A COURSE ON ADVANCED CALCULUS

JASON MURPHY

Introduction. The basic purpose of an 'advanced calculus' course is to put the concepts of differential and integral calculus on a firm mathematical footing. To give context for such an undertaking, let us briefly attempt to put ourselves in the setting of mathematics in the early 1800s. At this point, the techniques of calculus (introduced largely in the late 1600s) had proven to be enormously effective in solving both physical and mathematical problems. At the same time, it was clear that many of the basic notions of calculus were rather shakily defined; accordingly, attempts at proofs of various results (known 'intuitively' to be correct) often contained serious gaps. In addition, the aggressive application of calculus techniques to new physical problems had led to solutions that challenged mathematicians' understanding, leading to serious disagreements over whether or not these solutions were acceptable. Such challenges led mathematicians of the 1800s to revisit the basic notions of calculus and to develop the subject anew in such a way that they were able not only to establish rigorously all of the essential results in calculus, but also to discover many mathematical phenomena that had never been imagined before. In fact, this program ultimately led to an investigation of the very foundations of mathematics, including our understanding of the real number system and of the infinite in general. The first semester of this course is a rigorous presentation of differential and integral calculus in the setting of real-valued functions on the line; the presentation is inspired by the historical development of the relevant ideas. The second semester of this course primarily concerns the extension of differential and integral calculus to the higher-dimensional setting.

These notes were prepared using several different sources. For the first semester content, the biggest influence was 'A Radical Approach to Real Analysis' by D. Bressoud; for the second semester content, the biggest influence was 'Advanced Calculus of Several Variables' by C. H. Edwards, Jr.

Contents

1. Techniques of calculus and some motivating examples

The techniques of calculus (e.g. derivatives, antiderivatives, and series expansions) were developed in the late 1600s to address a range of problems, including the computation of instantaneous velocities; optimization problems; computing tangents to curves; and the computation of lengths, areas, volumes, and centers of gravity. Such problems had wide applicability in areas such as optics and astronomy. In the century that followed, the techniques of calculus continued to be developed and applied to a wide range of mathematical and physical problems. In this section, we will look at several representative applications of calculus techniques that bring to light some of the mathematical issues that were ultimately resolved through the work of Cauchy and others throughout the 19th century.

1.1. Original notions of derivatives and integrals.

As students of calculus know, the definition of the derivative depends on the notion of a *limit.* Calculus students may at least be *shown* the formal definition of a limit (the 'ε-δ definition'), although it typically plays no real role in a first calculus course. In fact, this definition was not introduced until the work of Cauchy in the 1820s. So, what did Newton and Leibniz think a derivative was?

Newton and Leibniz used different language and had somewhat different perspectives, but ultimately they were both trying to get at the notion of an instantaneous rate of change of one quantity with respect to another. Newton used the language of fluents, fluxions, moments, and ultimate ratios, while Leibniz worked with infinitesimals and differentials. Neither was able to write down fully satisfactory definitions, but between the two of them they succeeded in establishing the algebraic rules of calculus, developing the important connection between areas and antiderivatives, and applying the techniques of calculus to many important mathematical and physical problems.

It is not our goal in this course to try to understand the original works of Newton and Leibniz. Nonetheless, it may be interesting to consider a few simple examples that demonstrate some of the early thinking about concepts in calculus. For example, here is an argument 'in the style of Newton' that if $y = x^2$, then $\frac{dy}{dx} = 2x$.

Example 1.1. Let $y = x^2$. Now let x 'flow' to $x + o$. Then

$$(x + o)^2 = x^2 + 2ox + o^2.$$

Thus, the increases of x and y are to one another as

$$o \quad \text{to} \quad 2ox + o^2, \quad \text{or equivalently} \quad 1 \quad \text{to} \quad 2x + o.$$

Thus, letting the increment vanish, the last proportion will be 1 to $2x$. Hence the 'fluxion' of x is to the 'fluxion' of x^2 as 1 to $2x$.

Here is an abbreviated presentation of what amounts to the product rule, as given in Book 2, Section 2, Lemma 2 of Newton's *Principia*:

Example 1.2. Suppose A and B are increasing or decreasing by continual motion, and that their 'moments' (that is, their 'instantaneous increment or decrement') are given by a and b, respectively. Then the moment of the generated rectangle AB is $aB + bA$.

Proof. When the halves of the moments (i.e. $\frac{1}{2}a$ and $\frac{1}{2}b$) are lacking from the sides of A and B, the rectangle is given by

$$(A - \tfrac{1}{2}a)(B - \tfrac{1}{2}b) = AB - \tfrac{1}{2}aB - \tfrac{1}{2}bA + \tfrac{1}{4}ab.$$

As soon as the sides A and B have been increased by the other halves of the moments, we obtain

$$(A + \tfrac{1}{2}a)(B + \tfrac{1}{2}b) = AB + \tfrac{1}{2}aB + \tfrac{1}{2}bA + \tfrac{1}{4}ab.$$

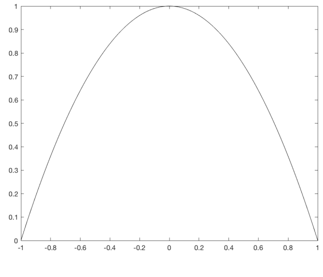
Subtracting the first rectangle from the second, there remains an excess of $aB+bA$. That is, a total increment of a and b to the sides generates an increment of $aB+bA$ of the rectangle. □

Along with derivatives, the other central topic of any calculus sequence is that of integration. In the early development of calculus, integration was essentially identified with antidifferentiation. That is,

$$\int_a^x f(t)\,dt \tag{1.1}$$

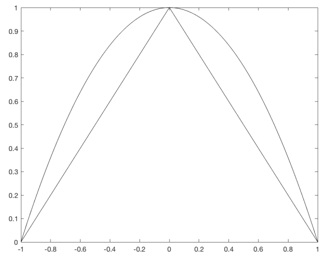
simply meant a function whose derivative was $f(x)$. At the same time, as mentioned above, mathematicians appreciated the fact that the antiderivative (1.1) also represented the area under the curve of f for $t \in [a, x]$ (this is essentially the 'Fundamental Theorem of Calculus'). The fact that a given function might not have an antiderivative seemed to be no real bother. For example, to evaluate (1.1) one could simply write a power series expansion for f and integrate term by term (we will return to this below). It was not until the work of Cauchy and Riemann in the 1800s that the integral came to be *defined* to represent the area under the curve.

1.2. **Geometric series and the Archimedean notion of convergence.** The following simple example serves to introduce several key concepts. We consider the problem of computing the area under the inverted parabola $y = 1 - x^2$, as shown in the following figure:



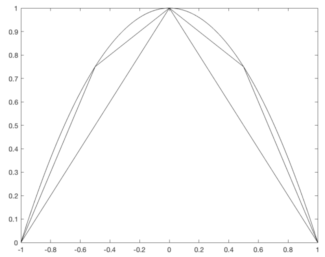
We follow an argument of Archimedes (from the 200s BC) (which is similar to arguments of Eudoxos from the 300s BC) to approximate this area by inscribing triangles under this curve, as follows:

First, inscribe the triangle with vertices at $(-1, 0)$, $(0, 1)$, and $(1, 0)$:



We readily compute the area of this triangle to be $A_0 = 1$. We take this as our first 'approximation' to the area under the curve.

Next, fill in two more triangles by placing x coordinates at $x = \pm\frac{1}{2}$, as follows:



Some basic geometry reveals that the total area under these two triangles is 1/4, so that our second approximation to the area is

$$A_1 = 1 + \tfrac{1}{4}.$$

The next step is to fill in four more triangles, placing x coordinates at $x = \pm\frac{1}{4}$ and $x = \pm\frac{3}{4}$. Some more geometry will show that the total area added is $\frac{1}{16}$, giving a new approximation to the area of

$$A_2 = 1 + \tfrac{1}{4} + \tfrac{1}{16}.$$

If we continue this process, then we can demonstrate that at each stage we will add an area equal to $\frac{1}{4}$ of the area added in the previous stage. (This is not supposed to be obvious, but it is true and it can be justified using geometry.) We can therefore arrive at a sequence of approximations to the total area, given by

$$A_n = 1 + \tfrac{1}{4} + \tfrac{1}{16} + \cdots + \tfrac{1}{4^n}, \quad \text{also written} \quad A_n = \sum_{j=0}^{n} \tfrac{1}{4^j}.$$

Now, one can show that

$$A_n = \sum_{j=0}^{n} \tfrac{1}{4^j} = \tfrac{4}{3} - \tfrac{1}{3\cdot 4^n} \quad \text{for any} \quad n \tag{1.2}$$

(see Exercise 1.1). From our modern viewpoint, it now seems perfectly reasonable to 'send $n \to \infty$' and declare that the area under the curve is given by the infinite series

$$\sum_{j=0}^{\infty} \tfrac{1}{4^j}, \quad \text{which equals} \quad \tfrac{4}{3}. \tag{1.3}$$

However, this was *not* the approach of Archimedes. Instead, he argued as follows: writing A for the area under the curve, he showed that neither $A < \frac{4}{3}$ nor $A > \frac{4}{3}$ is possible. To do this, he showed by geometric arguments that the process of adding additional triangles as above always reduces the uncovered area by at least half. Thus if $A > \frac{4}{3}$, we can eventually inscribe enough triangles so that $A_n > \frac{4}{3}$; however, this contradicts (1.2), which clearly shows $A_n < \frac{4}{3}$ for each n. On the other hand, if $A < \frac{4}{3}$, then we can find an n large enough that $A_n > A$ (again by (1.2)); however, this contradicts that we always have $A_n \leq A$ by construction (since we are inscribing triangles under the curve).

In the particular case of the 'geometric' series (1.3), it is relatively straightforward to give an interpretation of the 'value' of this infinite series, motivated by the approach just described. In particular, we may make the following definition:

Definition 1.1 (Value of an infinite series)**.** Let a_j be a sequence of real numbers. We say that the infinite series $\sum_{j=0}^{\infty} a_j$ equals L if for any $L_0 < L < L_1$, there exists an N such that

$$n \geq N \implies L_0 < \sum_{j=0}^{n} a_j < L_1.$$

As we will see, this definition states that the value of an infinite series is equal to the limit of the partial sums of the series (provided that the limit exists).

Early works on calculus concerning the calculation of areas could be viewed (or defended) as merely providing a short-hand for a complicated Archimedean argument of the type above. The techniques of calculus do indeed greatly simplify the problem above (and lead to the correct answer): Supposing we know that (i) $\int x^k = \frac{x^{k+1}}{k+1}$ and (ii) integrals represent areas under curves, then the desired area is given by

$$A = \int_{-1}^{1} (1 - x^2)\, dx = [x - \tfrac{1}{3}x^3]\Big|_{x=-1}^{1} = \tfrac{4}{3}.$$

1.3. **Newton's binomial series and applications.** The coefficients appearing in the expansion of expressions such as

$$(a + b)^n \tag{1.4}$$

are known as the *binomial coefficients.* They arise frequently in mathematical problems arising in algebra, combinatorics, probability theory, and other areas. These coefficients may be read off from the following diagram, often called *Pascal's triangle* (although he was not actually the first to work out binomial coefficients or even to write down such a diagram):

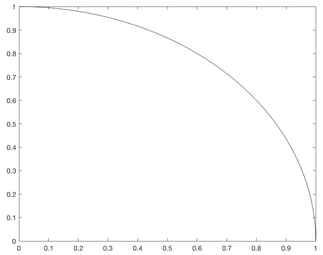
$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

Thus, for example,

$$(a + b)^3 = a + 3ab^2 + 3a^2b + b^3.$$

Newton was interested in finding an analogous expansions for expressions such as (1.4) in the case that n was not necessarily a positive integer. One reason for doing this was an application to computing digits of π. For example, using the geometric interpretation of the integral (see the figure below), we may write

$$\int_0^1 (1-t^2)^{\frac{1}{2}}\,dt = \tfrac{1}{4}\cdot(\text{area of the unit circle}) = \tfrac{\pi}{4}. \tag{1.5}$$



One quarter of the unit circle is given by the curve $y = (1-t^2)^{1/2}$, with $t \in [0,1]$.

Thus, if we had a representation of the function $(1-t^2)^{\frac{1}{2}}$ in powers of t, we could integrate as many terms of the series as desired in order to construct series approximations to the value of π.

We are therefore led to the problem of constructing a *power series representation* for the function $(1-t^2)^{\frac{1}{2}}$. As a matter of fact, Newton was able to derive something even more general:

Proposition 1.1 (Newton's binomial series)**.** *For any $a \in \mathbb{R}$ and $x \in \mathbb{R}$ with $|x| < 1$,*

$$(1+x)^a = 1 + ax + \tfrac{a(a-1)}{2!}x^2 + \tfrac{a(a-1)(a-2)}{3!}x^3 + \cdots$$

In particular, when a is an integer, the right-hand side reduces to a finite sum and we recover the usual binomial expansion. However, when a is not an integer, the right-hand side becomes an infinite series of functions of x, which at this point is not something we understand particularly well.

At any rate, we can *use* this expansion (with $a = \frac{1}{2}$ and $x = -t^2$) to approximate π. Integrating term by term, we have:

$$\begin{aligned} \tfrac{\pi}{4} &= \int_0^1 (1-t^2)^{\frac{1}{2}}\,dt \\ &= \int_0^1 1 - \tfrac{1}{2}t^2 - \tfrac{1}{2^2\cdot 2!}t^4 - \tfrac{3}{2^3\cdot 3!}t^6 - \tfrac{3\cdot 5}{2^4\cdot 4!}t^8 - \cdots\,dt \\ &= 1 - \tfrac{1}{2\cdot 3} - \tfrac{1}{2^2\cdot 2!\cdot 5} - \tfrac{3}{2^3\cdot 3!\cdot 7} - \tfrac{3\cdot 5}{2^4\cdot 4!\cdot 9} - \ldots \end{aligned}$$

In particular, this yields the following sequence of (crude) approximations to π (where we keep the first four decimals only):

$$4,\ 3.3333,\ 3.2333,\ 3.1976,\ 3.1803,\ 3.1703,\ 3.1640,\ 3.1597,\ 3.1566,\ 3.1543,\ \ldots$$

Using enough terms of the series, we should in theory be able to approximate π as accurately as we wish. In fact, although this approximation is very crude, Newton had much more sophisticated versions of this approach that yielded much better series approximations to π. However, this simple example hopefully suffices to illustrate the general idea. This example also leads us to some fundamental questions:

(i) How can we compute the power series that represents a given function?
(ii) Given a power series representation for a given function, is it actually permitted to integrate the series term by term in order to integrate the given function? What about the other operations in algebra and calculus (like taking sums, products, derivatives, and so on)?

Students of calculus already know the answer to some of these questions, for example, the fact that we compute the coefficients of the power series expansion of a function in terms of the higher order derivatives of the function (Taylor series). Similarly, mathematicians working in the early days of calculus understood the general procedure for computing such expansions. The other questions, however, proved to be quite subtle and were not settled until well into the 1800s.

1.4. **An infinitely differentiable function with no power series expansion.** In the early history of calculus, there was basically no distinction between the notion of a 'function' and that of an 'analytic function' (that is, one admitting a power series representation). That is to say, functions were implicitly assumed to admit power series representations. The general prescription for computing such power series was first put in print by Taylor in 1715 (although the derivation was based on an interpolation formula found before by both Newton and Gregory). This is the familiar 'Taylor series' formula:

$$f(x) = f(a) + f'(a)(x-a) + \tfrac{f''(a)}{2!}(x-a)^2 + \tfrac{f'''(a)}{3!}(x-a)^3 + \cdots , \tag{1.6}$$

where 'primes' denote derivatives (still somewhat murkily defined at this point). In fact, if any series expansion of the form

$$f(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots$$

should hold, one can see fairly quickly that the coefficients should be given by $c_k = \frac{f^{(k)}(a)}{k!}$ (see Exercise 1.2).

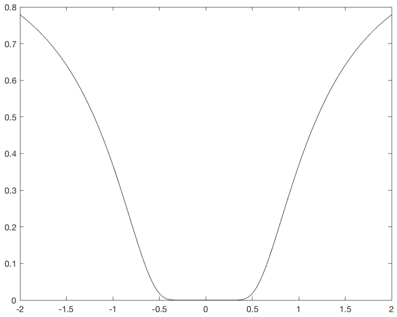
Understanding the sense in which the identity (1.6) holds and the notion of convergence in general is a subtle problem that mathematicians did consider in the 1700s. Of particular interest was the problem of establishing bounds on the difference between a function and its Taylor polynomial approximations. In the meantime, however, mathematicians were happy to use series expansions freely to solve a wide variety of problems (for example, by constructing power series solutions to differential equations arising in physics).

Unfortunately, the following example of Cauchy shows that one must give up the hope that *all* functions (even 'nice' or 'continuous' ones) actually admit power series representations. This was a big problem for those (like Lagrange) who had hoped to base their very definition of derivatives on the power series expansion.

Example 1.3. Define the function

$$f(x) = \begin{cases} e^{-1/x^2} & x \neq 0 \\ 0 & x = 0, \end{cases}$$

as pictured below:

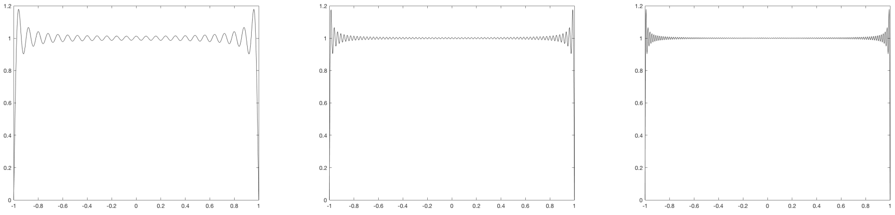


In general, one finds that any derivative of f is of the form $e^{-1/x^2}P(1/x)$ for some polynomial P. In particular, by using L'Hospital's rule, we can establish that all derivatives of f exist and equal zero at $x=0$. However, this means that f *cannot* admit a power series representation at $x=0$, for every coefficient of the series would equal zero!

1.5. **Trigonometric series solutions.** In addition to the power series solutions mentioned above, mathematicians began to introduce other series solutions to physical problems, including *trigonometric series* involving sums of sines and cosines of higher and higher frequencies. One such problem was that of heat propagation through a thin metal sheet, which was investigated by Fourier in the early 1800s. He considered a problem in which the temperature was held equal to zero on two sides of a sheet (at $x=\pm1$, say) and equal to one at one base of the sheet (on the interval $[-1,1]$, say). The general solution to the underlying physical model involved products of exponentially decaying functions and (in this case) cosines. In particular, to produce a solution in this particular case required that Fourier express the constant function 1 as an infinite series of cosines. Remarkably, he arrived at the solution

$$1=\tfrac{4}{\pi}\big[\cos(\tfrac{\pi x}{2})-\tfrac{1}{3}\cos(\tfrac{3\pi x}{2})+\tfrac{1}{5}\cos(\tfrac{5\pi x}{2})+\cdots\big],\quad x\in(-1,1), \tag{1.7}$$

and he even provided a general method for computing the coefficients in such series expansions. In fact, (1.7) looks like it works reasonably well:



Approximations to the constant function 1 using (1.7) with 25, 75, and 125 terms.

Originally, however, (1.7) was *not* accepted as a legitimate solution to Fourier's problem. In fact, there are several potential objections to this solution (beyond just a general unwillingness to even consider an infinite sum of cosines). First, observing that the general term of the series is

$$\tfrac{4}{\pi}\tfrac{(-1)^{n-1}}{2n-1}\cos\big[\tfrac{(2n-1)\pi x}{2}\big], \tag{1.8}$$

we can see that the general term is roughly of size $1/n$. However (as we will see in the next section), the infinite series with general term $1/n$ is divergent! Second, if this series is meant to represent the constant function 1, then its derivative should certainly equal zero. Using (1.8), we should therefore have

$$0 = 2\sum_{n=1}^{\infty}(-1)^n \sin\left[\tfrac{(2n-1)\pi x}{2}\right], \quad x \in (-1,1).$$

However, here the general coefficients do not decay at all, so how can the series possibly converge? In fact, if we imagine trying to evaluate at or near the endpoint $x = -1$ (say), then all of the summands tend to one and it seems that the series should diverge.

Nonetheless, Fourier's ideas ultimately won out. Before this could happen, however, many fundamental concepts of calculus had to be completely reimagined. Indeed, the controversy surrounding Fourier's trigonometric series was a catalyst for much of the progress that was to be made throughout the 1800s.

1.6. **The harmonic series and the existence of limits.** The final example we will consider in this section is the *harmonic series*

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \cdots + \tfrac{1}{n} + \cdots,$$

which we already mentioned in the previous section. Although the individual summands tend to zero, we can show that the partial sums

$$S_n = 1 + \tfrac{1}{2} + \cdots + \tfrac{1}{n}, \quad \text{also written} \quad \sum_{j=1}^{n} \tfrac{1}{j}$$

do not converge to any finite value in the sense of Definition 1.1. One quick way to see this is due to the 17th century mathematician Mengoli: Since

$$\tfrac{1}{3j-1} + \tfrac{1}{3j} + \tfrac{1}{3j+1} = \tfrac{27j^2-1}{27j^3-3j} > \tfrac{1}{j},$$

we see that

$$1 + \left[\tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4}\right] + \cdots + \left[\tfrac{1}{3n-1} + \tfrac{1}{3n} + \tfrac{1}{3n+1}\right] > 1 + [1] + \cdots + \left[\tfrac{1}{n}\right].$$

Thus, we deduce that

$$S_{3n+1} > 1 + S_n, \tag{1.9}$$

which one can check is incompatible with the notion of convergence in Definition 1.1.

In fact, (1.9) actually shows that the partial sums S_n 'increase without bound'. Indeed, noting that $S_{4n} > S_{3n+1}$, we can use (1.9) to obtain the lower bound

$$S_{4^m} \geq m \quad \text{for} \quad m \geq 1.$$

This leads to the following definition:

Definition 1.2 (Divergent series)**.** A series of positive terms a_j *diverges* if for any $M > 0$, there exists N such that

$$n \geq N \implies \sum_{j=1}^{n} a_j > M.$$

We may write $\sum_{j=1}^{\infty} a_j = \infty$.

