for a parametrized surface $H\colon D \to \Sigma$, Stokes' theorem follows from Green's theorem by using the formula $H^*(d\omega) = dH^*(\omega)$.

Lastly there is the Divergence Theorem. Some cases of this theorem were stated by Gauss as early as 1813, and a version was also stated by Ostrogradsky in 1826.

Theorem 11.35 (Divergence Theorem)**.** *Let R be a region in $\mathbb{R}^3$, in other words a 3-manifold with boundary. Orient ∂R by the outward normal, in other words the one which points away from R. Let $\mathbf{F} = (F_1, F_2, F_3)$ be a C^∞ vector field on $\mathbb{R}^3$. Then*

$$\iint_{\partial R} \mathbf{F} = \iiint_R \nabla \cdot \mathbf{F}.$$

In differential forms notation, if $\mathbf{F}$ corresponds to the 2-form $\omega = F_1 dy \wedge dz + F_2 dz \wedge dx + F_3 dx \wedge dy$, then

$$\iint_{\partial R} \omega = \iiint_R d\omega.$$

For example, if R is the region enclosed by two concentric spheres, then ∂R is oriented as follows: for the outer sphere, we take the normal which points away from the center, but for the inner sphere we take the normal pointing toward the center.

The fact that all of these theorems look essentially the same when written out in differential forms notation suggests that there is one common result:

Theorem 11.36 (Generalized Stokes' Theorem)**.** *Let $X \subseteq \mathbb{R}^n$ be a k-manifold with boundary ∂X. Suppose that X is orientable. Let ω be a $(k-1)$-form in $\mathbb{R}^n$. Then there are consistent choices of orientations for X and ∂X so that*

$$\int_{\partial X} \omega = \int_X d\omega.$$

The following is a corollary of Stokes' theorem:

Proposition 11.37. *If ω is exact, then $\displaystyle\int_X \omega$ only depends on ∂X.*

Proof. Suppose that $\omega = d\varphi$. If X_1 and X_2 are two manifolds with boundary such that $\partial X_1 = \partial X_2$, then

$$\int_{X_1} \omega = \int_{X_1} d\varphi = \int_{\partial X_1} \varphi = \int_{\partial X_2} \varphi = \int_{X_2} d\varphi = \int_{X_2} \omega.$$

□

In case X is closed, in other words $\partial X = \emptyset$, then, using the fact that the integral over the empty set of a k-form is always zero (by logic or convention or by approximation), Stokes' theorem gives the following:

Proposition 11.38. *If X is a smooth compact oriented k-manifold such that $\partial X = \emptyset$, then, for all $(k-1)$-forms ψ, $\int_X d\psi = 0$.*

We shall just give a proof of the special case of Green's theorem where D is a rectangle $[a, b] \times [c, d]$. In this case, let $\omega = F_1 dx + F_2 dy$. Consider for example the term $\int_{\partial D} F_1 dx$. Now ∂D has 4 pieces:

1. $\mathbf{r}_1(t) = (x, c), x \in [a, b]$;
2. $\mathbf{r}_2(t) = (b, y), y \in [c, d]$;
3. $\mathbf{r}_3(t) = (x, d), x \in [a, b]$;
4. $\mathbf{r}_4(t) = (a, y), y \in [c, d]$.

Here the first two curves are oriented consistently with our conventions for ∂D, whereas the other two have the opposite orientations. We will work out the integrals with the wrong orientations and then add in the minus signs where necessary. Now

$$\int_{\mathbf{r}_1(t)} F_1 dx = \int_a^b F_1(x, c) dx.$$

On the other hand, $\int_{\mathbf{r}_2(t)} F_1 dx = 0$ since x is constant on $\mathbf{r}_2$. Likewise,

$$-\int_{\mathbf{r}_3(t)} F_1 dx = -\int_a^b F_1(x, d) dx$$

and $\int_{\mathbf{r}_2(t)} F_1 dx = 0$ as before. Putting this together, we see that

$$\int_{\partial D} F_1 dx = \int_a^b F_1(x, c) dx - \int_a^b F_1(x, d) dx.$$

On the other hand,

$$\iint_D -\frac{\partial F_1}{\partial y} dxdy = \int_a^b \left\{ \int_c^d -\frac{\partial F_1}{\partial y} dy \right\} dx$$
$$= -\int_a^b (F_1(x,y)]_c^d dx = \int_a^b (F_1(x,c)F_1(x,d))dx,$$

where we have used the fundamental theorem of calculus to evaluate the integral $\int_c^d \frac{\partial F_1}{\partial y} dy$. Comparing, we see that

$$\int_{\partial D} F_1 dx = \iint_D -\frac{\partial F_1}{\partial y} dxdy.$$

A similar calculation shows that

$$\int_{\partial D} F_2 dy = \iint_D \frac{\partial F_2}{\partial x} dxdy.$$

This proves Green's theorem for a rectangle. Once the theorem has been proven for a rectangle, the idea is that it follows for more general regions D by cutting D up into many pieces which are diffeomorphic to rectangles or to pieces of rectangles and then by using the property that $H^*(d\omega) = dH^*(\omega)$.

We end with one final comment. The Poincaré lemma guarantees that, for a convex open subset of $\mathbb{R}^n$, if ω is a smooth k-form with $d\omega = 0$, then there exists a smooth $(k-1)$-form ψ such that $\omega = d\psi$. But this fails to hold if U is not convex. One way to check that a closed form ω with $d\omega = 0$ is **not** exact, in other words is not of the form $d\psi$, is to find a compact manifold X with $\partial X = \emptyset$ such that $\int_X \omega \neq 0$. For, in this case, if $\omega = d\psi$, then by Proposition 11.38, we would have

$$\int_X \omega = \int_X d\psi = 0,$$

a contradiction. For example, generalizing the 1-form $\frac{-y\,dx}{x^2+y^2} + \frac{x\,dy}{x^2+y^2}$ in $\mathbb{R}^2 - \{\mathbf{0}\}$, for all n, there exists an $(n-1)$-form ω defined on $\mathbb{R}^n - \{\mathbf{0}\}$ such that $d\omega = 0$ but $\int_{S^{n-1}} \omega \neq 0$. Intuitively, we think that removing the origin from $\mathbb{R}^n$ has introduced an $(n-1)$-dimensional "hole" (of dimension $n-1$ because it is detected by the $(n-1)$-sphere), and it is the presence of this hole which prevents the closed form ω from being exact. One can also

show that, for all $k \neq n-1$ and every k-form φ such that $d\varphi = 0$, there exists a $(k-1)$-form α such that $d\alpha = \varphi$. In other words, $\mathbb{R}^n - \{\mathbf{0}\}$ has just one hole and it only appears in dimension $n-1$. This idea leads in two different but related directions. First, we can try to use the failure of the Poincaré lemma to measure the "holes" in an open subset U of $\mathbb{R}^n$, or more generally, giving us a tool to quantify the topology of U. Second, we can see that the topology of U influences when certain systems of equations such as $d\alpha = \varphi$ (viewed as a set of equations for the unknown $(k-1)$-form α) have a solution.