It is worth thinking for a moment about whether there could be any other alternatives beyond convergence in the sense of Definition 1.1 or divergence in the sense of Definition 1.2, at least for the case of positive series. We may equivalently ask the following: Suppose a series of positive terms a_j does *not* diverge, so that there exists some fixed $M > 0$ that bounds *all* of the partial sums S_n. Is the infinite series then guaranteed to have some value (in the sense of Definition 1.1)?

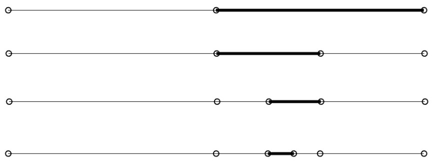
This turns out to be a tricky question. If we believe that the series converges, then we need to find a way to identify its value. We may attempt to do so as follows: Let S_n denote the partial sums, and suppose $S_n < M$ for each n.

We first observe that there must exist some integer N_0 large enough that

$$n \geq N_0 \implies S_n < S_{N_0} + 1.$$

Indeed, if this were false, then we could find a sequence of partial sums that increase by at least 1 between subsequent terms. In this case, however, there would be no way that the partial sums remain bounded by M.

With this observation in mind, let us change our assumptions slightly and just assume that $0 < S_n < 1$ for all n (this just amounts to fixing the sum up to S_{N_0} and only considering the rest of the series). Now let us argue as follows. Because the partial sums can only increase, either all of the partial sums are contained in $(0, \frac{1}{2})$, or eventually the partial sums are all contained the interval $[\frac{1}{2}, 1)$. In the first case, we can repeat this argument to see that either all of the sums are contained in $(0, \frac{1}{4})$, or eventually they reach $[\frac{1}{4}, \frac{1}{2})$. Similarly, in the second case, we see that eventually the partial sums are all contained in $(\frac{1}{2}, \frac{3}{4})$ or $[\frac{3}{4}, 1)$. Repeating this argument, we can deduce the following: there exists a *nested* sequence of intervals of the form $I_j = [k_j 2^{-j}, (k_j + 1)2^{-j}]$, where $0 \leq k_j < 2^{-j}$, and an increasing sequence of integers N_j such that $S_n \in I_j$ for all $n > N_j$. Here *nested* refers to the fact that I_{j+1} is a subset of I_j. The first few steps in a typical situation are depicted in the following figure (where the darkened intervals show where the partial sums are eventually 'trapped'):



It now seems pretty clear that the partial sums S_n must be getting 'squeezed' towards some particular value. But how do we prove it? As it turns out, this relies on a fundamental property of the real numbers called *completeness*. To show that this property holds actually requires that we stop and think about how we really define the real numbers in the first place. It was not until the late 1800s that mathematicians such as Dedekind and Cantor provided rigorous constructions of the real number system that guaranteed this completeness property.

To drive this point home, let us end with a simple example that shows how completeness may *fail*.

Example 1.4 (The rational numbers are not complete)**.** We construct a sequence of numbers of the form $\frac{a}{b}$, where the first term has $a = b = 1$ and subsequent terms are defined via

$$\tfrac{a}{b} \mapsto \tfrac{3a+4b}{2a+3b}. \tag{1.10}$$

In particular, each term in this sequence is a rational number. We can also show that this sequence is increasing and bounded. Boundedness is straightforward, since

$$\tfrac{3a+4b}{2a+3b} = \tfrac{3a+(9/2)b}{2a+3b} - \tfrac{(1/2)b}{2a+3b} < \tfrac{3}{2} \quad \text{for all} \quad a, b > 0.$$

To see that the sequence is increasing, we need to show

$$\tfrac{3a+4b}{2a+3b} > \tfrac{a}{b}, \quad \text{which requires} \quad 2b^2 > a^2.$$

For this, we observe that this property is initially true (when $a = b = 1$), and we show that it is preserved under the rule (1.10). Indeed, since

$$2(2a+3b)^2 = 8a^2 + 18b^2 + 24ab, \quad \text{while} \quad (3a+4b)^2 = 9a^2 + 16b^2 + 24ab,$$

we may observe that

$$2b^2 > a^2 \implies 2(2a+3b)^2 > (3a+4b)^2.$$

Now we are in a similar position to the discussion above, that is, we have a bounded, increasing sequence whose limiting value we would like to identify.

To this end, let us write the sequence in the form $\frac{a_n}{b_n}$ and suppose this sequence converges to a limit L. Then, the shifted sequence $\frac{a_{n+1}}{b_{n+1}}$ also converges to L. However,

$$\frac{a_{n+1}}{b_{n+1}} = \frac{3a_n + 4b_n}{2a_n + 3b_n} = \frac{3\frac{a_n}{b_n} + 4}{2\frac{a_n}{b_n} + 3} \quad \text{must converge to} \quad \frac{3L+4}{2L+3}.$$

Thus we must have

$$L = \tfrac{3L+4}{2L+3}, \quad \text{which requires} \quad L^2 = 2.$$

However it has been known since the time of the ancient Greeks that there is no *rational* number satisfying $L^2 = 2$ (see Exercise 1.3). In particular, we have demonstrated a bounded, increasing sequence of rational numbers with no *rational* limit.

1.7. **Exercises.**

Exercise 1.1. Let $a > 0$. Derive a formula for the sum

$$1 + a + a^2 + \cdots + a^n, \quad \text{also denoted} \quad \sum_{j=0}^{n} a^j.$$

Exercise 1.2. Suppose that

$$f(x) = c_0 f(a) + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \cdots$$

Show that $c_k = \frac{f^{(k)}(a)}{k!}$, where $f^{(k)}$ denotes the k^{th} derivative of f.

Exercise 1.3. Show that $\sqrt{2}$ is irrational. (*Hint:* Suppose $\sqrt{2} = p/q$ where p and q are relatively prime integers, with $q > 0$. Then show that p and q must be even, contradicting that they were chosen to be relatively prime.)

2. Differentiability and continuity

2.1. **Limits and differentiability.**

Our starting point for the rigorous development of calculus is to introduce Cauchy's definition of the derivative. This definition relies on the notion of a *limit*, which is similar to the 'Archimedean' notion of the value of an infinite series (see Definition 1.1).

Before stating the definition, let us introduce one convenient technical term: we call an interval with one point removed a *punctured interval.* We denote this by $I\backslash\{a\}$, where a is the missing point.

Definition 2.1 (Limit, function version)**.** Let f be a real-valued function defined on some punctured interval $I\backslash\{a\}$. We say that $f(x)$ *has limit* ℓ *as* x *approaches* a, written

$$\lim_{x\to a} f(x) = \ell, \tag{2.1}$$

if for any $\ell_1 < \ell < \ell_2$, we have

$$\ell_1 < f(x) < \ell_2 \quad \text{for all} \quad x \quad \text{sufficiently close to} \quad a.$$

A few remarks are in order. First, note that we only require f to be defined on the punctured interval, not at the point a itself. Next, there are many ways to denote (2.1). For example we may write

$$f(x) \to \ell \quad \text{as} \quad x \to a, \quad \text{or} \quad f(x) \xrightarrow{x\to a} \ell.$$

Finally, it is sufficient to take $\ell_1 = \ell - \varepsilon$ and $\ell_2 = \ell + \varepsilon$ for some $\varepsilon > 0$, and it is also convenient to quantify what we mean by 'sufficiently close' by introducing another parameter $\delta > 0$. In particular, we have the alternate version of Definition 2.1 above, which is the the version that we will really use in what follows:

Definition 2.2 (Limit, alternate version)**.** We say that

$$\lim_{x\to a} f(x) = \ell$$

if for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$0 < |x-a| < \delta \implies |f(x) - \ell| < \varepsilon.$$

There is a related notion of the limit of a sequence of numbers, which we will discuss below. In the exercises, you will work out some standard algebraic 'limit laws', which should help you get used to working with this 'ε-δ' definition. You may find the following theorem convenient to use, as well:

Theorem 2.1 (The '$C\varepsilon$' Theorem)**.** *Suppose that there exists $C > 0$ so that for all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$0 < |x-a| < \delta \implies |f(x) - \ell| < C\varepsilon.$$

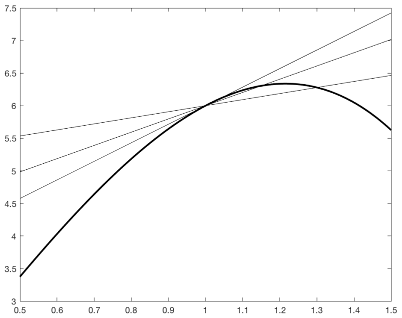
Then $\lim_{x\to a} f(x) = \ell$.

Proof. Let C be as in the statement of the theorem. Now choose any $\varepsilon > 0$. By assumption, we may choose a $\delta > 0$ (corresponding to the positive number $\frac{\varepsilon}{C}$) so that

$$0 < |x-a| < \delta \implies |f(x) - \ell| < C \cdot \tfrac{\varepsilon}{C} = \varepsilon.$$

As $\varepsilon > 0$ was arbitrary, we have satisfied the condition required by Definition 2.2. □

Let us finally introduce the formal definition of the derivative. It is defined as the limit of the slopes of the secant lines to a curve at a given point (as depicted in the following figure):



Definition 2.3. Let f be a real-valued function defined on an interval I containing a point a. We say that f is *differentiable at $x = a$ with derivative ℓ* if

$$\lim_{x\to a} \frac{f(x) - f(a)}{x - a} = \ell \tag{2.2}$$

We write $\ell = f'(a)$, or alternately $\ell = \frac{df}{dx}\big|_{x=a}$.

Note that the ratio in (2.2) is defined in the punctured interval $I\backslash\{a\}$, and hence it makes sense to speak of its limit as $x \to a$. We may equivalently express the limit in (2.2) as

$$\lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = \ell.$$

With these definitions in mind, let us revisit the two examples originally presented in the style of Newton in Section 1.1.

Example 2.1. Let $f(x) = x^2$. Then

$$(x+h)^2 = x^2 + 2xh + h^2,$$

which yields.

$$\frac{(x+h)^2 - x^2}{h} = 2x + h \quad \text{for all} \quad h \neq 0.$$

Thus

$$f'(x) = \lim_{h\to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h\to 0}(2x+h) = 2x \quad \text{for each} \quad x \in \mathbb{R}.$$

Example 2.2. Suppose f and g are both differentiable at $x = a$. Then the product fg is differentiable at $x = a$, with

$$\tfrac{d(fg)}{dx}\big|_{x=a} = f'(a)g(a) + g'(a)f(a).$$

Proof. Let us define the functions $E_f(h)$ and $E_g(h)$ via

$$f(a+h) = f(a) + hf'(a) + E_f(h) \quad \text{and} \quad g(a+h) = g(a) + hg'(a) + E_g(h), \tag{2.3}$$

respectively. By definition of differentiability, we then have

$$\lim_{h\to 0} \tfrac{E_f(h)}{h} = 0 \quad \text{and} \quad \lim_{h\to 0} \tfrac{E_g(h)}{h} = 0.$$

We can then expand

$$\begin{aligned} f(a+h)g(a+h) &= [f(a)+hf'(a)+E_f(h)][g(a)+hg'(b)+E_g(h)] \\ &= f(a)g(a)+h[f'(a)g(a)+f(a)g'(a)]+R(h), \end{aligned}$$

where

$$\begin{aligned} R(h) := f(a)E_g(h)+g(a)E_f(h)+h[g'(a)E_f(h) \\ + f'(a)E_g(h)]+E_f(h)E_g(h)+h^2f'(a)g'(b). \end{aligned}$$

Now, using (2.3) and the algebraic limit laws (see Exercise 2.2), we can see that

$$\lim_{h\to 0} \tfrac{R(h)}{h} = 0.$$

Thus, we deduce

$$\lim_{h\to 0} \frac{f(a+h)g(a+h)-f(a)g(a)}{h} = f'(a)g(a)+f(a)g'(a),$$

as desired. □

In the exercises, you will work out some of the other basic rules for derivatives (see Exercise 2.3). For example, the sum of two differentiable functions f and g is again differentiable, with $(f+g)'=f'+g'$. This immediately implies that any finite sum of any finite collection of differentiable functions $f_1,\dots,f_N$ is differentiable, with

$$\Big(\sum_{n=1}^N f_n\Big)' = \sum_{n=1}^N f_n'.$$

Whether or not we can extend this to the case of *infinite* series is a question we will study later.

Using the definition of the derivative (as in Example 2.1), it is straightforward to compute derivatives of some simple functions, for example:

$$\tfrac{d}{dx}[1]=0, \quad \tfrac{d}{dx}[x]=1, \quad \tfrac{d}{dx}[x^2]=2x.$$

More generally, utilizing the product rule (see Example 2.2) and proof by induction, we can establish the power rule

$$\tfrac{d}{dx}[x^n]=nx^{n-1} \quad \text{for any} \quad n=1,2,3,\dots \tag{2.4}$$

as follows:

Proof of (2.4). We know the result holds for $n=1$. Now suppose we know $\frac{d}{dx}[x^m]=mx^{m-1}$ for some integer m. Then, by the product rule,

$$\tfrac{d}{dx}[x^{m+1}] = x\tfrac{d}{dx}[x^m]+x^m\tfrac{d}{dx}[x] = x\cdot[mx^{m-1}]+x^m\cdot[1]=(m+1)x^m.$$

Therefore (by induction) the result holds for all positive integers. □

By applying the product rule, we may also compute the derivative of negative powers. Indeed, if $f(x)=x^{-n}$, then we have $x^nf(x)=1$, and we can then apply the product rule and solve for $f'(x)$. In order to compute derivatives of fractional powers (like $\sqrt{x}$, $x^{1/3}$, and so on), we will need to introduce technique known as the *chain rule*, which tells us how to compute the derivative of the composition of two functions.

Definition 2.4. The *composition* of two functions f and g is given by

$$(f \circ g)(x) = f(g(x)).$$

It is defined on the domain of g, provided the range of g is a subset of the domain of f.

Theorem 2.2 (Chain rule)**.** *Suppose $g'(a)$ and $f'(g(a))$ both exist. Then $f \circ g$ is differentiable at $x = a$ and*

$$(f \circ g)'(a) = f'(g(a))g'(a).$$

Remark 2.3. One way to *discover* this formula is to suppose that $g(a+h) = g(a) + hg'(a) + Ch^2 + \dots$ and $f(b+k) = f(b) + kf'(b) + Ck^2 + \dots$ and to work out the term in $f(g(a+h))$ that is linear in h. Once we have the right candidate for the derivative, we can prove the equality as follows:

Proof. Proceeding as in (2.2), let us define the function $E_g(h)$ via

$$g(a+h) = g(a) + hg'(a) + E_g(h), \quad \text{so that} \quad \lim_{h\to 0} \tfrac{E_g(h)}{h} = 0.$$

We now let $\varepsilon > 0$ and choose $\delta_1 > 0$ small enough that

$$0 < |k| < \delta_1 \implies \big|f(g(a)+k) - f(g(a)) - kf'(g(a))\big| < |k|\varepsilon.$$

Now, for h sufficiently small (say $0 < |h| < \delta_2$ for suitable $\delta_2 > 0$), we may guarantee

$$|hg'(a) + E_g(h)| < \delta_1. \tag{2.5}$$

In this case, we may obtain (adding and subtracting $E_g(h) \cdot f'(g(a))$) and using (2.5)):

$$\begin{aligned}
&\big|f(g(a+h)) - f(g(a)) - hf'(g(a))g'(a)\big| \\
&\quad \le \big|f(g(a) + hg'(a) + E_g(h)) - f(g(a)) - [hg'(a) + E_g(h)]f'(g(a))\big| \\
&\qquad + |E_g(h)f'(g(a))| \\
&\quad \le [|h|\,|g'(a)| + |E_g(h)|]\varepsilon + |E_g(h)|\,|f'(g(a))|.
\end{aligned}$$

Rearranging, this implies

$$\left|\frac{f(g(a+h)) - f(g(a))}{h} - f'(g(a))g'(a)\right| \le |g'(a)|\varepsilon + |\tfrac{E_g(h)}{h}|\varepsilon + |\tfrac{E_g(h)}{h}|\,|f'(g(a))|.$$

Choosing h possibly even smaller (than some $\delta_3 > 0$, say), we can therefore guarantee that

$$0 < |h| < \delta_3 \implies \left|\frac{f(g(a+h)) - f(g(a))}{h} - f'(g(a))g'(a)\right| \le C\varepsilon$$

for some $C > 0$. By the '$C\varepsilon$ Theorem', we deduce that $f \circ g$ is differentiable at $x = a$, with $(f \circ g)'(a) = f'(g(a))g'(a)$. □

With the chain rule in hand, we can compute the derivative of any rational power of x:

Example 2.3. Suppose $f(x) = x^{p/q}$ for some integers p, q. Then $f(x^q) = x^p$, and so

$$px^{p-1} = f'(x^q) \cdot qx^{q-1} \implies f'(x^q) = \tfrac{p}{q}x^{p-q} \implies f'(x) = \tfrac{p}{q}x^{\frac{p}{q}-1},$$

extending our 'power rule' to the case of rational powers.

We will extend this 'power rule' to arbitrary powers later, when we discuss exponential functions and logarithms.

Up to this point, we have established some useful rules for computing derivatives, including integer powers (and hence arbitrary polynomials). However, it is not so clear how to compute derivatives of other special functions. As a simple example, let us give a geometric derivation of the fact that $\frac{d}{dx}\sin x = \cos x$ (when x is given in radians).

Example 2.4. To find the derivative of $\sin x$, we first rely on the following trigonometric identity (which can be derived geometrically):

$$\sin(a+b) = \sin a \cos b + \sin b \cos a.$$

We can therefore write

$$\sin(x+h) = \sin x \cos h + \sin h \cos x,$$

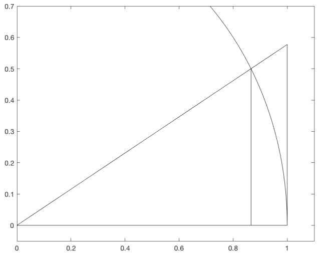
which rearranges to

$$\tfrac{\sin(x+h)-\sin(x)}{h} = \sin x \cdot \tfrac{\cos h - 1}{h} + \tfrac{\sin h}{h}\cos x, \quad h \neq 0. \tag{2.6}$$

We now claim that

$$\lim_{h\to 0} \tfrac{\sin h}{h} = 1. \tag{2.7}$$

To see this, consider the following diagram:



Comparing the heights of the two vertical line segments and the length of the arc, we can read off the inequality

$$\sin h \leq h \leq \tfrac{\sin h}{\cos h} \implies \cos h \leq \tfrac{\sin h}{h} \leq 1 \tag{2.8}$$

for non-zero h within $(-\frac{\pi}{2}, \frac{\pi}{2})$. Using the fact that $\cos h \to 1$ as $h \to 0$, we therefore derive (2.7).

Next, we claim that

$$\lim_{h\to 0} \tfrac{\cos h - 1}{h} = 0. \tag{2.9}$$

To see this, let us rely on another trigonometric identity (derived from sum formulas for sine and cosine), namely

$$\cos h - 1 = -\tfrac{\sin^2(h/2)}{1/2}.$$

Indeed, from this identity, (2.7), the fact that $\lim_{h\to 0}\sin(h/2) = 0$, and the product law for limits, we obtain (2.9).

Returning to (2.6), we now take the limit as $h \to 0$ to obtain $\frac{d}{dx}\sin x = \cos x$.

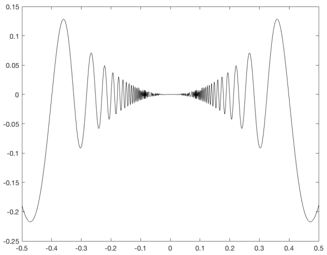
A few remarks are in order. First, you may have derived (2.7) in a calculus course by applying L'Hospital's rule and the fact that the derivative of $\sin h$ is $\cos h$. However, here we are *using* (2.7) to determine the derivative of $\sin h$, so we needed a different argument. Second, the claim (2.8) was essentially based off of looking at a diagram. By first computing areas, it is possible to establish this inequality more rigorously. We will also consider a different approach later in these notes, which entails defining $\sin x$ as a suitable power series and using that series to determine the derivative.

Let us consider one last example in which we cannot simply apply derivative rules to find the correct answer.

Example 2.5. Let

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x^2}) & x \neq 0 \\ 0 & x = 0. \end{cases}$$

The graph of this function is given in the following figure:



To compute $f'(x)$, we apply the product rule, chain rule, power rule, and the result of the previous example. This leads to

$$f'(x) = 2x \sin(\tfrac{1}{x^2}) - \tfrac{1}{x} \cos(\tfrac{1}{x^2}) \quad \text{for any} \quad x \neq 0.$$

If we try to evaluate this at $x = 0$, we run into a problem, since the first term tends to zero x approaches zero but the second term becomes unbounded. Nonetheless, if we look at the graph, the function seems to be flat at $x = 0$, suggesting that perhaps we have $f'(0) = 0$. In fact,

$$\left| \frac{f(h) - f(0)}{h} \right| = |h \sin(\tfrac{1}{h^2})| \leq |h|,$$

which implies that

$$f'(0) = \lim_{h \to 0} \frac{f(h) - f(0)}{h} = 0.$$

We turn now to one of the most important applications of derivatives, namely polynomial approximations of functions. This is closely related to the idea of power series (or Taylor series) representations, which was mentioned in Section 1.4. To get the discussion started, we first need to introduce the notion of higher order derivatives. The idea is straightforward.

Definition 2.5 (Higher order derivatives)**.** Suppose f is a function defined on an open interval I such that $f'(x)$ exists for all $x \in I$. Then f' defines a function on I. If f' is differentiable at a point a, then we call $f''(a)$ the *second derivative* of f at $x = a$. Higher order derivatives are defined the same way. We write f', f'', f''', but eventually switch to notation such as $f^{(4)}$ or $f^{(k)}$ to denote the fourth or k^{th} derivative, respectively.

The idea of Taylor polynomial approximation is to use the polynomials

$$P_n(x) = f(a) + f'(a)(x-a) + \tfrac{1}{2!}f''(a)(x-a)^2 + \cdots + \tfrac{1}{n!}f^{(n)}(a)(x-a)^n$$

as approximations to f, at least in some interval around the point $x = a$. We may also make the notation more compact and write

$$P_n(x) = \sum_{k=0}^{n} \tfrac{1}{k!}f^{(k)}(a)(x-a)^k.$$

The precise coefficients in these polynomials are chosen precisely so that the polynomials P_n satisfy

$$P_n^{(k)}(a) = f^{(k)}(a) \quad \text{for} \quad k = 0, \dots, n.$$

To understand approximation properties of Taylor polynomials (and eventually power series), we will endeavor to prove the following *Lagrange remainder theorem.*

Theorem 2.4 (Lagrange Remainder Theorem)**.** *Suppose f is a function that is n times differentiable on an open interval I. For any $a, x \in I$, there exists c between a and x such that*

$$f(x) = \sum_{k=0}^{n-1} \tfrac{1}{k!}f^{(k)}(a)(x-a)^k + \tfrac{1}{n!}f^{(n)}(c)(x-a)^n.$$

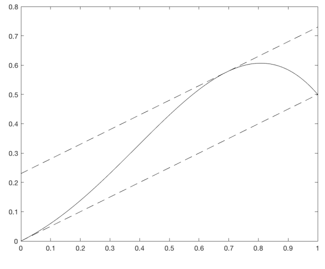
This theorem shows that the degree to which the polynomial P_{n-1} approximates f depends on the size of the n^{th} order of derivatives of f. Note that in the special case $n = 1$, we obtain the statement that there exists c between x and a so that

$$f(x) = f(a) + f'(c)(x-a), \quad \text{or} \quad \tfrac{f(x)-f(a)}{x-a} = f'(c).$$

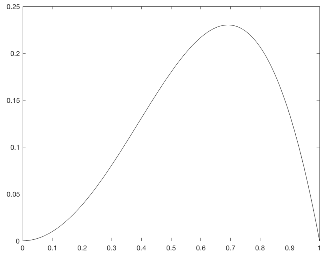
This is the well-known Mean Value Theorem, asserting that there exists c so that the slope of the tangent line at $x = c$ agrees with the slope of the line joining $(a, f(a))$ and $(x, f(x))$ (see the graph on the left in the figure below). Perhaps you remember from your calculus course that the Mean Value Theorem can be deduced from a result known as Rolle's Lemma. This will be our approach, as well. We state the result we will use as follows.

Lemma 2.5 (Rolle's lemma)**.** *Suppose f is differentiable on an open interval I. Suppose $a, x \in I$ are such that $f(a) = f(x)$. Then there exists c between a and x such that $f'(c) = 0$.*

We depict the Mean Value Theorem and Rolle's Lemma in the following figure:



The Mean Value Theorem



Rolle's Lemma

Rolle's Lemma is one of those results that just seems intuitively obvious once you draw the picture. Alas, such results can be the most dangerous of all! Indeed, to prove this result, we will ultimately be forced to introduce several completely new and fundamental concepts beyond differentiability. In this section, our approach will be to *use* Lemma 2.5 to prove Theorem 2.4 (which, as mentioned above, includes the Mean Value Theorem as a special case). We will then see exactly what a proof of Rolle's Lemma might require of us, and we will develop the necessary concepts in the next section.

Proof of Theorem 2.4, assuming Lemma 2.5. Let us first prove the case $n = 1$, which (as stated above) corresponds to the Mean Value Theorem. The idea is to reduce the Mean Value Theorem to Rolle's theorem basically by subtracting the appropriate line from the function (essentially tilting our heads until the graph on the left above looks like the graph on the right).

To this end, we define

$$g(y) = [f(x) - f(y)] - \tfrac{x-y}{x-a}[f(x) - f(a)]$$

for y between a and x. Then we have

$$g(a) = 0 \quad \text{and} \quad g(x) = 0,$$

so that by Rolle's Lemma we may find c between a and x satisfying $g'(c) = 0$. Observing that

$$g'(c) = -f'(c) + \tfrac{f(x)-f(a)}{x-a},$$

we obtain the result in the case $n = 1$ (that is, the Mean Value Theorem).

The general case can also be reduced to Rolle's Lemma. Let us define the remainder $R_n(x, y)$ by

$$R_n(x, y) = f(x) - \sum_{k=0}^{n-1} \tfrac{f^{(k)}(y)}{k!}(x-y)^k.$$

Our goal is then to prove that there exists c between a and x so that

$$R_n(x, a) = \tfrac{f^{(n)}(c)}{n!}(x-a)^n.$$

First, by the product rule, we can compute

$$\begin{aligned} \tfrac{d}{dy} R_n(x, y) &= -\left[\sum_{k=0}^{n-1} \tfrac{f^{(k+1)}(y)}{k!}(x-y)^k - \sum_{k=1}^{n-1} \tfrac{f^{(k)}(y)}{(k-1)!}(x-y)^{k-1}\right] \\ &= -\tfrac{f^{(n)}(y)}{(n-1)!}(x-y)^{n-1} \end{aligned} \tag{2.10}$$

(see Exercise 2.4).

We now define the function

$$g(y) = R_n(x, y) - \tfrac{(x-y)^n}{(x-a)^n} R_n(x, a).$$

(In fact, when $n = 1$, this is exactly the same construction as above.) We then have

$$g(a) = R_n(x, a) - R_n(x, a) = 0, \quad \text{while} \quad g(x) = R_n(x, x) = 0,$$

so that by Rolle's Lemma we may find c between a and x so that $g'(c) = 0$. Using (2.10), this becomes

$$0 = g'(c) = -\tfrac{f^{(n)}(c)}{(n-1)!}(x-c)^{n-1} + \tfrac{n(x-c)^{n-1}}{(x-a)^n} R_n(x, a),$$

which rearranges to yield

$$R_n(x,a) = \tfrac{f^{(n)}(c)}{n!}(x-a)^n,$$

as desired. □

The proof above relies on Rolle's Lemma, to which we now turn.

First attempt at proving Rolle's Lemma. Let's recall the setup of Rolle's Lemma. We let f be a differentiable function on an interval I, and we let $a < x$ be two points in I satisfying $f(a) = f(x)$. Why must there be a point c so that $f'(c) = 0$?

We could try to argue as follows: Imagine connecting $f(a)$ to $f(x)$ with some 'continuous' curve. The most trivial case is a horizontal line, but in this case the derivative is *always* zero, so we're done. Otherwise, to get back to the same value, the curve will have to go up and come back down again (or vice versa). However, at any moment when the function changes from being increasing to decreasing (or vice versa), it will have a horizontal tangent line, that is, its derivative will equal zero. This argument seems to rest on several 'obvious' facts:

Obvious Fact #1. A differentiable function on an interval must be 'continuous'.

Obvious Fact #2. A non-constant 'continuous' function on will obtain some maximum or minimum on an interval.

Obvious Fact #3. The derivative at a maximum or minimum will equal zero.

Of course, the use of the word 'obvious' above is meant to be a bit facetious. What is really meant is that our intuition strongly suggests that these things are true. In fact, we will need to develop some new ideas to be able to write down rigorous proofs of these facts. The key new concept that we need is that of *continuity*, which is the topic of our next section. □

2.2. **Continuity.** The main goal of this section is to introduce the concept of continuity. In particular, we will use this concept to complete the proof of Rolle's Lemma from the previous section. We will also explore several other important consequences of continuity.

We should take a moment to ponder what the right definition of continuity even ought to be. Historically, the notion of continuity arose much later than the notion of differentiability, and indeed it turns out to be a much more general concept that extends to far more abstract settings.

A reasonable candidate for what 'continuity' might mean is the following 'intermediate value property':

Definition 2.6 (Intermediate value property)**.** Let f be a function defined on an interval $[a,b]$. We say that f has the *intermediate value property* if for any $x_1, x_2 \in [a,b]$ and any m satisfying

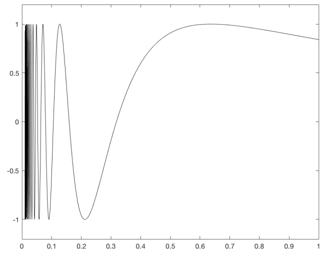
$$f(x_1) < m < f(x_2) \quad \text{or} \quad f(x_2) < m < f(x_1),$$

there exists c between x_1 and x_2 such that $f(c) = m$.

This definition seems reminiscent of the old expression that 'a continuous function is one you can draw without picking up your pencil'. As we will see, this property alone does not constitute an acceptable definition of continuity. For one thing, we can find functions satisfying the intermediate value property that we would never consider to be continuous. Consider, for example:

$$f(x) = \begin{cases} \sin(\frac{1}{x}) & x \in (0,1] \\ 0 & x = 0, \end{cases} \tag{2.11}$$

depicted (as best we can) in the following figure:



In particular, this function satisfies the intermediate value property (indeed, it takes on on every value between -1 and 1 infinitely many times as $x \to 0$), but there is no way we could call this function continuous at $x = 0$.

Another defect of the intermediate value property is that it is not preserved under simple operations like addition! Indeed, we can basically re-use the same example above, defining a second function

$$g(x) = \begin{cases} -\sin(\frac{1}{x}) & x \in (0,1] \\ 1 & x = 0. \end{cases}$$

Then g satisfies the intermediate value property for the same reasons f does, but

$$(f+g)(x) = \begin{cases} 0 & x \in (0,1] \\ 1 & x = 0, \end{cases}$$

which certainly fails to have the intermediate value property.

We will instead define continuity as follows.

Definition 2.7 (Continuity)**.** We say that a function f is *continuous* at a point $x = a$ if

(i) f is defined in an interval containing a, and
(ii) $\lim_{x\to a} f(x)$ exists and equals $f(a)$.

In particular continuity at a point $x = a$ means that the function has a limiting value as x approaches a, and that that limiting value agrees with the value of the function at $x = a$. Note that in item (i), it is permitted that a be the endpoint of the interval on which f is defined; in such cases, we only need to consider a 'one-sided' limit in part (ii).

In the case of (2.11), we see that continuity fails because the function has no limit as $x \to 0$.

Using the standard limit laws (see Exercise 2.2), we can immediately see that continuity is a property that is preserved under scalar multiplication, finite sums, products, and quotients (provided the function in the denominator is nonzero). Whether or not continuity is preserved under infinite sums will be addressed later.

As with differentiability, continuity is preserved under the composition of functions.

Proposition 2.6. *Suppose the composition $f\circ g$ is defined in an interval containing $x=a$. Suppose that g is continuous at a and f is continuous at $g(a)$. Then $f\circ g$ is continuous at a.*

Proof. Let $\varepsilon>0$. By continuity of f at $g(a)$, there exists $\delta>0$ so that

$$0<|y-g(a)|<\delta \implies |f(y)-f(g(a))|<\varepsilon.$$

By continuity of g at a, there exists $\eta>0$ so that

$$0<|x-a|<\eta \implies |g(a)-g(x)|<\delta.$$

Thus, for $0<|x-a|<\eta$, we have

$$|f(g(x))-f(g(a))|<\varepsilon.$$

As $\varepsilon>0$ was arbitrary, we deduce $\lim_{x\to a} f(g(x))=f(g(a))$. □

Using this definition of continuity, we would now like to show that (i) continuous functions satisfy the intermediate value property and (ii) differentiable functions are continuous.

Our first main goal is therefore the following:

Theorem 2.7 (Intermediate Value Theorem)**.** *Suppose f is a continuous function on an interval $[a,b]$. Then f has the intermediate value property on $[a,b]$.*

It is convenient at this point to formally introduce the notion of sequences of real numbers and their limits. A sequence of numbers is really a function mapping the natural numbers $\{1,2,3,\dots\}$ into the real numbers. We typically denote sequences by $\{x_k\}$, where $k=1,2,\dots$ (as opposed to the usual functional notation $x(k)$, say). Informally, we may say that we have an 'infinite, ordered list of numbers'.

We will jump straight to the formal definition of limits (the ε-N definition). It once again corresponds to the 'Archimedean' notion of limits.

Definition 2.8 (Limit, sequence version)**.** We say

$$\lim_{n\to\infty} a_n=\ell$$

if for any $\varepsilon>0$, there exists N such that

$$n\geq N \implies |a_n-\ell|<\varepsilon.$$

We may also write $a_n\to\ell$ as $n\to\infty$.

In particular, we can now observe that the statement that an infinite series has value ℓ (in the sense of Definition 1.1) is equivalent to the statement that the limit of the sequence of partial sums exists and equals ℓ.

In the exercises, you will prove an analogue of the '$C\varepsilon$ Theorem' (Theorem 2.1) for sequences of real numbers, as well as the standard algebraic 'limit laws'. You will also prove the following handy lemma:

Lemma 2.8 (Sequential version of continuity)**.** *The following are equivalent:*

- *f is continuous at a point $x=a$.*
- *If $\{x_k\}$ is a sequence such that $\lim_{k\to\infty} x_k=a$, then $\lim_{k\to\infty} f(x_k)$ exists and equals $f(a)$.*

We are now in a position to start the proof of the Intermediate Value Theorem.

Proof of Theorem 2.7. Fix any $x_1, y_1 \in [a,b]$ with $x_1 < y_1$ and suppose that M is between $f(x_1)$ and $f(y_1)$.

Let m_1 be the midpoint of x_1 and y_1. If $f(m_1) = M$, then we are done. Otherwise, $f(m_1)$ will be on the same side of M as either $f(x_1)$ or $f(y_1)$.

- If $f(m_1)$ is on the same side as $f(y_1)$, let $x_2 = x_1$ and $y_2 = m_1$.
- If $f(m_1)$ is on the same side as $f(x_1)$, let $x_2 = m_1$ and and $y_2 = y_1$.

Then we have

$$x_1 \leq x_2 < y_2 \leq y_1, \quad y_2 - x_2 = \tfrac{1}{2}(y_1 - x_1),$$

and M is between $f(x_2)$ and $f(y_2)$.

Now we repeat this step: We let m_2 be the midpoint of x_2 and y_2. If $f(m_2) = M$, we are done. Otherwise, we define x_3 and y_3 according to the same criterion above, and we obtain

$$x_1 \leq x_2 \leq x_3 < y_3 \leq y_2 \leq y_1, \quad y_3 - x_3 = \tfrac{1}{2}(y_2 - x_2) = \tfrac{1}{4}(y_1 - x_1),$$

and M is between $f(x_3)$ and $f(y_3)$. Proceeding in this way, we construct sequences $\{x_k\}$ and $\{y_k\}$ so that

$$x_1 \leq \cdots \leq x_k < y_k \leq \cdots \leq y_1, \quad y_{k+1} - x_{k+1} = \tfrac{1}{2^k}(y_1 - x_1),$$

and M is always between $f(x_k)$ and $f(y_k)$.

Now, if one of the midpoints ever satisfies $f(m_k) = c$, then we stop this process and the proof is done. Otherwise, this process constructs infinite sequences x_k, y_k such that the intervals (x_k, y_k) are getting 'squeezed down' and appear to be collapsing down to some point (similar to the situation discussed in Section 1.6). In this case, we hold our breath and **posit** that

$$\text{there exists} \quad c \in \mathbb{R} \quad \text{so that} \quad c \in (x_k, y_k) \quad \text{for every} \quad k,$$

and that

$$\lim_{k\to\infty} x_k = c \quad \text{and} \quad \lim_{k\to\infty} y_k = c. \tag{2.12}$$

Taking this for granted for the moment, we see that by the continuity of f (in the form of Lemma 2.8) then implies

$$\lim_{k\to\infty} f(x_k) = f(c) \quad \text{and} \quad \lim_{k\to\infty} f(y_k) = f(c). \tag{2.13}$$

We now claim that we must have $f(c) = M$, which will complete the proof (modulo the claim in (2.12), which we return to below). To see this, suppose first that $f(c) < M$. Then, by (2.13), we would have that $f(x_k) < M$ and $f(y_k) < M$ for all k sufficiently large. However, this contradicts the fact that M is always between $f(x_k)$ and $f(y_k)$. The assumption that $f(c) > M$ leads to a similar contradiction, and hence we conclude that $f(c) = M$, as desired. □

We now have a proof of the intermediate value property for continuous functions, but it rests on the condition (2.12). As we will see, this condition ultimately follows from the *completeness* property of the real numbers that was mentioned briefly in Section 1.6. Indeed, the following example shows that the intermediate value theorem could *fail* without this property!

Example 2.6. Let f be defined on the rational numbers by $f(x) = x^2$. Then f is continuous (this is clear from the sequential version of compactness in Lemma 2.8), $f(0) = 0$, $f(2) = 4$, but there is no x such that $f(x) = 2$.

Rigorous justification for the claim made in (2.12) was not given until the late 1800s in work of mathematicians such as Dedekind and Cantor. We will consider this foundational work in a later section. For the time being, let us state the precise fact about the real numbers that we need, and show how it implies (2.12).

To formally state the completeness property of $\mathbb{R}$, we need a new definition. It refers to sequences whose terms get closer and closer to one another.

Definition 2.9 (Cauchy sequence)**.** A sequence x_n of real numbers is said to be *Cauchy* if for all $\varepsilon > 0$, there exists N such that

$$n, m \geq N \implies |x_n - x_m| < \varepsilon.$$

For example, every convergent sequence is Cauchy (see Exercise 2.10). So is the sequence constructed in Example 1.4.

The completeness property of $\mathbb{R}$ is now straightforward to state. We have the following:

Proposition 2.9 (Completeness)**.** *The real numbers, $\mathbb{R}$, are complete.*

Example 1.4, on the other hand, shows that the rational numbers $\mathbb{Q}$ are *not complete.* We will discuss this proposition in a later section. For now, let us use it to prove (2.12).

Proof of (2.12)*, assuming Proposition 2.9.* We let $\varepsilon > 0$. By construction, we may find N so that $|y_N - x_N| < \varepsilon$. Again by construction, we have that $x_n \in [x_N, y_N]$ for all $n \geq N$, and so $|x_n - x_m| < \varepsilon$ for all $n, m \geq N$. Similarly, $|y_n - y_m| < \varepsilon$ for all $n, m \geq N$. Thus $\{x_n\}$ and $\{y_n\}$ are both Cauchy and hence converge to some limits c_x and c_y. Using

$$\lim_{n\to\infty} |x_n - y_n| = 0,$$

we can quickly deduce that $c_x = c_y$, so that we may drop the subscript and denote the common limit by c. We now observe that $c \leq y_k$ for every k, for if we had $c > y_k$ then we would obtain $x_n > y_k$ for large enough n (contradicting the construction of these sequences). Similarly, $c \geq x_k$ for every k. This completes the proof of all of the claims made in the proof above. □

Next, let us demonstrate that differentiable functions are continuous. (This was one of the 'obvious facts' that we needed with when trying to construct a proof of Rolle's Lemma.)

Proposition 2.10 (Differentiability implies continuity)**.** *Suppose f is differentiable on an open interval I. Then f is continuous at each point $a \in I$.*

Proof. Let $a \in I$. By assumption,

$$\lim_{x\to a} \frac{f(x) - f(a)}{x - a} = f'(a).$$

Thus, by limit laws,

$$\lim_{x\to a} [f(x) - f(a)] = \lim_{x\to a} [x - a] \cdot \lim_{x\to a} \frac{f(x) - f(a)}{x - a} = 0 \cdot f'(a) = 0,$$

so that $\lim_{x\to a} f(x) = f(a)$. □

The converse to Proposition 2.10 is false. That is, continuity does *not* imply differentiability. A simple example is $f(x) = |x|$, which is continuous but not differentiable at $x = 0$ (see Exercise 2.1). In fact, there are much weirder examples than this (e.g. functions that are continuous but nowhere differentiable), but one needs to have an understanding of infinite series of functions before it is really possible to discuss them.

Similarly, the derivative of a differentiable function need not be continuous. We saw such an example in Example 2.5, where (with $f(x) = x^2 \sin(x^{-2})$) we had $f'(0) = 0$ but $\lim_{x\to 0} f'(x)$ did not exist. Despite this possibility, we will see later that any function that arises as a derivative is still guaranteed to have the intermediate value property (a result known as *Darboux's Theorem*).

In fact, there are some bizarre examples that already challenge our intuitive understanding of 'continuity'.

Example 2.7 (Dirichlet function)**.** The function

$$f(x) = \begin{cases} x & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

is continuous at $x = 0$ but discontinuous for all $x \neq 0$. The proof relies on the fact that the rational numbers are *dense* in the real numbers, which means for any real number x and any $\varepsilon > 0$, there exists a rational number y such that $|x - y| < \varepsilon$.

Example 2.8. In the following, we write nonzero rational numbers uniquely as $x = p/q$, with $q > 0$ and p and q relatively prime. If we then define

$$g(x) = \begin{cases} 1 & x = 0, \\ 1/q & \text{if } x = p/q \text{ is a rational number,} \\ 0 & \text{if } x \text{ is irrational,} \end{cases}$$

then g is continuous precisely at the irrational numbers.

At the same time, our original thoughts about 'continuity' (namely, that it should be closely linked to the intermediate value property), are not so far off. Indeed, combined with one additional property, the intermediate value property *does* guarantee continuity. The additional property is that of *monotonicity.*

Definition 2.10 (Monotone function)**.** A function f is *monotone increasing* if

$$x_1 < x_2 \implies f(x_1) \leq f(x_2).$$

It is *monotone decreasing* if

$$x_1 < x_2 \implies f(x_1) \geq f(x_2).$$

Theorem 2.11. *Suppose f is monotonic and satisfies the intermediate value property on an interval $[a, b]$. Then f is continuous at every $c \in (a, b)$.*

Proof. Without loss of generality, let's suppose f is increasing. We let $c \in (a, b)$ and $\varepsilon > 0$. We let

$$\varepsilon' = \min\{\varepsilon, f(c) - f(a), f(b) - f(c)\} > 0.$$

By the intermediate value property and the fact that f is increasing, we may find $c_1 < c < c_2$ so that

$$f(c_1) = f(c) - \tfrac{1}{2}\varepsilon' \quad \text{and} \quad f(c_2) = f(c) + \tfrac{1}{2}\varepsilon'.$$

Using the monotonicity of f again, it follows that

$$x \in (c_1, c_2) \implies |f(x) - f(c)| < \varepsilon' \leq \varepsilon,$$

which implies continuity. □

Our next goal is to establish two crucial properties of continuous functions on closed, bounded intervals. In particular, we will show that such functions are *bounded* and that they *achieve their extreme values.*

Theorem 2.12 (Continuity implies boundedness)**.** *If f is continuous on $[a, b]$, then f is* bounded: *that is, there exists M_1, M_2 so that*

$$M_1 \leq f(x) \leq M_2 \quad \textit{for all} \quad x \in [a, b].$$

The hypotheses of this theorem are all necessary: First, we need continuity here, as opposed to just the intermediate value property, say. To see this, consider the function $f(x) = \frac{1}{x}\sin(\frac{1}{x})$ for $x \in (0, 1]$, with $f(0) = 0$. This satisfies the intermediate value property but is unbounded. We also need the interval to be closed (otherwise, consider the continuous function $f(x) = \frac{1}{x}$ on $(0, 1]$) and bounded (otherwise, consider $f(x) = x$ on $[0, \infty)$).

Proof of Theorem 2.12. Suppose f is not bounded. Let $x_1 = a$, $y_1 = b$, and $m_1 = \frac{1}{2}(x_1 + y_1)$. Then f must be unbounded on either $[x_1, m_1]$ or $[m_1, y_1]$ (or both). Choose one of these intervals on which f is unbounded and denote this interval by $[x_2, y_2]$. Then we have

$$x_1 \leq x_2 < y_2 \leq y_1 \quad \text{and} \quad y_2 - x_2 = \tfrac{1}{2}(y_1 - x_1).$$

Now repeat the process, defining the midpoint m_2 of $[x_2, y_2]$ and choosing one of the resulting halves $[x_3, y_3]$ on which f is unbounded. Proceeding in this way we construct sequences $\{x_k\}$ and $\{y_k\}$ so that

$$x_1 \leq x_2 \leq \cdots \leq x_k < y_k \leq \cdots \leq y_2 \leq y_1, \quad y_{k+1} - x_{k+1} = \tfrac{1}{2^k}(y_1 - x_1),$$

and f is unbounded on each interval $[x_k, y_k]$. Arguing as we did on page 26, we deduce that there exists $c \in (x_k, y_k)$ for all k such that $x_k \to c$ and $y_k \to c$ as $k \to \infty$. By continuity, we have $f(x_k) \to f(c)$ and $f(y_k) \to f(c)$.

Now, by continuity, there exists $\delta > 0$ so that

$$|z - c| < \delta \implies |f(z) - f(c)| < 1. \tag{2.14}$$

However, for large enough k, we have $|y_k - x_k| < \delta$, while f is unbounded on $[x_k, y_k]$. In particular, we may find $z \in [x_k, y_k]$ (so that $|z - c| < \delta$) such that

$$f(z) > f(c) + 1 \quad \text{or} \quad f(z) < f(c) - 1,$$

contradicting (2.14). □

Next, we would like to show that continuous functions achieve maximum and minimum values on closed, bounded intervals. Here we once again stumble into some subtleties that must be resolved using the completeness of the real numbers. There are actually two questions we must answer. First, given that f is bounded on the interval $[a, b]$, what do we even mean by the 'extreme values' or the 'best possible bounds' for f? Second, how can we show that f attains these values?

For the first question, we need to introduce the notion of *least upper bounds* and *greatest lower bounds*, also called *suprema* and *infima*, respectively.

Definition 2.11 (Least upper bound, greatest lower bound)**.** Suppose S is a bounded set of real numbers. We say that m is a *least upper bound* of S if

- m is an upper bound for S, that is, $x \in S \implies x \leq m$, and
- if y is any upper bound for S, then $m \leq y$.

We say that b is a *greatest lower bound* of S if

- b is a lower bound for S, that is, $x \in S \implies x \geq b$, and
- if z is any lower bound for S, then $z \leq b$.

This definition suggests two questions: First, do least upper bounds always exist? Second, if they do, are they unique?

The uniqueness question is straightforward, if m_1 and m_2 are both 'least upper bounds' for S, then the second condition means we must have $m_1 \leq m_2$ and $m_2 \leq m_1$, so that $m_1 = m_2$. The situation is similar for greatest lower bounds. We may speak of *the* least upper bound or greatest lower bound, if they exist. We also call these the *supremum* and *infimum*, respectively, denoted by sup and inf.

If a set has a maximal element, this element is necessarily the supremum. However, in general the question of existence is more subtle. In fact, existence can fail for sets of rational numbers:

Example 2.9. Let S be the set of positive rational numbers x such that $x^2 < 2$. This set is bounded above (e.g. $\frac{3}{2}$ is an upper bound). However, it has no least upper bound within the set of rational numbers. Indeed, suppose $m > 0$ were a (rational) least upper bound for S. Then either $m^2 > 2$ or $m^2 < 2$ (since there is no rational number whose square equals 2). If $m^2 < 2$, then we may use the sequence constructed in Example 1.4 to find a rational number $x > 0$ such that $x^2 > m^2$ (and so which yields $x > m$). In particular, m cannot not be an upper bound for S. On the other hand, if $m^2 > 2$, for n sufficiently large we have $(m - \frac{1}{n})^2 > 2$, which implies that $m - \frac{1}{n}$ is an upper bound for the set S for large enough n. In particular, m cannot be the *least* upper bound. We conclude that S *has* no least upper bound.

If we consider the set of *real* numbers x such that $x^2 < 2$, then the supremum is simply given by $\sqrt{2}$. The fact that something goes wrong for the rationals but not the reals suggests that the existence of suprema is connected to the completeness property of the real numbers. In fact, we have the following theorem, the proof of which relies on the completeness of the real numbers (Proposition 2.9):

Theorem 2.13 (Least upper bound property)**.** *If S is a set of real numbers with an upper bound, then S has a least upper bound. (A similar statement holds for greatest lower bounds.)*

Proof. We'll use the nested interval trick that we've seen a few times by now. We let x_1 be any number that is *not* an upper bound for S, and let y_1 be any upper bound for S. Then take the midpoint $m_1 = \frac{1}{2}(x_1 + y_1)$, and:

- If m_1 is an upper bound for S, we let $x_2 = x_1$ and $y_2 = m_1$.
- If m_1 is not an upper bound for S, we let $x_2 = m_1$ and $y_2 = y_1$.

This yields

$$x_1 \leq x_2 < y_2 \leq y_1 \quad \text{and} \quad y_2 - y_1 = \tfrac{1}{2}(x_2 - x_1).$$

Now repeat the process, using the midpoint $m_2 = \frac{1}{2}(x_2 + y_2)$. In this way, we construct sequences x_k and y_k satisfying

$$x_1 \leq x_2 \leq \cdots x_k < y_k \leq \cdots \leq y_2 \leq y_1, \quad y_{k+1} - x_{k+1} = \tfrac{1}{2^k}(y_1 - x_1),$$

with x_k *not* an upper bound for S and y_k an upper bound for S.

Now, as we have argued before, there must exist $c \in \mathbb{R}$ satisfying $c \in (x_k, y_k)$ for all k and $x_k, y_k \to c$ as $k \to \infty$. We claim that c is the least upper bound of S.

First we show that c is an upper bound. Indeed, if $x > c$, then there exists k such that $y_k < x$. As y_k is an upper bound for S, this means $x \notin S$. Stated in the contrapositive, we have just shown $x \in S \implies x \leq c$.

Next, if $y < c$, then there exists k such that $x_k > y$. As x_k is not an upper bound for S, y cannot be an upper bound either. Stated in the contrapositive, we have just shown that if y is an upper bound for S, then $y \geq c$. □

We are now in a position to show that continuous functions achieve their extreme values on on closed, bounded intervals. (This is another one of the 'obvious facts' that we needed for the proof of Rolle's Lemma.)

Theorem 2.14 (Extreme value theorem)**.** *Let f be a continuous function on an interval $[a, b]$. Then there exists $x \in [a, b]$ such that*

$$f(x) = M := \sup\{f(y) : y \in [a, b]\}.$$

(Similarly, f attains its minimal value.)

Proof. We'll use our nested interval trick again. Let $x_1 = a$, $y_1 = b$, and $m_1 = \frac{1}{2}(x_1 + y_1)$. Then M must be the supremum of f over either $[x_1, m_1]$ or $[m_1, y_1]$ (or both). If it is the supremum over $[x_1, m_1]$, we let $x_2 = x_1$ and $y_2 = m_1$; if not, we let $x_2 = m_1$ and $y_2 = y_1$.

We now repeat this process, always choosing the subinterval on which M is the supremum. This leads to sequences obeying

$$x_1 \leq x_2 \leq \cdots x_k < y_k \leq \cdots \leq y_2 \leq y_1, \quad y_{k+1} - x_{k+1} = \tfrac{1}{2^k}(y_1 - x_1),$$

and such that $\sup_{[x_k, y_k]} f = M$ for all k. We then let c be the common point of all of these intervals, that is, the common limit of the sequences x_k and y_k.

We now show $f(c) = M$. By construction, the only other option is $f(c) < M$. Now, if $f(c) < M$, then we can find ε so that $0 < \varepsilon < M - f(c)$. By continuity of f, we can therefore choose $\delta > 0$ so that

$$|x - c| < \delta \implies |f(x) - f(c)| < \varepsilon.$$

For k sufficiently large, we have that $|z - c| < \delta$ for all $z \in [x_k, y_k]$. In this case,

$$f(z) < f(c) + \varepsilon < M \quad \text{for all} \quad z \in [x_k, y_k].$$

That is, $f(c) + \varepsilon$ is an upper bound for f on $[x_k, y_k]$ that is strictly less than M, contradicting that $\sup_{[x_k, y_k]} f = M$. We therefore conclude that $f(c) = M$, as desired. □

We are finally almost in a position to complete the proof of Rolle's Lemma from the previous section. The final 'obvious fact' that we needed is the following:

Theorem 2.15 (Fermat's theorem on max/min)**.** *Suppose f is a differentiable function on (a, b) that attains a maximum or minimum at $c \in (a, b)$. Then $f'(c) = 0$.*

Proof. We will show that if $f'(c) \neq 0$, then there exists $x_1, x_2 \in (a,b)$ so that

$$f(x_1) < f(c) < f(x_2).$$

In particular, $f'(c) \neq 0$ implies that $f(c)$ is neither a maximum or minimum.

Without loss of generality, suppose $f'(c) > 0$. It follows (by definition of the derivative) that for all x sufficiently close to c, we have

$$\frac{f(x) - f(c)}{x - c} > 0.$$

In particular, we can find $x_1 < c$ so that $f(x_1) < c$ and $x_2 > c$ so that $f(x_2) > c$. □

Finally, we can prove Rolle's Lemma. Let us state it again carefully.

Lemma 2.16 (Rolle's Lemma)**.** *Suppose f is a continuous function on $[a,b]$ that is differentiable on (a,b). If $f(a) = f(b)$, then there exists $c \in (a,b)$ so that $f'(c) = 0$.*

Proof. If $f(x) = f(a)$ for all $x \in [a,b]$, then f is constant and $f'(x) = 0$ for all x. Otherwise, f attains a maximum or minimum at some point $c \in (a,b)$, at which its derivative necessarily vanishes. □

With Rolle's Lemma in place, the arguments of the previous section also yield the Mean Value Theorem and the Lagrange Remainder Theorem.

Using Rolle's Lemma, we can also establish the ever-useful 'L'Hospital's Rules' (which are actually due to Johann Bernoulli!).

Theorem 2.17 (L'Hospital's $\frac{0}{0}$ Rule)**.** *Suppose that f and g are differentiable on an open interval I containing a. Suppose further that*

$$\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0.$$

Finally, suppose that g, g' are nonzero on $I \backslash \{a\}$. Then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)},$$

provided this latter limit exists.

Proof. First note that differentiability of f and g implies continuity, and hence we must have $f(a) = g(a) = 0$.

Now let $\varepsilon > 0$ and denote

$$\ell = \lim_{x \to a} \tfrac{f'(x)}{g'(x)}.$$

We may then choose $\delta > 0$ so that

$$0 < |x - a| < \delta \implies |\tfrac{f'(x)}{g'(x)} - \ell| < \varepsilon.$$

Now fix x obeying $0 < |x - a| < \delta$ and define the function

$$F(y) = [f(y) - f(a)][g(x) - g(a)] - [g(y) - g(a)][f(x) - f(a)].$$

Then $F(a) = F(x) = 0$, and hence by Rolle's Lemma there exists a c between a and x so that $F'(c) = 0$. Using the definition of F, that means

$$f'(c)[g(x) - g(a)] = g'(c)[f(x) - f(a)], \quad \text{or} \quad \tfrac{f'(c)}{g'(c)} = \tfrac{f(x) - f(a)}{g(x) - g(a)} = \tfrac{f(x)}{g(x)},$$

where we have used that $f(a) = g(a) = 0$ and the fact that g', g are nonzero. Collecting the above and observing that $0 < |x - a| < \delta$ guarantees $0 < |c - a| < \delta$, we have just established that for $0 < |x - a| < \delta$,

$$|\tfrac{f(x)}{g(x)} - \ell| < \varepsilon,$$

which completes the proof. □

There is also an "$\frac{\infty}{\infty}$" version of L'Hospital's Rule. You may try to work out its statement and proof, or perhaps you could pay somebody else to do it for you!

In the rest of this section, we will establish a few more results and also discuss an alternate approach to proving Rolle's Lemma. This not only serves to demonstrate the interconnectedness of many of these results, but also to warn how easily one can stumble into circular reasoning.

We first establish Darboux's theorem. There is an approach to proving this result that relies on the Mean Value Theorem, but since we would like to use this theorem to provide an alternate proof of Rolle's Lemma, we will present a proof that relies only on the Extreme Value Theorem and Fermat's Theorem.

Theorem 2.18 (Darboux's Theorem)**.** *If f is differentiable on $[a, b]$, then f' has the intermediate value property on $[a, b]$.*

Proof. Without loss of generality, we suppose $f'(b) < m < f'(a)$ and seek a point c such that $f'(c) = m$. We define the auxiliary function $g(x) = f(x) - mx$, which is continuous on $[a, b]$ and hence attains a maximum at some $c \in [a, b]$. Since $g'(a) = f'(a) - m > 0$, the maximum cannot occur at a, and similarly since $g'(b) < 0$ the maximum cannot occur at b. Thus $c \in (a, b)$ and $g'(c) = 0$, which yields $f'(c) = m$. □

The next result we'll prove is another one of these 'obvious' facts that turns out to be surprisingly subtle.

Lemma 2.19. *If $f'(x) > 0$ for all $x \in [a, b]$, then f is increasing over $[a, b]$. (Similarly, if $f' < 0$ then f is decreasing.)*

Note that we must interpret $f'(a)$ and $f'(b)$ as one-sided limits in this context.

Proof. First, observe the Mean Value Theorem provides a very slick proof: If f is not increasing over $[a, b]$, then we may find $a \leq x_1 < x_2 \leq b$ so that $f(x_1) \geq f(x_2)$. Then by the Mean Value Theorem, there exists $c \in (x_1, x_2)$ so that

$$f'(c) = \tfrac{f(x_2) - f(x_1)}{x_2 - x_1} \leq 0,$$

contradicting the assumption that $f' > 0$.

However, as we would like to use this result to give an alternate proof of Rolle's Lemma, let us also give a proof of this result that does *not* use the Mean Value Theorem. We argue as follows:

Fix $x_2 \in (a, b]$ and let

$$S = \{x : x \in [a, x_2) \quad \text{and} \quad f(x) \geq f(x_2)\}.$$

We need to show that S is empty. Suppose instead that S is non-empty. As S is bounded, it has a least upper bound m, which satisfies $m \leq x_2$.

We first claim that $f(m) \geq f(x_2)$. To see this, observe that for any n, the number $m - \frac{1}{n}$ is *not* an upper bound for S, and so there exists $y_n \in [m - \frac{1}{n}, m]$

such that $y_n \in S$, that is, $f(y_n) \geq f(x_2)$. As $y_n \to m$, we deduce $f(y_n) \to f(m)$ by continuity of f, and hence it follows that $f(m) \geq f(x_2)$, as claimed.

We next claim that $m < x_2$. Suppose instead that $m = x_2$. Then, in particular, $m \notin S$, and (as above) we may find a sequence $y_n \in [x_2 - \frac{1}{n}, x_2)$ such that $f(y_n) \geq f(x_2)$ for all n. However, along this sequence we have

$$\frac{f(x_2) - f(y_n)}{x_2 - y_n} \to f'(x_2) > 0,$$

which implies $f(x_2) > f(y_n)$ for all n sufficiently large. This is a contradiction, and so we must have $m < x_2$.

Finally, we claim that there exists $y > m$ such that $y \in S$, contradicting that $m = \sup S$. This is equivalent to finding $m < y < x_2$ so that $f(y) \geq f(x_2)$. Suppose instead that $f(y) < f(x_2)$ for all $y \in (m, x_2)$. Then, we may find a sequence $y_n > m$ with $y_n \to m$ such that $f(y_n) < f(x_2)$. Recalling $f(x_2) \leq f(m)$, this implies

$$\frac{f(y_n) - f(m)}{y_n - m} < 0 \quad \text{for all} \quad n.$$

However, along this sequence we have

$$\frac{f(y_n) - f(m)}{y_n - m} \to f'(m) > 0,$$

and thus we obtain a contradiction for all sufficiently large n.

We conclude that S is empty, and hence for all $a \leq x_1 < x_2 \leq b$, we obtain $f(x_1) < f(x_2)$, as desired. □

Let us now give another proof of Rolle's Lemma, this time relying on the last two results proven. Again, this is why we insisted on giving proofs of the previous results that did *not* rely on the mean value theorem! Indeed, using the Mean Value Theorem on the way to proving Rolle's Lemma would be completely circular, given that we use Rolle's Lemma to prove the Mean Value Theorem!

Another proof of Rolle's Lemma. Suppose $f(a) = f(b)$, with $a < b$. As before, we assume f is not constant (otherwise the result follows immediately). Now pick $y \in (a, b)$. If $f'(y) = 0$, we are done, so suppose instead that $f'(y) > 0$ (say). Now, if $f' > 0$ for all other points in $[a, b]$, then by the previous lemma we have that f is increasing on $[a, b]$, which is incompatible with the fact that $f(a) = f(b)$. In particular, there must be some point $z \in (a, b)$ so that $f'(z) \leq 0$. If $f'(z) = 0$, then we are done. If $f'(z) < 0$, then by Darboux's Theorem there must be some point c between y and z such that $f'(c) = 0$, and again we are done. □

To close this section, let us make the simple observation that with Darboux's Theorem in hand, it becomes clear that identifying integrals with antiderivatives cannot give a satisfactory theory of integration. Indeed, any function with a jump discontinuity *cannot* be the derivative of a function, while it can still be straightfoward to interpret the integral of such a function as an area under a curve.

2.3. **Exercises.**

Exercise 2.1. (i) Define the function f by

$$f(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases}$$

Show that

$$\lim_{x\to 0} f(x) \quad \text{does not exist.}$$

(ii) Show that the function $f(x)=|x|$ is not differentiable at $x=0$.

Exercise 2.2. Prove the following limit laws:

- If $\lim_{x\to a} f(x)=\ell$ and $\lim_{x\to a} g(x)=m$, then
$$\lim_{x\to a}[f(x)+g(x)]=\ell+m.$$
- If $\lim_{x\to a} f(x)=\ell$ and c is a real number, then
$$\lim_{x\to a} c\cdot f(x)=c\ell.$$
- If $\lim_{x\to a} f(x)=\ell$ and $\lim_{x\to a} g(x)=m$, then
$$\lim_{x\to a} f(x)g(x)=\ell m.$$
- If $\lim_{x\to a} f(x)=\ell$ and $\ell\neq 0$, then
$$\lim_{x\to a} \tfrac{1}{f(x)}=\tfrac{1}{\ell}.$$
(Part of this problem requires that you show $f\neq 0$ in some punctured interval around $x=a$.)

Exercise 2.3. Prove the following derivative laws:

- If $f'(a)=\ell$ and $g'(a)=m$, then $[f+g]'(a)=\ell+m$.
- If $f'(a)=\ell$ and c is a real number, then $[c\cdot f]'(a)=c\ell$.
- If $f'(a)=\ell$, $g'(a)=m$, and g is nonzero in an interval containing a, then
$$(\tfrac{f}{g})'(a)=\tfrac{g(a)\ell-f(a)m}{[g(a)]^2}.$$

Exercise 2.4. Work out the details of (2.10). Try working out the cases $n=2,3$, to see the pattern.

Exercise 2.5. Let f be defined on an open interval I, with $a\in I$. Show that f is differentiable at $x=a$ if and only if there exists $\ell\in\mathbb{R}$ such that the function

$$g(x)=\begin{cases}\frac{f(x)-f(a)}{x-a} & x\in I\backslash\{a\}\\ \ell & x=a\end{cases}$$

is continuous at $x=a$. In this case, show that $\ell=f'(a)$.

Exercise 2.6. Prove an analogue of the '$C\varepsilon$ Theorem' (Theorem 2.1) for sequences of real numbers.

Exercise 2.7. Prove the following limit laws for sequences:

- If $a_n\to a$ and $b_n\to b$ as $n\to\infty$, then
$$ca_n\to ac,\quad a_n+b_n\to a+b,\quad a_nb_n\to ab,\quad\text{and}\quad \tfrac{a_n}{b_n}\to\tfrac{a}{b}$$
as $n\to\infty$. Here c is an arbitrary real number, and for the final limit we require that $b\neq 0$.

Exercise 2.8. Show that limits of convergent sequences are unique. That is, if $x_n\to\ell$ and $x_n\to m$ as $n\to\infty$, then $\ell=m$. (*Hint:* Argue that it is enough to show that for any $\varepsilon>0$, $|\ell-m|<2\varepsilon$. Then show that this holds.)

Exercise 2.9. Prove Lemma 2.8.

Exercise 2.10. Show that any convergent sequence of real numbers is a Cauchy sequence.

Exercise 2.11. Prove the continuity properties claimed in Examples 2.7 and 2.8.

Exercise 2.12. Show that if $\lim_{x\to a} f'(x)$ exists, then in fact f is differentiable at $x = a$ and $\lim_{x\to a} f'(x) = f'(a)$.

Exercise 2.13. Use the Mean Value Theorem to show that if f is a function on an open interval with $f' = 0$ at each point, then f is constant.

3. Infinite series and integration

In the preceding sections, we have encountered the concept of infinite series (both of numbers and of functions) at several points. In this section, we will put these concepts on firm mathematical footing. We will then discuss how the study of infinite series of functions provided the motivation for the theories of integration introduced by Cauchy and Riemann, which will be the final topic of this section.

3.1. **Infinite series.**

Let us first recall the definition of a convergent infinite series.

Definition 3.1 (Convergent series)**.** Suppose a_k is a sequence of real numbers. We say that the *infinite series* $\sum_{k=0}^{\infty} a_k$ *converges* to the value ℓ if

$$\text{for any}\quad \varepsilon > 0,\quad \text{there exists}\quad N \quad \text{such that}\quad n \geq N \implies \left|\sum_{k=0}^{n} a_k - \ell\right| < \varepsilon.$$

That is, $\sum a_k = \ell$ if the sequence of partial sums $\sum_{k=0}^{n} a_k$ converges to ℓ. If the sequence of partial sums does not converge, then we say the series *diverges.*

The notion of convergence of series can be subtle. For example, we saw in Section 1.6 that the harmonic series (given by $a_n = \frac{1}{n}$) has its individual terms tend to zero, but the series itself *diverges.* On the other hand, if instead of considering

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \cdots$$

we take

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \cdots, \tag{3.1}$$

then, even though the summands are the same size, this latter series converges to $\ln 2$ (we'll show this later).

To get started, we can at least show that if there is any hope for convergence, the individual summands must tend to zero. This is the 'divergence test' from calculus.

Proposition 3.1 (Divergence Test)**.** *If* $\sum a_k$ *converges, then* $\lim_{k\to\infty} a_k \to 0$.

Proof. Let $\varepsilon > 0$ and choose $N > 0$ so that

$$n \geq N \implies |S_n - \ell| < \varepsilon,$$

where $S_n = \sum_{k=0}^{n} a_k$ and ℓ is the value of the infinite series. Then since

$$a_n = S_{n+1} - S_n,$$

we obtain

$$n \geq N \implies |a_n| \leq |S_{n+1} - \ell| + |S_n - \ell| < 2\varepsilon.$$

□

To check our definition of convergence currently requires that we know the value of the infinite series. This is rather inconvenient. Fortunately, there is a very useful criterion due to Cauchy (around 1820) that allows us to check whether or not a series converges.

Theorem 3.2 (Cauchy criterion for convergence)**.** *Let* a_n *be a sequence of real numbers and* S_n *the corresponding partial sums. Then the series* $\sum a_n$ *converges if and only if for any* $\varepsilon > 0$, *there exists* N *so that*

$$n, m \geq N \implies |S_n - S_m| < \varepsilon.$$

Proof. The condition described in the theorem is just the statement that the partial sums form a Cauchy sequence (in the sense of Definition 2.9). Thus the convergence follows from the completeness of the real numbers, Proposition 2.9. On the other hand, convergence of the partial sums implies that the partial sums are Cauchy via Exercise 2.10. □

The Cauchy criterion implies the useful fact that if a series converges 'absolutely', then the series converges.

Definition 3.2 (Absolute Convergence)**.** Let a_n be a sequence of real numbers. If $\sum |a_n|$ converges, then we say that $\sum a_n$ *converges absolutely.*

Corollary 3.3 (Absolute Convergence Theorem)**.** *If a series converges absolutely, then it converges (in the usual sense).*

Proof. Let S_n be the partial sums of the a_n and T_n be the partial sums of the $|a_n|$. Then

$$|S_n - S_m| = \left|\sum_{k=m+1}^{n} a_k\right| \leq \sum_{k=m+1}^{n} |a_k| = T_n - T_m.$$

Thus if $\{T_n\}$ is a Cauchy sequence, so is $\{S_n\}$. □

The converse to this theorem, of course, is false, as shown by the series in (3.1). Thus we need one more definition:

Definition 3.3 (Conditional Convergence)**.** A series *converges conditionally* if it converges but does not converge absolutely.

We can produce many conditionally convergent series using the following:

Corollary 3.4 (Alternating series test)**.** *Let a_n be a nonnegative, decreasing sequence. Then the alternating series*

$$\sum_{n=1}^{\infty}(-1)^n a_n$$

converges if and only if $\lim_{n\to\infty} a_n = 0$.

Proof. The hypotheses guarantee that for any $m \geq n$,

$$S_n \leq S_m \leq S_{n+1} \quad \text{or} \quad S_{n+1} \leq S_m \leq S_n$$

(depending on the parity of n). Thus

$$|S_m - S_n| \leq |S_{n+1} - S_n| = |a_n|,$$

which yields convergence provided $a_n \to 0$. The converse follows from the divergence test. □

Example 3.1. The alternating series test guarantees convergence of the series in (3.1) (although it does not identify the value).

Example 3.2. The series

$$1 - \tfrac{1}{2} + \tfrac{1}{2} - \tfrac{1}{4} + \tfrac{1}{3} - \tfrac{1}{8} + \cdots + \tfrac{1}{n} - \tfrac{1}{2^n} + \cdots$$

has alternating summands that tend to zero. However, this series diverges. Indeed, the positive terms yield the harmonic series, while the negative terms yield a convergent geometric series. This does not contradict the alternating series test, because the absolute values of the terms are not strictly decreasing.

Another useful way to determine convergence of series is to make comparisons with known series. For example:

Theorem 3.5 (Comparison Test)**.** *Suppose a_n and b_n are positive sequences such that $a_n \leq b_n$ for all n. If $\sum b_n$ converges, then so does $\sum a_n$. If $\sum a_n$ diverges, so does $\sum b_n$.*

Proof. Let S_n be the partial sums for the a_n and T_n the partial sums for the b_n. Since the partial sums are increasing and $a_n \leq b_n$, we have

$$0 \leq S_n - S_m = \sum_{k=m+1}^{n} a_k \leq \sum_{k=m+1}^{n} b_k \leq T_n - T_m.$$

Thus if $\{T_n\}$ is Cauchy, so is $\{S_n\}$, that is, convergence of $\sum b_n$ guarantees convergence of $\sum a_n$. Stated in the contrapositive, we see that divergence of $\sum a_n$ guarantees divergence of $\sum b_n$ (see also Exercise 3.1). □

Note that we need the sequences to be positive in Theorem 3.5 Indeed, just consider $b_n \equiv 0$ and $a_n = -\frac{1}{n}$.

Other 'comparison' type tests rely on comparison with *geometric* series. Based on (1.1), we can see that series of the form $\sum r^n$ (called geometric series) will converge if $r < 1$ and diverge if $r \geq 1$. The ratio and root tests essentially allow us to make use of this fact.

Theorem 3.6 (Ratio test)**.** *Let a_n be a sequence of nonzero real numbers and let*

$$r(n) = \left| \frac{a_{n+1}}{a_n} \right|.$$

- (i) *If there exists $c < 1$ so that $r(n) \leq c$ for all n, then the series $\sum a_n$ converges absolutely.*
- (ii) *If $r(n) \geq 1$ for all n, then the series $\sum a_n$ diverges.*

Remark 3.7. Since convergence or divergence does not depend on the first N terms of the sequence for any fixed N, we see that the conditions in Theorem 3.6 only need to hold for all n sufficiently large.

Proof of Theorem 3.6. Suppose (i) holds. Then

$$|a_n| \leq c^n |a_1|,$$

and hence (since $c < 1$) we obtain convergence via the comparison test. On the other hand, if (ii) holds then $|a_n| \geq |a_1|$ for all n, and hence we cannot have $a_n \to 0$. Thus the series diverges. □

This result yields the following 'limit ratio test', which you will prove in Exercise 3.2.

Corollary 3.8 (Limit ratio test)**.** *Suppose a_n is a sequence of nonzero real numbers and $r(n) = \left|a_{n+1}/a_n\right|$. Suppose $\lim_{n\to\infty} r(n) = L$.*

- *If $L < 1$ then the series converges absolutely.*
- *If $L > 1$ then the series diverges.*
- *If $L = 1$ then the test is inconclusive.*

We next state the root test and limit root test, which you will prove in Exercise 3.3.

Theorem 3.9 (Root test, limit root test)**.** *Let a_n be a sequence of real numbers. Let*

$$\rho(n) = |a_n|^{1/n}.$$

Then:

(i) *If there exists $c < 1$ and $N \geq 1$ so that $\rho(n) \leq c$ for all $n \geq N$, then the series $\sum a_n$ converges absolutely.*
(ii) *If there exist arbitrarily large n so that $\rho(n) \geq 1$, then the series diverges.*

Consequently, if $\rho(n) \to L$, then we obtain:

- *If $L < 1$, then the series converges absolutely.*
- *If $L > 1$, then the series diverges.*
- *If $L = 1$, then the test is inconclusive.*

In the exercises, you will work out a few examples. As a rule of thumb, we may say that if the ratio test works for a series, the root test will too (e.g. if $r(n)$ has a limit, then so does $\rho(n)$ and the limits are equal), but that the root test is a more robust test. For example, if we consider $a_n = 2^{(-1)^n - n}$, then we have

$$r(n) = |\tfrac{a_{n+1}}{a_n}| = \tfrac{1}{2} \cdot 2^{-2(-1)^n} = \begin{cases} \frac{1}{8} & n \text{ odd}, \\ 2 & n \text{ even} \end{cases},$$

so that the ratio test cannot be applied. On the other hand,

$$\rho(n) = |a_n|^{\frac{1}{n}} = 2^{-1+(-1)^n/n} \to \tfrac{1}{2} \quad \text{as} \quad n \to \infty,$$

so that the root test yields absolute convergence. Of course, sometimes neither test works. For example, if $a_n = 1/n$ then we have both $r(n) \to 1$ and $\rho(n) \to 1$, so that both are inconclusive.

Cauchy devised one more test for convergence:

Theorem 3.10 (Cauchy's condensation test)**.** *Let a_n be a positive, decreasing sequence of real numbers. Then $\sum a_n$ converges if and only if the series*

$$a_1 + 2a_2 + 4a_4 + \cdots + 2^k a_{2^k} + \cdots$$

converges.

Proof. Consider a term like $4a_4$. By assumption, we can obtain

$$4a_4 = 2 \cdot (a_4 + a_4) \leq 2(a_4 + a_3).$$

Similarly,

$$8a_8 = 2 \cdot (a_8 + a_8 + a_8 + a_8) \leq 2 \cdot (a_8 + a_7 + a_6 + a_5),$$

and in general

$$2^n a_{2^n} \leq 2 \sum_{k=2^{n-1}+1}^{2^n} a_k.$$

From this, we obtain

$$\sum_{k=1}^{2^n} 2^k a_{2^k} \leq 2 \sum_{k=1}^{\infty} a_k \quad \text{for all} \quad n.$$

On the other hand, by the assumptions we have

$$a_2 + a_3 \leq 2a_2, \quad a_4 + a_5 + a_6 + a_7 \leq 4a_4,$$

and more generally

$$\sum_{k=2^n}^{2^{n+1}-1} a_k \leq 2^n a_n,$$

which yields

$$\sum_{k=1}^{m} a_k \leq \sum_{n=1}^{\infty} 2^n a_{2^n} \quad \text{for all} \quad m.$$

The result follows. □

As an application, we can show that the series

$$\sum_{n=2}^{\infty} \tfrac{1}{n \ln n}$$

diverges. Indeed, the convergence or divergence is equivalent to that of the series with general term

$$\tfrac{2^n}{2^n \ln(2^n)} = \tfrac{1}{n \ln 2},$$

which we know diverges. Another important application is the following:

Corollary 3.11 (p-test)**.** *The series*

$$\sum_{n=1}^{\infty} \tfrac{1}{n^p}$$

converges for $p > 1$ and diverges for $p \leq 1$.

Proof. We use Cauchy's condensation test, which shows that convergence of $\sum n^{-p}$ is equivalent to the convergence of the series with general term $2^n(2^n)^{-p} = 2^{-n(p-1)}$. This is a geometric series that converges if and only if $p > 1$. □

We will wait to discuss the last common test, namely, the *integral test*, until we have properly introduced integrals below.

Before turning to the important topic of series of *functions*, we discuss some additional questions related to summing infinite series.

There are a few simple observations that we can make immediately. For example, if $\sum a_n = A$ and $\sum b_n = B$, then $\sum(a_n + b_n) = A + B$. Similarly, $\sum(ca_n) = cA$.

More interesting is the question of what happens if we start regrouping or rearranging the terms of an infinite series? We start with a silly example, namely, what happens when we try to sum the series $1 - 1 + 1 - 1 + \cdots$. If we group this as

$$1 - 1 + (1 - 1) + (1 - 1) + \cdots$$

then we seem to get the answer zero, while if we group this as

$$1 + (-1 + 1) + (-1 + 1) + \cdots$$

then we seem to get the answer one. Of course, the problem here is that the series does not converge! (This follows immediately from the divergence test.) If we start with a convergent series, however, then regrouping is fine:

Theorem 3.12 (Regrouping of convergent series)**.** *If $\sum a_n$ converges to A, then we can regroup summands without changing the value of the series.*

We leave the proof as an exercise (see Exercise 3.6).

A more interesting situation arises if we consider what happens if we *rearrange* the terms of a convergent series. It was Riemann who ultimately figured out what happens, with the results appearing posthumously in the late 1860s. As it turns out, the situation is very different for absolutely convergent versus conditionally convergent series.

For absolutely convergent series, there is no problem with rearrangements:

Theorem 3.13 (Rearranging Absolutely Convergent Series)**.** *Suppose $\sum a_n$ converges absolutely, with value A. If $\{b_n\}$ is any rearrangement of the sequence $\{a_n\}$, then $\sum b_n$ converges absolutely to A.*

Proof. We first establish absolute convergence. In fact, this is straightforward, since any partial sum $\sum_{n=1}^{N} |b_n|$ is bounded by some partial sum $\sum_{n=1}^{M} |a_n|$, which is in turn uniformly bounded (by absolute convergence).

Next, let $\varepsilon > 0$ and let $N > 0$ to be determined below. We then consider a partial sum $T_n = \sum_{k=1}^{n} b_k$ for some $n \geq N$. Now, if we choose m large enough, then the summands in the partial sum $S_m = \sum_{k=1}^{m} a_k$ include all of the terms $\{b_k\}_{k=1}^{n}$. In particular, any extra summands must correspond to some b_k for $k \geq n \geq N$. Thus we may write

$$S_m - T_n = \sum_{k \in X_{n,m}} b_k$$

for some set $X_{n,m} \subset \{N, N+1, \cdots\}$. Using convergence of S_m to A and absolute convergence of $\sum b_k$, we can therefore choose N sufficiently large and then m sufficiently large to obtain

$$|T_n - A| \leq |T_n - S_m| + |S_m - A| \leq \sum_{k \in X_{n,m}} |b_k| + \varepsilon \leq 2\varepsilon$$

for all $n \geq N$. We conclude that $\sum b_k = A$. □

For conditionally convergent series, we instead have the following striking result, which tells us that a conditionally convergent series can be rearranged to take on any value we like!

Theorem 3.14 (Rearranging Conditionally Convergent Series)**.** *Suppose $\sum a_n$ converges conditionally. Then for any $L \in \mathbb{R}$, there exists a rearrangement $\{b_n\}$ of the sequence $\{a_n\}$ such that $\sum b_n = L$.*

Proof. Let a_n^+ be the positive terms in $\{a_n\}$ and a_n^- the negative terms. We first observe that the series $\sum a_n^{\pm}$ must both diverge. Indeed, suppose $\sum a_n^+$ converges. Then since $\sum a_n$ converges, we deduce that $\sum a_n^-$ converges as well. However, this implies

$$\sum a_n^+ - \sum a_n^- = \sum_n |a_n|$$

converges, contradicting the hypotheses. Similarly, $\sum a_n^-$ must diverge. We also note that we may assume without loss of generality that both of $\{a_n^{\pm}\}$ are in decreasing order (in magnitude). By convergence of $\sum a_n$, we have that $a_n^{\pm} \to 0$ as $n \to \infty$.

Now fix $L \in \mathbb{R}$. Let's suppose (without loss of generality) that $L > 0$. Since $\sum a_n^+$ diverges and $\{a_n^+\}$ is in decreasing order, we may choose the smallest m_1 so that

$$\sum_{n\leq m_1} a_n^+ \geq L.$$

We take the first m_1 terms of our rearrangement to be $\{a_n^+\}_{n\leq m_1}$. We then take the smallest m_2 so that

$$\sum_{n\leq m_1} a_n^+ + \sum_{n\leq m_2} a_n^- < L,$$

and take $\{a_n^-\}_{n\leq m_2}$ to be the next m_2 terms in our rearrangement. We now repeat this, choosing the minimal $m_3 > m_1$ so that

$$\sum_{n\leq m_1} a_n^+ + \sum_{n\leq m_2} a_n^- + \sum_{m_1<n\leq m_3} a_n^+ \geq L,$$

taking $\{a_n^+\}_{m_1<n\leq m_3}$ as the next terms in our rearrangement, and then choosing minimal $m_4 > m_2$ so that

$$\sum_{n\leq m_1} a_n^+ + \sum_{n\leq m_2} a_n^- + \sum_{m_1<n\leq m_3} a_n^+ + \sum_{m_2<n\leq m_4} a_n^- < L.$$

The fact that $\sum a_n^+$ and $\sum a_n^-$ each diverge means that we can continue this process indefinitely, thus building our rearrangement $\{b_n\}$.

It remains to show that $\sum b_n = L$. To this end, we let $\varepsilon > 0$ and (recalling $a_n \to 0$ as $n\to\infty$) firstly choose N_1 large enough that $|b_n| < \varepsilon$ for all $n \geq N_1$. We then take $N_2 \geq N_1$ to be the next index such that

$$\sum_{n\leq N_2} b_n \geq L \quad \text{but} \quad \sum_{n\leq N_2+1} b_n < L. \tag{3.2}$$

It follows that for all $N \geq N_2$, we have

$$\Big|\sum_{n\leq N} b_n - L\Big| < \varepsilon.$$

Indeed, from (3.2) we see that L differs both from $\sum_{n\leq N_2} b_n$ and $\sum_{n\leq N_2+1} b_n$ by less than ε. Each time we add a new summand, we either move closer to L or jump across the value of L by an amount less than ε. □

3.2. **Series of functions.** Some of the most important applications of applications involve taking infinite summations not just of numbers, but of functions. We have seen two such classes of examples, namely, *power series* and *trigonometric/Fourier series.*

We must first make precise the notion of convergence of an infinite series of functions.

Definition 3.4 (Pointwise convergence)**.** Let $f_n(x)$ be a sequence of real-valued functions. Then we say the series $\sum f_n(x)$ *converges pointwise* to a function $F(x)$ if for each x, we have $\sum f_n(x) = F(x)$ in the sense of infinite series of real numbers. That is:

For all x and for all $\varepsilon > 0$, there exists N so that

$$n \geq N \implies \Big|F(x) - \sum_{k=1}^{n} f_n(x)\Big| < \varepsilon.$$

Remark 3.15. In general, we say that a sequence of functions F_n converges pointwise to a function F if for all x we have $F_n(x) \to F(x)$ as $n \to \infty$. Then the definition above just says that the partial sums converge pointwise.

Given a series of functions, we would like to describe the values of x for which the series converges. We first consider the important special case of *power series*, in which case $f_n(x) = a_n x^n$ for some real coefficients a_n. One of our main motivating problems is that of finding power series representations for various functions. Indeed, one of our first examples (in Proposition 1.1) was the following:

Example 3.3. Let $a \in \mathbb{R}$ and $x \in \mathbb{R}$ with $|x| < 1$. Then Newton tells us that

$$(1+x)^a = 1 + ax + \tfrac{a(a-1)}{2!}x^2 + \tfrac{a(a-1)(a-2)}{3!}x^3 + \cdots \tag{3.3}$$

We expect this to hold because the coefficients appearing on the right-hand side are precisely given by $\frac{1}{n!}\frac{d^n}{dx^n}(1+x)^a\big|_{x=0}$. However, the first question we should ask is whether the series on the right-hand side even converges. For this, we can use the ratio test. We find that the radio of consecutive terms is given by

$$r(n) = \big|\tfrac{a-(n-1)}{n}\big|\,|x| = \big[1 - \tfrac{1+a}{n}\big]|x|.$$

In particular, $\lim_{n\to\infty} r(n) = |x|$, and so we obtain convergence for $|x| < 1$ and divergence for $|x| > 1$. The case $|x| = 1$ is unclear at this point.

Note that even if the series converges, this does not guarantee us the equality claimed in (3.3). That is actually a more difficult question, which we may return to later.

Example 3.4 (Exponential function)**.** The infinite series

$$\sum_{n=0}^{\infty} \tfrac{1}{n!}x^n$$

is the power series for the exponential function e^x (which we have not formally introduced yet). In this case,

$$r(n) = \lim_{n\to\infty} \tfrac{|x|}{n} = 0 \quad \text{for all} \quad x \in \mathbb{R},$$

so that the series converges for all $x \in \mathbb{R}$.

Both of the previous examples were centered around $x = 0$, although we can readily speak of power series centered around other points (so that the terms x^n are replaced by $(x - x_0)^n$ for some $x_0 \in \mathbb{R}$).

Power series (when they converge) have some very nice properties. The first is the fact that when power series converge, they will do so in an interval that is symmetric about the center point (although the situation at the endpoints typically needs to be investigated separately). The definition we need is the following:

Definition 3.5 (Radius of convergence)**.** Consider a power series of the form $\sum a_n(x - x_0)^n$. We call B the *radius of convergence* for the series if the series converges for $|x - x_0| < B$ and diverges for $|x - x_0| > B$.

To determine the radius of convergence, we can often rely on the root test. To make this precise, we introduce one additional notion (basically due to Cauchy) known as upper limits (or ‘lim sup’s).

Definition 3.6 (Upper limit)**.** Let $\{x_1, x_2, \dots\}$ be a bounded sequence, and for each k let M_k denote the supremum of $\{x_k, x_{k+1}, \dots\}$. Then the *upper limit* (or $\limsup$) of the sequence $\{x_k\}$ is defined to be the infimum of the set $\{M_k\}$. We denote this quantity by

$$\limsup_{k\to\infty} x_k.$$

There is a related notion of lower limits (or '$\liminf$'s). For example, for the sequence

$$\{x_k\} = \{.9,\ 2.1,\ .99,\ 2.01,\ .999,\ 2.001, \dots\}$$

to compute the upper limit we first consider the sequence of suprema

$$\{M_k\} = \{2.1,\ 2.1,\ 2.01,\ 2.01,\ 2.001,\ 2.001, \dots\},$$

the infimum of which is 2. Thus $\limsup x_k = 2$. Similarly, $\liminf x_k = 1$. If a sequence converges, then the $\limsup$ and $\liminf$ both equal the limit; the converse of this is true as well.

Here are the essential facts we will need about the $\limsup$ in a moment: *If $M > \limsup x_k$, then there exists N such that $x_n < M$ for all $n \geq N$. If $m < \limsup x_k$, then there are infinitely many n such that $x_n \geq m$.*

Now consider some power series, say $\sum a_n x^n$, and take the sequence

$$S = \{|a_n|^{1/n}\}.$$

If S is unbounded above (in which case we say the upper limit is infinite), then so is $\{|a_n|^{1/n}|x|\}$ so by the root test the series diverges for all $x \neq 0$. We say the radius of convergence is zero. If S is bounded, then it has a finite upper limit. We will prove the following:

Theorem 3.16 (Radius of convergence)**.** *Consider the power series $\sum a_n x^n$ and define*

$$R = \frac{1}{\limsup |a_n|^{1/n}}.$$

Then the series converges absolutely for $|x| < R$ and diverges for $|x| > R$.

If $\limsup |a_n|^{1/n} = 0$, the series converges for all $x \in \mathbb{R}$.

If $\limsup |a_n|^{1/n} = \infty$, then the series converges only at $x = 0$.

Proof. We write

$$\ell = \limsup_{n\to\infty} |a_n|^{1/n}.$$

(i) If $|x| < \frac{1}{\ell}$, then we can find $\alpha < 1$ and $\varepsilon > 0$ so that

$$|x| < \tfrac{\alpha}{\ell+\varepsilon}.$$

Consequently,

$$|a_n x^n|^{1/n} < |a_n|^{1/n}|x| < \tfrac{|a_n|^{1/n}}{\ell+\varepsilon}\alpha.$$

Now, by the first fact about $\limsup$s above, we have that

$$|a_n|^{1/n} < \ell + \varepsilon \quad \text{for all} \quad n \quad \text{sufficiently large.}$$

This implies

$$|a_n x^n|^{1/n} < \alpha < 1 \quad \text{for all} \quad n \quad \text{sufficiently large,}$$

which yields convergence by the root test.

(ii) If $|x| > \frac{1}{\ell}$, then we can find $\varepsilon > 0$ so that

$$|x| > \tfrac{1}{\ell - \varepsilon}.$$

Then

$$|a_n x^n|^{1/n} = |a_n|^{1/n}|x| > \tfrac{|a_n|^{1/n}}{\ell - \varepsilon}.$$

By the second fact about lim sups above, there are infinitely many n such that $|a_n|^{1/n} \geq \lambda - \varepsilon$. Thus there are infinitely many n so that $|a_n x^n|^{1/n} \geq 1$. By the root test, this yields divergence.

The remaining statements follow similarly, so we will end the proof here. (You are encouraged to check the details!) □

The situation is basically the same for power series centered at some other $x_0 \in \mathbb{R}$. We will not bother writing down everything in detail.

We have not so far considered the interesting question of what happens at the endpoints of the interval of convergence. There are many possibilities, for example: convergence or divergence at both endpoints, or convergence at just one of the end points. When convergence holds, it could be absolute or conditional. In practice, we basically have to proceed on a case by case basis, but there is at least one special class of series that can be understood completely. These are so-called *hypergeometric series*, for which the ratio of consecutive terms satisfies

$$\tfrac{a_{n+1}}{a_n} = \tfrac{P(n)}{Q(n)} \quad \text{for some polynomials} \quad P, Q.$$

This special class, which includes many familiar examples, was studied in depth by Gauss, who developed a comprehensive test for convergence around 1815. However, this particular result will not be important for our applications, so we will not go into it in any detail.

We briefly turn to another important class of infinite series of functions, namely, *Fourier* or *trigonometric series.* As described in Section 1.5, these series were introduced in order to solve certain physically-motivated problems. The basic premise is to represent a function $f(x)$ as a linear combination of sines and cosines of higher and higher frequencies. Let us return to the specific example described in (1.5), i.e. the claim that

$$\tfrac{\pi}{4} = \cos(\tfrac{\pi x}{2}) - \tfrac{1}{3}\cos(\tfrac{3\pi x}{2}) + \tfrac{1}{5}\cos(\tfrac{5\pi x}{2}) + \cdots \quad \text{for} \quad x \in (-1, 1) \tag{3.4}$$

None of the tests we have developed so far can say much about convergence of this series. For $x = 0$, we get the alternating harmonic series (which we have seen before), but in general we get a series with absolute values decaying like $1/n$ but which does not alternate. Techniques for determining the convergence of series like (3.4) were developed by the mathematicians Abel and Dirichlet (in the 1820s). For example, we have:

Theorem 3.17 (Abel's Lemma)**.** *Consider an infinite series of the form*

$$\sum a_k b_k,$$

where the b_k are positive and decreasing. Suppose there exists $M > 0$ so that

$$\left|\sum_{k=1}^{n} a_k\right| \leq M \quad \textit{for all} \quad n.$$

Then

$$\left|\sum_{k=1}^{n} a_k b_k\right| \leq M b_1 \quad \textit{for all} \quad n.$$

Corollary 3.18 (Dirichlet's Test). *Assume the same hypotheses as in Abel's Lemma. If in addition $\lim_{n\to\infty} b_n = 0$, then the series $\sum a_k b_k$ converges.*

We will not prove these results here. Instead, let us just briefly show how Dirichlet's test may be applied to (3.4). We fix $x \in (-1,1)$ and view the terms in (3.4) as $a_k b_k$ where $b_k = \frac{1}{2k-1}$ and $a_k = (-1)^{k+1}\cos(\frac{k\pi x}{2})$. Now, to make sense of the partial sums $\sum a_k$, one can derive the following trigonometric identity:

$$\sum_{k=1}^{n} (-1)^{k-1} \cos\left[\frac{(2k-1)\pi x}{2}\right] = \frac{1-(-1)^n \cos(\pi n x)}{2\cos(\pi x/2)}.$$

This can be deduced by applying Euler's formula $e^{iy} = \cos y + i \sin y$ and the formula for summing a geometric series. Using this identity, we obtain the bound

$$\left|\sum_{k=1}^{n} (-1)^{k-1} \cos\left[\frac{(2k-1)\pi x}{2}\right]\right| \leq |\sec(\pi x/2)| \quad \text{for any} \quad n.$$

In particular, for any $x \in (-1,1)$, we obtain the uniform bounds required by Abel's Lemma and Dirichlet's Test, and so we conclude that this strange series in (3.4) does in fact converge for $x \in (-1,1)$. Technically, we have not identified the values the series converges to, but let us take for granted that the value is $\frac{\pi}{4}$ for each $x \in (-1,1)$ (recall the graph plotted in Section 1.5). Now, at the endpoints, $x = \pm 1$, our bound seems to blow up. However, if we plug in $x = \pm 1$, the summands in the series (3.4) are identically zero. So something strange is happening here. In particular, we are taking an infinite series of continuous functions and obtaining the *discontinuous* limit

$$f(x) = \begin{cases} \frac{\pi}{4} & x \in (-1,1) \\ 0 & x = \pm 1. \end{cases} \tag{3.5}$$

It seems that is is finally time to take a closer look at some questions mentioned briefly above, namely, whether properties of functions (like continuity and differentiability) are preserved under taking infinite sums. Often these are described as questions of the *interchange of limits.* Let us first formally state the question of continuity.

Question 3.1. *Suppose the infinite series $\sum_n f_n(x)$ converges to the function $F(x)$. If each f_n is continuous, is the limit F necessarily continuous?*

Evidently, the answer to this question (as currently stated) is 'no', as the Fourier series example (3.4) showed us. But let us at least see what this question has to do with interchanging limits: Recall that continuity of F at a point x_0 is the statement that

$$\lim_{x\to x_0} F(x) = F(x_0).$$

Writing $F(x) = \lim_{n\to\infty} S_n(x)$, with $S_n(x) = \sum_{k=1}^{n} f_k(x)$, and recalling that each f_k is continuous (and hence so is each S_n), we see that

$$\lim_{x\to x_0} F(x) = \lim_{x\to x_0}\lim_{n\to\infty} S_n(x),$$
$$F(x_0) = \lim_{n\to\infty}\lim_{x\to x_0} S_n(x),$$

so that this question exactly boils down to whether or not we can interchange the two limits.

Cauchy's 1821 work on analysis actually included the statement (and proof) that the infinite sum of continuous functions is a continuous function. The Fourier series we just studied shows that this is wrong (although technically we did not prove that the limit function is given by (3.5)). We can provide a simpler counterexample as well:

Example 3.5. Consider the infinite series

$$S(x) = \sum_{k=1}^{\infty} \frac{x^2}{(1+kx^2)(1+(k-1)x^2)},$$

whose summands are continuous and whose coefficients decay like $\frac{1}{k^2}$ (yielding convergence). The partial sums can be computed explicitly:

$$S_n(x) = \tfrac{nx^2}{1+nx^2},$$

which shows that the limit $S(x)$ is the *discontinuous* function

$$S(x) = \begin{cases} 1 & x \neq 0, \\ 0 & x = 0. \end{cases}$$

The fact that Cauchy's theorem was not correct as stated was observed by Abel in 1826. It was around 25 years later that Cauchy finally corrected his error (after some clarifying work by other mathematicians), with more clarification coming from the work of Weierstrass in the 1860s. In particular, to conclude that the infinite sum of continuous functions is again continuous, one needs a slightly stronger notion than pointwise convergence (in the sense of Definition 3.4). The notion is that of *uniform* convergence.

Definition 3.7 (Uniform convergence)**.** Let $f_n(x)$ be a sequence of real-valued functions. Then we say the series $\sum f_n(x)$ *converges uniformly* to a function $F(x)$ if for all $\varepsilon > 0$, there exists N so that

$$n \geq N \implies \left| F(x) - \sum_{k=1}^{n} f_k(x) \right| < \varepsilon \quad \text{for all} \quad x.$$

Remark 3.19. In general, we say that a sequence of functions F_n converges uniformly to F if for all $\varepsilon > 0$, there exists N so that

$$n \geq N \implies |F(x) - F_n(x)| < \varepsilon \quad \text{for all} \quad x.$$

Then the definition above just says that the partial sums converge uniformly.

The difference between this definition and Definition 3.4 is that for a given ε, we must be able to find a single choice of N that works simultaneously for *all* choices of x. In Definition 3.4, the choice of N was implicitly allowed to depend on both $\varepsilon > 0$ and the choice of the point x.

Example 3.6. Let us return to the series in (3.5). In this case, the partial sums

$$S_n(x) = \tfrac{nx^2}{1+nx^2}$$

do *not* converge uniformly to the limit

$$S(x) = \begin{cases} 1 & x \neq 0, \\ 0 & x = 0. \end{cases}$$

To see this consider

$$|S(x) - S_n(x)| = \begin{cases} \frac{1}{1+nx^2} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

In particular, for any fixed n, we have

$$\lim_{x\to 0} |S(x) - S_n(x)| = \lim_{x\to 0} \tfrac{1}{1+nx^2} = 1.$$

Thus for any $\varepsilon \in (0,1)$ and any $N \geq 0$, we can x sufficiently close to zero that

$$|S(x) - S_N(x)| > \varepsilon.$$

That is, given $\varepsilon \in (0,1)$, there is no single choice of N that can work simultaneously for all choices of x.

On the other hand, uniform convergence *is* enough to establish continuity of the limit.

Theorem 3.20 (Continuity for Infinite Series)**.** *Suppose f_n is a sequence of continuous real-valued functions on an interval (a,b) and the infinite series $\sum_n f_n$ converges uniformly to the function F on (a,b). Then F is continuous on (a,b).*

Proof. Let $x \in (a,b)$ and $\varepsilon > 0$. By *uniform* convergence, we may find N so that

$$\left| F(z) - \sum_{k=1}^N f_k(z) \right| < \varepsilon \quad \text{for all} \quad z \in (a,b).$$

By continuity of the *finite* sum $\sum_{k=1}^N f_k$, we may find $\delta > 0$ so that

$$|y - x| < \delta \implies \left| \sum_{k=1}^N [f_k(x) - f_k(y)] \right| < \varepsilon.$$

Thus for $|x - y| < \delta$, we have by the triangle inequality that

$$\begin{aligned} &|F(x) - F(y)| \\ &\leq \left| F(x) - \sum_{k=1}^N f_k(x) \right| + \left| \sum_{k=1}^N [f_k(x) - f_k(y)] \right| + \left| \sum_{k=1}^N f_k(y) - F(y) \right| \qquad (3.6) \\ &\leq 3\varepsilon. \end{aligned}$$

□

If we return to the series in Example 3.5, this proof would break down precisely because we cannot find a single choice fo N that works simultaneously for all values of y in (3.6).

Uniform convergence is a natural condition to check that suffices to establish continuity of the limit. Although, as it turns out, it is not a *necessary* condition. In particular, there are examples of series of continuous functions that converge non-uniformly to a continuous limit. See Exercise 3.11.

We turn to the question of differentiating infinite series. Let us state the question formally:

Question 3.2. *Suppose the infinite series $\sum_n f_n(x)$ converges to the function $F(x)$. If each f_n is differentiable, is the limit F necessarily differentiable, with $F'(x) = \sum_n f_n'(x)$?*

This is once again a question of the interchange of limits. Indeed, we are asking whether

$$\tfrac{d}{dx}\sum_n f_n = \sum_n \tfrac{d}{dx} f_n,$$

where both of $\frac{d}{dx}$ and $\sum_n$ are defined via limits. Given the subtleties involved with understanding continuity, it should not be surprising that the answer to Question 3.2 (as stated) is 'no'. Let's see what can go wrong with a few examples.

Example 3.7. Let's return to our favorite Fourier series (3.4). Taking for granted that

$$\tfrac{\pi}{4} = \cos(\tfrac{\pi x}{2}) - \tfrac{1}{3}\cos(\tfrac{3\pi x}{2}) + \tfrac{1}{5}\cos(\tfrac{5\pi x}{2}) + \cdots \quad \text{for} \quad x \in (-1,1),$$

we differentiate term by term and arrive at the claim

$$0 = -2[\sin(\tfrac{\pi x}{2}) - \sin(\tfrac{3\pi x}{2}) + \sin(\tfrac{5\pi x}{2}) - \cdots] \quad \text{for} \quad x \in (-1,1).$$

Well, this is embarrassing. Unless x is an even integer, these coefficients do not tend to zero. Thus, by the divergence test, the series on the right-hand side does not even converge! Whoops...

Example 3.8. Let

$$f_k(x) = \frac{x^3}{(1+kx^2)(1+(k-1)x^2)} = \frac{kx^3}{1+kx^2} - \frac{(k-1)x^3}{1+(k-1)x^2},$$

which satisfy

$$f_k'(0) = 0 \quad \text{for all} \quad k.$$

The series $\sum f_k$ converges for each x (the coefficients decay like $\frac{1}{k^2}$). In fact, the partial sums are given by

$$S_n(x) = \tfrac{nx^3}{1+nx^2}, \quad \text{which converges to} \quad S(x) = x.$$

In particular, $S'(x) \equiv 1$, and so we have

$$\tfrac{d}{dx}\sum_{k=1}^{\infty} f_k = 1, \quad \text{while} \quad \sum_{k=1}^{\infty} \tfrac{d}{dx} f_k = 0 \quad \text{at} \quad x = 0.$$

As with continuity, requiring *uniform* convergence is enough to obtain a positive result. This time, we need uniform convergence of the series of *derivatives.*

Theorem 3.21 (Term-by-term differentiation)**.** *Let f_k be a sequence of functions so that $\sum f_k$ converges at $x = a$ and the series of derivatives $\sum f_k'$ converges uniformly on an open interval I containing a. Then:*

- *The series $F(x) := \sum_{k=1}^{\infty} f_k(x)$ converges uniformly on I,*
- *The function F is differentiable on I, and*
- *$F'(x) = \sum_{k=1}^{\infty} f_k'(x)$ for $x \in I$.*

Proof. We define

$$g_k(x) = \begin{cases} \frac{f_k(x)-f_k(a)}{x-a} & x \in I\backslash\{a\}, \\ f_k'(a) & x = a, \end{cases}$$

which are continuous on I by Exercise 2.5.

We let $F_n(x) = \sum_{k=1}^{\infty} f_k(x)$ denote the partial sums of the f_k. Then each F_n is differentiable on I, with $F_n'(x) = \sum_{k=1}^{\infty} f_k'(x)$.

We will first show that the series $\sum_k g_k$ is uniformly Cauchy on I. To see this, note that by applying the definition of g_k and the Mean Value Theorem (to the function $F_n - F_m$), a difference of partial sums at some $x \neq a$ is of the form

$$\begin{aligned}\sum_{k=m+1}^{n} g_k(x) &= \tfrac{F_n(x)-F_n(a)}{x-a} - \tfrac{F_m(x)-F_m(a)}{x-a} \\ &= \tfrac{F_n(x)-F_m(x)-[F_n(a)-F_m(a)]}{x-a} \\ &= F_n'(t) - F_m'(t) \\ &= \sum_{k=m+1}^{n} f_k'(t)\end{aligned}$$

for some t between x and a. At $x = a$ we simply have the sum $\sum_{k=m+1}^{n} f_k'(a)$. In particular, using *uniform* convergence of $\sum f_k'$, we deduce that for any $\varepsilon > 0$ there exists N so that

$$m, n \geq N \implies \Big| \sum_{k=m+1}^{n} g_k(x) \Big| < \varepsilon \quad \text{for any} \quad x \in I.$$

This means that the series $\sum_k g_k$ is 'uniformly Cauchy'. By Exercise 3.10, this implies that $\sum_k g_k$ converges uniformly.

The uniform convergence of $\sum g_k$ now implies uniform convergence of $\sum f_k$ on the entire interval I, since

$$f_k(x) = g_k(x)[x-a] + f_k(a).$$

We write the limit of the series $\sum f_k(x)$ as $F(x)$. In particular, we have that

$$\sum_{k=1}^{\infty} g_k(x) = \frac{F(x)-F(a)}{x-a} \quad \text{for} \quad x \in I \backslash \{a\}.$$

We will now show that $F'(a)$ exists and equals $\sum f_k'(a)$. (In fact, this will actually be enough to conclude that $F'(x) = \sum f_k'(x)$ for all $x \in I$, since we have now established convergence of $\sum f_k$ at each $x \in I$.) For any $x \in I\backslash\{a\}$ and any n, we write

$$\begin{aligned}&\Big| \frac{F(x)-F(a)}{x-a} - \sum_{k=1}^{\infty} f_k'(a) \Big| \\ &\leq \Big| \frac{F(x)-F(a)}{x-a} - \frac{F_n(x)-F_n(a)}{x-a} \Big| \\ &\quad + \Big| \frac{F_n(x)-F_n(a)}{x-a} - F_n'(a) \Big| + \Big| F_n'(a) - \sum_{k=1}^{\infty} f_k'(a) \Big| \\ &\leq \Big| \sum_{k=n+1}^{\infty} g_k(x) \Big| + \Big| \frac{F_n(x)-F_n(a)}{x-a} - F_n'(a) \Big| + \Big| \sum_{k=n+1}^{\infty} f_k'(a) \Big|.\end{aligned}$$

Now let $\varepsilon > 0$. By the (uniform) convergence of the series $\sum g_k$ and $\sum f_k'$, we may choose n sufficiently large that the first and third terms above are less than ε, uniformly in the choice of x. Given this fixed value of n, we may then choose $\delta > 0$ so that

$$0 < |x-a| < \delta \implies \left| \frac{F_n(x) - F_n(a)}{x-a} - F_n'(a) \right| < \varepsilon.$$

Putting this all together, we find that

$$0 < |x-a| < \delta \implies \left| \frac{F(x) - F(a)}{x-a} - \sum_{k=1}^{\infty} f_k'(a) \right| < 3\varepsilon,$$

yielding $F'(a) = \sum_k f_k'(a)$, as desired. □

If we return to Example 3.8, then we can see that the issue is that the partial sums of the derivatives, namely,

$$S_n'(x) = \tfrac{3nx^2 + n^2x^4}{n^2x^4 + 2nx^2 + 1}$$

do not converge uniformly (see Exercise 3.12).

We have seen that the notion of uniform convergence is important if we wish to understand properties of convergent series. In general, this fact was not fully appreciated until around the 1860s. We will discuss a test for uniform convergence due to Weierstrass in 1880 known as the 'Weierstrass M-test', which we will then apply in the setting of power series.

Theorem 3.22 (Weierstrass M-test)**.** *Suppose $f_k(x)$ is a sequence of functions on an interval I and M_k is a sequence of real numbers such that*

$$|f_k(x)| \le M_k \quad \text{for all} \quad x \in I \quad \text{and} \quad k \ge 1. \tag{3.7}$$

If $\sum_k M_k$ converges, then $\sum f_k$ converges uniformly on I.

Proof. We use the Cauchy criterion for uniform convergence (see Exercise 3.10). We let $\varepsilon > 0$ and choose N sufficiently large that

$$n, m \ge N \implies \sum_{k=m+1}^{n} M_k < \varepsilon.$$

Then for $n, m \ge N$, we obtain

$$\left| \sum_{k=m+1}^{n} f_k(x) \right| \le \sum_{k=m+1}^{n} |f_k(x)| \le \sum_{k=m+1}^{n} M_k < \varepsilon$$

for all $x \in I$. This implies that the series is 'uniformly Cauchy' on I, and hence uniformly convergent on I. □

It is a straightforward extension to see that (3.7) only needs to hold for all k sufficiently large. Let's see what this result says about power series.

Corollary 3.23 (Uniform convergence for power series)**.** *Suppose $\sum_{n=0}^{\infty} a_n x^n$ is a power series with a radius of convergence R.*

(i) *If $0 < \alpha < R$, then the series converges uniformly on $[-\alpha, \alpha]$.*
(ii) *If R is finite and the series converges at $x = R$, then it converges uniformly on $[0, R]$.*

We will only prove part (i) of this result. Part (ii) relies on Abel's Lemma (Theorem 3.17), which we did not actually prove.

Proof of (i). We will focus on the case $R < \infty$, leaving the details for $R = \infty$ to you. By definition of the radius of convergence, we have

$$\limsup_{k\to\infty} |a_k|^{1/k} = \tfrac{1}{R}.$$

We fix $0 < \alpha < R$. We then take $|x| < \alpha$ and set $\varepsilon := (\alpha - |x|)/|x|$. We may find N so that

$$k \geq N \implies |a_k|^{1/k} < \tfrac{1+\varepsilon}{R} = \tfrac{\alpha}{R|x|}.$$

In particular,

$$|a_k x^k| \leq (\tfrac{\alpha}{R})^k.$$

Since $\frac{\alpha}{R} < 1$, we may apply the Weierstrass M-test to deduce uniform convergence on $(-\alpha, \alpha)$. Combining this with convergence at the endpoints, we obtain uniform convergence on $[-\alpha, \alpha]$ (see Exercise 3.13). □

The last question we will pose in this section about infinite series and interchange of limit operations is the following:

Question 3.3. *Suppose the infinite series $\sum_n f_n(x)$ converges on an interval $[a, b]$. Does it follow that*

$$\int_a^b \sum_n f_n(x)\,dx = \sum_n \int_a^b f_n(x)\,dx \quad ?$$

Considering we have not even given a proper definition of integration yet, we will not try to answer this question right now. However, let us see why it is an important question. It has to do with the formula derived by Fourier for trigonometric series. Suppose f is an even function on $(-1, 1)$ that we would like to expand in a cosine series of the form

$$f(x) = \sum_{k=1}^{\infty} a_k \cos\bigl[\tfrac{(2k-1)\pi x}{2}\bigr]. \tag{3.8}$$

If we accept that such a decomposition exists, then we can determine what the coefficients a_k must be. It is based on the fact that

$$\int_{-1}^{1} \cos\bigl[\tfrac{(2k-1)\pi x}{2}\bigr] \cos\bigl[\tfrac{(2m-1)\pi x}{2}\bigr]\,dx = \begin{cases} 1 & k = m, \\ 0 & k \neq m, \end{cases}$$

which can be derived using Euler's formula, as long as you are comfortable taking integrals of complex-valued functions. Then, *if* (3.8) holds *and* we are allowed to interchange intergration and summation, we must have

$$\begin{aligned} &\int_{-1}^{1} f(x) \cos\bigl[\tfrac{(2m-1)\pi x}{2}\bigr]\,dx \\ &\quad = \int_{-1}^{1} \sum_{k=1}^{\infty} a_k \cos\bigl[\tfrac{(2k-1)\pi x}{2}\bigr] \cos\bigl[\tfrac{(2m-1)\pi x}{2}\bigr]\,dx \\ &\quad = \sum_{k=1}^{\infty} a_k \cdot \begin{cases} 1 & k = m \\ 0 & k \neq m \end{cases} \quad = \quad a_m, \end{aligned}$$

providing a formula for computing the coefficients.

At this point we see that a better understanding of integration (including questions related to interchange of limits) is going to be necessary in order to resolve issues related to the existence/convergence of Fourier series. This will be the topic of our next section.

3.3. **Integration.** Up until the 1820s, integration was essentially identified with antidifferentiation. This is already unsatisfactory if we wish to integrate functions that are not the derivative of some other function (e.g. functions with jump discontinuities). Many questions that arise in the study of convergence of Fourier series also demand a better theory of integration. In this section, we introduce the theories of integration developed by Cauchy in the 1820s and Riemann in the 1850s. (In fact, an even more robust theory of integration was later introduced by Lebesgue in 1904; however, we will focus our attention here primarily on the theory of Riemann integration, which is the one we encounter in calculus courses.) In all of these theories of integration, the idea is to make precise the notion that the integral $\int_a^b f\,dx$ should represent the 'area under the curve f for x between a and b'.

We first present Cauchy's definition of the integral.

Definition 3.8 (Cauchy integral)**.** A real-valued function f is *Cauchy integrable* on an interval $[a,b]$ with value V, denoted

$$\int_a^b f(x)\,dx = V,$$

if the following holds: for any $\varepsilon > 0$, there exists $\delta > 0$ so that for any partition

$$a = x_0 < x_1 < \cdots < x_n = b$$

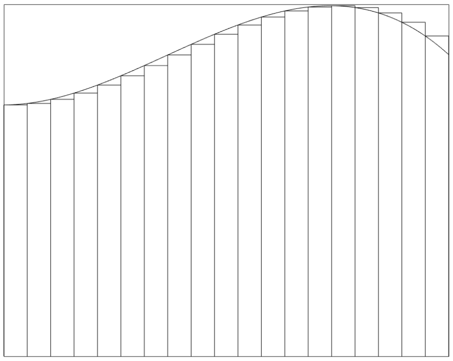
of $[a,b]$ satisfying

$$|x_j - x_{j-1}| < \delta \quad \text{for all} \quad j,$$

we have

$$\left| \sum_{j=1}^{n} f(x_{j-1}) \cdot (x_j - x_{j-1}) - V \right| < \varepsilon.$$

The situation is depicted in the following figure.



The Riemann definition is very similar, but is a bit more flexible. In particular, instead of evaluating f at the left endpoint of the interval $[x_{j-1}, x_j]$, it can be evaluated at any point in this interval.

Definition 3.9 (Riemann integral)**.** A real-valued function f is *(Riemann) integrable* on an interval $[a,b]$ with value V, denoted

$$\int_a^b f(x)\,dx = V,$$

if the following holds: for any $\varepsilon > 0$, there exists $\delta > 0$ so that for any partition

$$a = x_0 < x_1 < \cdots < x_n = b$$

of $[a,b]$ satisfying

$$|x_j - x_{j-1}| < \delta \quad \text{for all} \quad j, \tag{3.9}$$

and any $x_j^* \in [x_{j-1}, x_j]$, we have

$$\left| \sum_{j=1}^{n} f(x_j^*) \cdot (x_j - x_{j-1}) - V \right| < \varepsilon. \tag{3.10}$$

It is straightforward to see that any Riemann integrable function is Cauchy integrable (you should make sure that this is clear to you). In what follows, we will focus on Riemann's notion of integrability, since this is the theory of integration that is most commonly used in the calculus setting.

You may recall from your calculus course that the sums appearing in (3.10) are known as *Riemann sums*. We will use the same terminology here. We will write $S[f]$ for a general Riemann sum adapted to a partition of $[a,b]$. If a partition obeys (3.9), we say it has *width* $< \delta$.

The Riemann integral satisfies many reasonable properties, which we list here (and whose proofs we relegate to the exercises).

Theorem 3.24 (Properties of the Riemann integral)**.** *Suppose f, g are Riemann integrable on $[a,b]$ and $c \in \mathbb{R}$. Then $f + g$ is integrable and $c \cdot f$ is integrable, with*

$$\int_a^b [f(x) + g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx, \qquad \int_a^b cf(x)\,dx = c\int_a^b f(x).$$

Moreover, if $f(x) \geq 0$ for all $x \in [a,b]$, then

$$\int_a^b f(x)\,dx \geq 0.$$

More generally, if $m \leq f(x) \leq M$ for all $x \in [a,b]$, then

$$m(b-a) \leq \int_a^b f(x)\,dx \leq M(b-a).$$

Our first question is how to determine whether a given function is Riemann integrable. The definition seems to require that we already know the value of the integral, so this is not particularly helpful in practice. Fortunately, there is a relatively straightforward 'Cauchy criterion' for establishing Riemann integrability. In particular, we have the following:

Lemma 3.25 (Cauchy criterion)**.** *A function f is Riemann integrable on an interval $[a,b]$ if and only if for any $\varepsilon > 0$, there exists $\delta > 0$ so that*

$$|S_1[f] - S_2[f]| < \varepsilon$$

whenever $S_1[f]$ and $S_2[f]$ are Riemann sums for f corresponding to partitions of $[a,b]$ of width $< \delta$.

Proof. $\Longrightarrow$: If f is integrable and $\varepsilon > 0$ is given, then we may find a width $\delta > 0$ so that

$$\left| S[f] - \int_a^b f(x)\,dx \right| < \varepsilon/2$$

whenever $S[f]$ is a Riemann sum corresponding to a partition of width $< \delta$. Then if $S_1[f]$, $S_2[f]$ are any two such partitions, we immediately obtain

$$\left| S_2[f] - S_1[f] \right| < \varepsilon$$

by the triangle inequality.

$\Longleftarrow$: Suppose the Cauchy criterion holds. For any $n = 1, 2, 3, \ldots$, we let $S_n[f]$ by a sequence of Riemann sums of f corresponding to partitions of width $< 1/n$. It follows that $\{S_n[f]\}$ is a Cauchy sequence of real numbers, and hence has a limit V. We will show that $\int_a^b f(x)\,dx$ exists and equals V.

To this end, let $\varepsilon > 0$ and choose $\delta > 0$ as in the Cauchy criterion. Now let $S[f]$ be any Riemann sum corresponding to a partition of width $< \delta$, and let n be large enough that $|S_n[f] - V| < \varepsilon$ and $1/n < \delta$. Then

$$|S[f] - V| \leq |S[f] - S_n[f]| + |S_n[f] - V| < 2\varepsilon,$$

showing that f is integrable with $\int_a^b f(x)\,dx = V$. □

The Cauchy criterion, while convenient for use in proofs, still does not give any real insight into what types of functions are integrable. Our next few results will give some more insight in this direction.

Our first main result shows that a function f is integrable on $[a, b]$ if and only if it can be bounded from above and below by two step functions. Here a *step function* is a function g such that there exists a partition $x_0, \ldots, x_N$ of $[a, b]$ with g equal to some constant c_i on each (x_{i-1}, x_i). Note that in this definition, we don't care what happens at the endpoints x_i. It is not difficult to show that step functions are Riemann integrable. With g as just described, we have

$$\int_a^b g(x)\,dx = \sum_{i=1}^{N} c_i(x_i - x_{i-1}).$$

See Exercise 3.16.

Proposition 3.26. *A function f on $[a, b]$ is integrable if and only if for any $\varepsilon > 0$, there exist step functions f_1 and f_2 on $[a, b]$ so that*

$$f_1(x) \leq f(x) \leq f_2(x) \quad \textit{for} \quad x \in [a, b], \quad \textit{with} \quad \int_a^b [f_2(x) - f_1(x)]\,dx < \varepsilon$$

Proof. $\Longleftarrow$: Let $\varepsilon > 0$, and select step functions f_1, f_2 as in the statement of the proposition. By integrability of f_1, f_2, we may find $\delta > 0$ so that any Riemann sum corresponding to a partition of width $< \delta$ for f_1, f_2 differ from the value of the integral by at most ε. Now consider a Riemann sum

$$S[f] = \sum_{i=1}^{N} f(x_i^*)(x_i - x_{i-1}), \quad x_i^* \in [x_{i-1}, x_i],$$

corresponding to a partition of width $< \delta$. By assumption,

$$\sum_{i=1}^{N} f_1(x_i^*)(x_i - x_{i-1}) \leq S[f] \leq \sum_{k=1}^{N} f_2(x_i^*)(x_i - x_{i-1}).$$

This implies

$$S[f] \in \biggl(\int_a^b f_1(x)\,dx - \varepsilon, \int_a^b f_2(x)\,dx + \varepsilon \biggr).$$

By the assumptions on f_1, f_2, this is an interval of length 3ε.

In particular, we have just shown that any two Riemann sums for f corresponding to partitions of width $< \delta$ belong to this interval, meaning their difference is bounded by 3ε. By the Cauchy criterion, we conclude that f is integrable.

$\Longrightarrow$: Now suppose that f is integrable and let $\varepsilon > 0$. Using the Cauchy criterion, we find a partition $x_0 < x_1 < \cdots < x_N$ so that

$$\biggl| \sum_{i=1}^{N} (f(x_i^*) - f(x_i^{**}))(x_i - x_{i-1}) \biggr| < \varepsilon$$

for any choice of $x_i^*, x_i^{**} \in [x_{i-1}, x_i]$. In particular, this implies (by suitable choice of x_i^*, x_i^{**}) that

$$|(f(x_j^*) - f(x_j))(x_j - x_{j-1})| < \varepsilon \quad \text{for each} \quad j.$$

In particular, by the triangle inequality,

$$|f(x_j^*)| < \tfrac{\varepsilon}{x_j - x_{j-1}} + |f(x_j)| \quad \text{for all} \quad j \quad \text{and all} \quad x_j^* \in [x_{j-1}, x_j].$$

This shows that f is bounded on each $[x_{j-1}, x_j]$, and hence on $[a, b]$. We can therefore define m_i and M_i to be the infimum and supremum of f on each $[x_{i-1}, x_i]$ and define the step functions as follows:

First, f_1 is equal to m_i on (x_{i-1}, x_i). At the endpoints we set $f_1 = \min\{m_1, \dots, m_N\}$.

Second, f_2 is equal to M_i on (x_{i-1}, x_i). At the endpoints we set $f_2 = \min\{M_1, \dots, M_N\}$.

It follows that $f_1 \leq f \leq f_2$ on $[a, b]$. It remains to show that

$$\int_a^b [f_2(x) - f_1(x)]\,dx \leq \varepsilon. \tag{3.11}$$

To this end, we fix arbitrary $\eta > 0$ and (using the definition of infimum and supremum) find choices of $x_i^*, x_i^{**} \in [x_{i-1}, x_i]$ so that

$$f(x_i^*) < m_i + \eta, \quad f(x_i^{**}) > M_i - \eta.$$

It follows that

$$\begin{aligned} \sum_{i=1}^{N} (f(x_i^{**}) - f(x_i^*))(x_i - x_{i-1}) &> \sum_{i=1}^{N} (M_i - m_i - 2\eta)(x_i - x_{i-1}) \\ &= \int_a^b (f_2(x) - f_1(x))\,dx - 2\eta(b - a). \end{aligned}$$

Since

$$\biggl| \sum_{i=1}^{N} (f(x_i^{**}) - f(x_i^*))(x_i - x_{i-1}) \biggr| < \varepsilon,$$

we deduce

$$\int_a^b (f_2(x) - f_1(x))\,dx < \varepsilon + 2\eta(b - a).$$

As η was arbitrary, this implies (3.11), as desired. □

As a corollary, we deduce one necessary condition for Riemann integrability:

Corollary 3.27. *If f is Riemann integrable on $[a,b]$, then f is bounded on $[a,b]$.*

Proof. This was derived in the course of the proof of the preceding proposition. □

At this point, you might feel a bit confused, since you probably were able to do computations like

$$\int_{-1}^{1} |x|^{-\frac{1}{2}}\, dx = 4, \tag{3.12}$$

even though $|x|^{-1/2}$ is not bounded on $[-1,1]$ (or even defined at $x=0$). To resolve this confusion, one can introduce the notion of an *improper integral.* This refers to an integral in which either the function or the interval is unbounded. Then, for example, the integral (3.12) must be interpreted as the following limit:

$$\int_{-1}^{1} |x|^{-\frac{1}{2}}\, dx = \lim_{\varepsilon_1\to 0^-} \int_{-1}^{\varepsilon_2} |x|^{-\frac{1}{2}}\, dx + \lim_{\varepsilon_2\to 0^+} \int_{\varepsilon_1}^{1} |x|^{-\frac{1}{2}}\, dx = 4.$$

For improper integrals over unbounded intervals, see the exercises.

As another important consequence of Proposition 3.26, we can establish a very useful *sufficient* condition for Riemann integrability, namely, that of *continuity.* However, there is a subtlety here, similar to the subtlety we encountered when we wanted to establish continuity for infinite series of continuous functions. In fact, Cauchy again missed this subtlety when presenting a proof that continuous functions were integrable. In particular, to establish integrability, we need something slightly stronger than continuity, namely, *uniform continuity.*

Let us first review the definition of continuity in its complete ε-δ glory:

Definition 3.10 (Continuity)**.** Let f be a real-valued function on an interval I. Then f is continuous on $[a,b]$ if for every $x\in I$ and every $\varepsilon>0$, there exists $\delta>0$ so that for any $y\in I$,

$$|x-y|<\delta \implies |f(x)-f(y)|<\varepsilon.$$

Here is the definition of *uniform* continuity:

Definition 3.11 (Uniform continuity)**.** Let f be a real-valued function on an interval I. Then f is *uniformly continuous* if for every $\varepsilon>0$, there exists $\delta>0$ so that for any $x,y\in I$,

$$|x-y|<\delta \implies |f(x)-f(y)|<\varepsilon.$$

Can you see the difference between these two definitions? If you look closely at the statement of continuity, you'll see that the 'δ' you pick could implicitly depend on the choice of $x\in[a,b]$ as well as $\varepsilon>0$. On the other hand, for a uniformly continuous function, we can find a single $\delta>0$ that works *uniformly* in x (that is, for all choices of x simultaneously).

Example 3.9. Consider the function $f(x)=\frac{1}{x}$ on $(0,1)$. This function is continuous, but not *uniformly continuous.* If we instead consider $f(x)=\frac{1}{x}$ on the interval $(\delta,1)$ for any strictly positive $\delta>0$, the function is uniformly continuous. See Exercise 3.18.

The reason that Cauchy's oversight was inconsequential in this particular case is due to the following fact:

Theorem 3.28. *If f is continuous on $[a,b]$, then f is* uniformly *continuous on $[a,b]$.*

Proof. We suppose f is continuous on $[a,b]$ and fix $\varepsilon>0$. For each $x\in[a,b]$, we then define the 'modulus of continuity' by

$$\delta(x,\varepsilon)=\sup\{\eta>0:|x-y|<\eta \implies |f(x)-f(y)|<\varepsilon\}. \tag{3.13}$$

By pointwise continuity, we have $\delta(x)>0$ for each $x\in[a,b]$. To establish uniform continuity, it suffices to show that

$$\inf_{x\in[a,b]}\delta(x,\varepsilon)>0.$$

Suppose instead that this infimum equals zero. This implies that there exists a sequence $\{x_n\}\subset[a,b]$ so that $\delta_n:=\delta(x_n,\varepsilon)\to 0$. Because the sequence $\{x_n\}$ is stuck inside the closed, bounded interval $[a,b]$, it must have a convergent 'subsequence'. For convenience, we are still going to denote this subsequence by x_n, and we'll denote the limit by $x_*\in[a,b]$. We'll return to this crucial point below. For now, let us finish the argument.

We may also define the 'modulus of continuity' $\delta_*:=\delta(x_*,\frac{\varepsilon}{2})>0$ as in (3.13) above. Using the fact that $x_n\to x_*$ and $\delta_n\to 0$, we have that for all n sufficiently large,

$$|x_n-x_*|<\tfrac12\delta_* \quad\text{and}\quad \delta_n<\tfrac14\delta_*.$$

These together guarantee (by the triangle inequality) that

$$\{y:|x_n-y|<\delta_n+\tfrac{1}{10}\delta_*\}\subset\{y:|x_*-y|<\delta_*\}.$$

On the other hand, by continuity of f at x_*, we have that $f(x_n)\to f(x_*)$. Thus, choosing n possibly even larger, we may also guarantee that

$$|f(x_n)-f(x_*)|<\tfrac{\varepsilon}{2}.$$

Thus we find that for n sufficiently large and

$$|x_n-y|<\delta_n+\tfrac{1}{10}\delta_*, \tag{3.14}$$

we have $|x_*-y|<\delta_*$, and hence

$$|f(x_n)-f(y)|\le|f(x_n)-f(x_*)|+|f(x_*)-f(y)|<\tfrac{\varepsilon}{2}+\tfrac{\varepsilon}{2}=\varepsilon. \tag{3.15}$$

In particular, (3.14)–(3.15) show that δ_n is *not* the supremum as in (3.13); indeed, we have found a strictly larger choice of 'η' that does the job! Thus we derive our desired contradiction. □

The proof above relied on the following important fact, sometimes called the Bolzano–Weierstrass theorem:

Theorem 3.29 (Bolzano-Weierstrass)**.** *If $\{x_n\}$ is a sequence of real numbers contained in a closed, bounded interval $[a,b]$, then a subsequence of x_n converges to some limit in $[a,b]$.*

This is our first real encounter with the concept of 'compactness'. Essentially, the theorem says that closed, bounded intervals in $\mathbb{R}$ are 'compact'. Note that this result would not hold if the interval were either unbounded or open. Indeed, the sequence $1,2,3,\dots$ in the interval $(0,\infty)$ has no convergent sequence; neither does the sequence $1, 1/2, 1/3, \dots$ in $(0,1)$. Actually, in the latter case, the sequence converges to 0, but the point 0 is not contained in the interval.

I am not going to provide a proof of Theorem 3.29 at this point. I encourage you to try to find one yourself (see Exercise 3.21)!

With Theorem 3.28 in place, we return to our discussion of integrability and prove the following:

Theorem 3.30 (Continuity implies integrability)**.** *If f is continuous on $[a,b]$, then f is Riemann integrable on $[a,b]$.*

Proof. We have just shown that in fact, f must be *uniformly* continuous. We will use the criterion in Proposition 3.26 to show integrability.

Let $\varepsilon > 0$ and, by uniform continuity, choose $\delta > 0$ so that

$$|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Now let $x_0 < x_1 < \cdots < x_N$ be a partition of $[a,b]$ of width $< \delta$. Write M_i and m_i for the maximum and minimum values of f on $[x_{i-1}, x_i]$ (which are attained, as shown in Theorem 2.14). By the uniform continuity, we have $M_i - m_i < \varepsilon$ for each i.

Now define step functions f_1, f_2 that are equal to m_i, M_i on each (x_{i-1}, x_i) and equal to $f(x)$ at the endpoints of the intervals. It follows that $f_1 \leq f \leq f_2$ on $[a,b]$. Moreover, we have

$$\int_a^b [f_2(x) - f_1(x)]\,dx = \sum_{i=1}^N [M_i - m_i](x_i - x_{i-1}) < \varepsilon \sum_{i=1}^N (x_i - x_{i-1}) = \varepsilon(b-a).$$

The result follows. □

The preceding result gives a very natural condition (continuity) that can be used to guarantee integrability. However, not every integrable function needs to be continuous. Indeed, we already saw that step functions are integrable. In fact, Riemann produced an example of an integrable function with *infinitely* many discontinuities!

It is also possible to write down explicit functions that are *not* Riemann integrable. The classic example is the Dirichlet function defined

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational,} \end{cases}$$

defined on $[0,1]$. To prove this relies on a fact that we have not proven yet (but will prove later):

- *Any interval $I \subset \mathbb{R}$ contains both a rational and irrational number.*

Now consider any partition of $[0,1]$ of width $\delta > 0$. We build two Riemann sums S_1, S_2 adapted to this partition, where in each subinterval we pick a rational point to build S_2 and an irrational point to build S_1. Then we get

$$\big|S_2 - S_1\big| = \sum_{j=1}^N [1 - 0] \cdot (x_j - x_{j-1}) = 1.$$

That is, no matter how small we take the width of partition, we can find two Riemann sums that differ by 1. This shows that the Cauchy criterion for integrability fails.

Now that we have some understanding of Riemann integration, let us briefly return to the motivating problem of the convergence of Fourier series. With a reasonable theory of integration in place, Dirichlet was able to prove the following result:

Theorem 3.31 (Dirichlet's Theorem)**.** *Suppose F is a bounded, piecewise continuous on $[-\pi,\pi]$. Suppose further that at any point x_0 of discontinuity, F takes on the value at the midpoint of $\lim_{x\to x_0^+}F(x)$ and $\lim_{x\to x_0^-}F(x)$. If we define the coefficients*

$$\begin{aligned} a_0 &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} F(x)\,dx, \\ a_k &= \tfrac{1}{\pi}\int_{-\pi}^{\pi} F(x)\cos(kx)\,dx, \quad k\geq 1, \\ b_k &= \tfrac{1}{\pi}\int_{-\pi}^{\pi} F(x)\sin(kx)\,dx, \quad k\geq 1, \end{aligned}$$

then for every $x\in(-\pi,\pi)$ we have

$$F(x) = a_0 + \sum_{k=1}^{\infty}[a_k\cos(kx)+b_k\sin(kx)].$$

We won't prove this result here. In fact, this result is far from the end of the story, as far as convergence of Fourier series goes. Really, one should proceed using the theory of integration due to Lebesgue. If you're interested, you will have to sign up for more courses in analysis.

Instead, we will turn to some other important results related to Riemann integration. Specifically, we will work towards the important Fundamental Theorem of Calculus. One of the forms of this theorem states that

$$\int_a^b f'(x)\,dx = f(b)-f(a).$$

For the original 'Newtonian' notion of integration, this is not a theorem—it's a definition! For the Riemann integral, we will need a proof.

We start with the following:

Proposition 3.32. *Let $a<b<c$ and let f be a real-valued function on $[a,c]$. Then f is integrable on $[a,c]$ if and only if f is integrable on $[a,b]$ and $[b,c]$, and in this case we have*

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx.$$

Proof. We use the 'step function' criterion for integrability. Note that the result we are trying to prove is immediate for step functions.

First suppose f is integrable on $[a,b]$ and $[b,c]$. Then given $\varepsilon>0$, we can find step functions adapted to each of these intervals and patch them together to find suitable step functions adapted to the entire interval adapted to $[a,c]$.

Conversely, if f is integrable on $[a,c]$, then given $\varepsilon>0$ we may find step functions as in Proposition 3.26. We can then split these step functions up to find that the 'step function criterion' holds on each of $[a,b]$ and $[b,c]$.

To show the equality, we note that if S_1 and S_2 are Riemann sums for f on $[a, b]$ and $[b, c]$, then $S_1 + S_2$ is a Riemann sum for f on $[a, c]$. Using this, it is not hard to show that the difference between

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx \quad \text{and} \quad \int_a^c f(x)\,dx$$

may be made arbitrarily small. □

We turn to the the Fundamental Theorem of Calculus in its first form:

Theorem 3.33 (Fundamental Theorem of Calculus)**.** *Let f be a continuous, real-valued function on $[a, b]$. Define F on $[a, b]$ by*

$$F(x) = \int_a^x f(t)\,dt.$$

Then F is differentiable and $F' = f$.

Proof. First note that by the continuity of f, the integral $\int_a^x f(t)\,dt$ exists for any $x \in [a, b]$. Here we use the convention that $\int_a^a f\,dt = 0$.

For fixed $x_0 \in [a, b]$ and $x \in [a, b]\backslash\{x_0\}$, we write

$$\begin{aligned}\frac{F(x) - F(x_0)}{x - x_0} - f(x_0) &= \frac{\int_a^x f(t)\,dt - \int_a^{x_0} f(t)\,dt}{x - x_0} - f(x_0) \\ &= \frac{\int_{x_0}^x f(t)\,dt - \int_{x_0}^x f(x_0)\,dt}{x - x_0} \\ &= \frac{\int_{x_0}^x [f(t) - f(x_0)]\,dt}{x - x_0}.\end{aligned}$$

Now, given $\varepsilon > 0$, we choose $\delta > 0$ so that

$$|t - x_0| < \delta \implies |f(t) - f(x_0)| < \varepsilon.$$

In particular, we have $|f(t) - f(x_0)| < \varepsilon$ for all t in the interval between x_0 and x. Here we are using the convention that if $a < b$,

$$\int_b^a f(t)\,dt := -\int_a^b f(t)\,dt.$$

We next use the fact that if g is integrable on an interval I with $|g| \leq M$, then

$$\left|\int_I g(x)\,dx\right| \leq M \cdot |I|,$$

where $|I|$ is the length of the interval (see the last inequality in Theorem 3.24).

Continuing from above, we deduce that for $|x - x_0| < \delta$, we have

$$\left|\frac{F(x) - F(x_0)}{x - x_0} - f(x_0)\right| \leq \left|\frac{\int_{x_0}^x [f(t) - f(x_0)]\,dt}{x - x_0}\right| \leq \varepsilon \cdot \tfrac{x - x_0}{x - x_0} = \varepsilon.$$

It follows that

$$\lim_{x \to x_0} \frac{F(x) - F(x_0)}{x - x_0} = f(x_0).$$

□

As a corollary, we obtain the other form of the Fundamental Theorem of Calculus:

Corollary 3.34 (Fundamental Theorem of Calculus)**.** *Suppose F is a real-valued function on an interval I with continuous derivative f. Then for any $a, b \in I$,*

$$\int_a^b f(t)\,dt = F(b) - F(a).$$

Proof. By the first form of the Fundamental Theorem of Calculus,

$$\tfrac{d}{dx}\left[\int_a^x f(t)\,dt - F(x)\right] = f(x) - f(x) = 0,$$

and hence the function

$$x \mapsto \int_a^x f(t)\,dt - F(x) \quad \text{is constant}$$

(see Exercise 2.13). In particular,

$$\int_a^x f(t)\,dt = F(x) + c \quad \text{for some} \quad c \in \mathbb{R}.$$

Evaluating at $x = a$, we obtain $c = -F(a)$, and hence

$$\int_a^x f(t)\,dt = F(x) - F(a) \quad \text{for all} \quad x.$$

□

The Fundamental Theorem of Calculus is, of course, extremely useful. This is how we evaluate integrals! We learned how this works in calculus, so let's move on to another application.

In particular, we are going to use the Riemann integral to give a very efficient presentation of the logarithm and exponential functions. In contrast to much of what we have presented above, this presentation decidedly does *not* follow the historical development of these functions. The basic idea behind the development of the logarithm was the search for a function satisfying $f(xy) = f(x)+f(y)$, so that large multiplication problems could be converted to simple addition problems. With a sufficiently comprehensive table of values $(x, f(x))$, one could compute efficiently in this way. The logarithm in the sense we understand it was introduced basically in the early 1600s. Given that we are about to define the logarithm as a Riemann integral, we are clearly deviating from the historical presentation at this point.

Definition 3.12. For $x \in \mathbb{R}$ and $x > 0$, define

$$\ln x = \int_1^x \tfrac{dt}{t}.$$

The properties of this function are collected in the following proposition, whose proof is left to you:

Proposition 3.35. *The function $x \mapsto \ln x$ on $(0, \infty)$ is differentiable with derivative $\frac{1}{x}$. It is a strictly increasing function, its range equals $\mathbb{R}$, and it satisfies:*

- $\ln xy = \ln x + \ln y$,
- $\ln x/y = \ln x - \ln y$,
- $\ln x^n = n \ln x$ *for any integer n.*

In particular, the properties of ln show that it has a well defined inverse:

Definition 3.13. We define $x \mapsto \exp(x)$ as the inverse of ln. That is,

$$\exp(x) = y \iff x = \ln y.$$

Later we will write $\exp(x) = e^x$ (see below).

The properties of the exponential function are collected in the following proposition, whose proof is also left to you.

Proposition 3.36. *The function $x \mapsto \exp(x)$ is differentiable with $\frac{d}{dx}\exp(x) = \exp(x)$. It is strictly increasing, assumes all positive values, and satisfies:*

- $\exp(x) \cdot \exp(y) = \exp(x+y)$,
- $\exp(x)/\exp(y) = \exp(x-y)$,
- $\exp(nx) = (\exp x)^n$ *for any integer n.*

Using the exponential and logarithm functions, we can finally precisely define the meaning of x^n for n not equal to an integer:

Definition 3.14. For $x, n \in \mathbb{R}$ with $x > 0$, we define $x^n = \exp(n \ln x)$.

Using this definition, you can readily check all of the usual properties of powers, like

$$x^n \cdot x^m = x^{n+m}, \quad (xy)^n = x^n y^n,$$

and so on. In addition, the chain rule implies $\frac{d}{dx}x^n = nx^{n-1}$.

We can also at this point introduce the number 'e':

$$e := \exp(1).$$

Then the notation $\exp(x) = e^x$ becomes consistent.

We end this discussion by remarking that it is possible to 'go backwards' in the construction of exponential and logarithm functions. That is, one can start by arguing for the existence of rational powers, obtaining irrational powers by limits, and then discovering e as the unique value of a so that $\frac{d}{dx}[a^x] = a^x$. With the exponential function in place, one can then define the logarithm.

To finish this section, we are going to return to two topics that were previously skipped because we had not yet introduced any reasonable theory of integration. The first is an integral test for convergence of series; the second is a result on term-by-term integration.

Theorem 3.37 (Integral Test)**.** *Let f be a positive, decreasing, integrable function on $[1,\infty]$. The series*

$$\sum_{k=1}^{\infty} f(k)$$

converges if and only if the improper integral

$$\int_1^{\infty} f(x)\,dx$$

converges.

Proof. As f is decreasing, we have

$$f(k+1) \le \int_k^{k+1} f(x)\,dx \le f(k).$$

Thus

$$\sum_{k=1}^{N} f(k+1) \leq \sum_{k=1}^{N} \int_k^{k+1} f(x)\,dx = \int_1^N f(x)\,dx \leq \sum_{k=1}^{N} f(k).$$

Using these inequalities, we see that the partial sums are bounded if and only if the integral converges. □

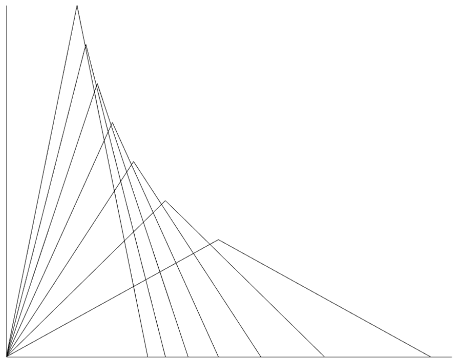
Example 3.10. The integral test shows that

$$\sum_{k=2}^{\infty} \tfrac{1}{k\ln k} \quad \text{diverges, while} \quad \sum_{k=2}^{\infty} \tfrac{1}{k[\ln k]^2} \quad \text{converges.}$$

Indeed, $\frac{1}{x\ln x}$ is the derivative of $\ln[\ln x]$, which diverges as $x \to \infty$, while $\frac{1}{x[\ln x]^2}$ is the derivative of $-\frac{1}{\ln x}$, which tends to zero as $x \to \infty$.

Next, we turn to term-by-term integration. By now it should not be too surprising that things can go wrong in general, but that the situation can be remedied with uniform convergence.

Example 3.11. Let $F_n(x)$ be the function on $[0,1]$ whose graph is an isosceles triangle whose base is $[0,\frac{1}{n}]$ and whose height is $2n$, as plotted in the following figure:



The functions F_n for $3 \leq n \leq 9$

In particular,

$$\int_0^1 F_n(x)\,dx = 1 \quad \text{for all} \quad n.$$

Now define $f_1 = F_1$ and $f_n = F_n - F_{n-1}$ for $n \geq 2$, so that

$$\sum_{k=1}^{n} f_k(x) = F_n(x).$$

Now, for any $x \in [0,1]$, we have $F_n(x) \to 0$ as $n \to \infty$. Thus we have

$$\int_0^1 \lim_{n\to\infty} F_n(x)\,dx = 0, \quad \text{while} \quad \lim_{n\to\infty} \int_0^1 F_n(x) = 1,$$

showing the failure of term-by-term integration in this case.

For the positive result, we have the following:

Theorem 3.38 (Term-by-term integration)**.** *Let $\sum_k f_k$ converge uniformly on $[a,b]$. If each f_k is integrable over $[a,b]$, then so is $\sum_k f_k$, with*

$$\int_a^b \sum_{k=1}^{\infty} f_k(x)\,dx = \sum_{k=1}^{\infty} \int_a^b f_k(x)\,dx.$$

Proof. Let

$$F(x) = \sum_{k=1}^{\infty} f_k(x) \quad \text{and} \quad F_n(x) = \sum_{k=1}^{n} f_k(x).$$

Now let $\varepsilon > 0$. By uniform convergence, we may find n so that

$$F_n(x) - \varepsilon \le F(x) \le F_n(x) + \varepsilon.$$

Since F_n is integrable, we may (by Proposition 3.26) find step functions g_1, g_2 so that $g_1 \le F_n \le g_2$ and

$$\int_a^b [g_2(x) - g_1(x)]\,dx < \varepsilon.$$

Then $g_1 - \varepsilon$ and $g_2 + \varepsilon$ are step functions satisfying

$$g_1(x) - \varepsilon \le F(x) \le g_2(x) + \varepsilon \quad \text{for} \quad x \in [a,b]$$

and

$$\int_a^b [g_2(x) + \varepsilon] - [g_1(x) - \varepsilon]\,dx < \varepsilon + 2\varepsilon(b-a).$$

Using Proposition 3.26, we deduce that F is integrable.

To identify the value of $\int F\,dx$, we note that for any $\varepsilon > 0$ we may find N so that

$$n \ge N \implies |F(x) - F_n(x)| < \varepsilon \quad \text{for all} \quad x \in [a,b].$$

Then for $n \ge N$, we have

$$\left| \int_a^b [F(x) - F_n(x)]\,dx \right| < \varepsilon(b-a).$$

Since

$$\int_a^b F_n(x) = \sum_{k=1}^{n} \int_a^b f_k(x)\,dx$$

for any fixed n, we find that

$$n \ge N \implies \left| \int_a^b \sum_{k=1}^{\infty} f_k(x)\,dx - \sum_{k=1}^{n} \int_a^b f_k(x)\,dx \right| < \varepsilon(b-a).$$

This implies that the series $\sum_k \int_a^b f_k\,dx$ converges, with value given by $\int_a^b \sum_k f_k\,dx$. □

Example 3.12. We can now sketch the justification of the procedure for approximating π given way back in Section 1.3. This required the series expansion for the binomial series $(1+x)^a$, which we revisited in Example 3.3. Recall we apply this series with $x = -t^2$ and $a = \frac{1}{2}$. We previously showed convergence on $(-1,1)$; here we need uniform convergence on $[-1,0]$. It turns out at that at $|x| = 1$ we can use a comparison with the p-series $n^{-3/2}$ to obtain convergence. Hence by Corollary 3.23 we have the desired uniform convergence to interchange integration and summation.

As for the fact that the series converges to $(1+x)^a$, there is a clever equation using a uniqueness result about differential equations, which we won't go into here.

So, it took about 200 years of work, but it seems as though Newton was justified in his term-by-term integration after all! Of course, I am being intentionally glib here. The rigorous theory of analysis was also essential in making sense of the much more complicated ideas of Fourier analysis, which has had truly remarkable impacts on our modern world. Indeed, just to give one example, this basic notion of decomposing functions into the waves (and the mathematical theory needed to make this precise) was central to the development of electromagnetism, which truly did shape the nature of modern life.

On the other hand, the deeper investigation into the possible behaviors of functions that was needed to understand concepts such as continuity and integrability revealed all sorts of strange possibilities that mathematicians had never considered before. This includes all sorts of bizarre functions introduced in the 1860s and later (e.g. integrable functions with infinitely many continuities, continuous functions with infinitely many points of non-differentiability, and so on). Research in this general direction is connected to the development of subjects like topology and measure theory (including the Lebesgue integral, which has become the dominant theory of integration in the last 100+ years). However, this is not the direction we will follow in this course.

Instead, we will turn to the work of mathematicians like Dedekind and Cantor from the 1870s onward, who realized that the work of Cauchy and others had uncovered some important questions about the very foundations of mathematics, namely, the nature of the real number system. You may recall that at many points above, we had to appeal to the 'completeness' of the real numbers, which so far we have basically assumed to be true because our 'intuition' tells us so. Well, by now you have hopefully also learned that intuition alone is not enough! It is time to finally talk about what we really mean by 'the real numbers', and to figure out if they really behave the way we think they do.

3.4. **Exercises.**

Exercise 3.1. Suppose a_n is a nonnegative sequence of real numbers. Show that $\sum a_n$ either converges or diverges to infinity. Equivalently, show that if there exists $M>0$ so that $\sum_{k=0}^n a_k \leq M$ for all n, then the series $\sum a_n$ converges.

Exercise 3.2. Prove the limit ratio test, Corollary 3.8. This includes showing that if $L=1$, then the test is inconclusive. To do this, find two series that yield the limit $L=1$, one of which is convergent and one of which is divergent.

Exercise 3.3. Prove the root test and limit root test, Theorem 3.9.

Exercise 3.4. Let $r(n)$ and $\rho(n)$ be the quantities from the ratio test and root test, respectively. Show that if $r(n)\to L$ as $n\to\infty$, then $\rho(n)\to L$ as $n\to\infty$.

Exercise 3.5. Determine convergence or divergence for the series whose general term is $\frac{n!(2n)!}{(3n)!}$.

Exercise 3.6. Prove Theorem 3.12.

Exercise 3.7. Prove the 'essential facts' about $\limsup$s stated on page 44.

Exercise 3.8. Determine the intervals of convergence for the following power series:

Exercise 3.9. Prove power series expansion for... sine? Use Lagrange remainder.

Exercise 3.10. Show that a series is 'uniformly Cauchy' if and only if it is uniformly convergent. Here we say that a series of functions $\sum_n f_n$ is uniformly Cauchy if for any $\varepsilon > 0$, there exists N so that

$$n, m \geq N \implies \left|\sum_{k=m+1}^{n} f_k(x)\right| < \varepsilon \quad \text{for all} \quad x.$$

Exercise 3.11. Show that the series

$$S(x) = \sum_{k=1}^{\infty} \frac{x + x^3(k - k^2)}{(1 + k^2x^2)(1 + (k-1)^2x^2)}$$

is a continuous, non-uniform limit of continuous partial sums.

Exercise 3.12. Show that the functions

$$f_n(x) = \tfrac{3nx^2 + n^2x^4}{1 + 2nx^2 + n^2x^4}$$

converge on the interval $[-1, 1]$, but not uniformly.

Exercise 3.13. Show that if a series of functions converges uniformly on $[0, \alpha)$ and converges at $x = \alpha$, then the series converges uniformly on $[0, \alpha]$.

Exercise 3.14. Suppose a power series has infinite radius of convergence. Show that for any $\alpha > 0$, the series converges uniformly on $[-\alpha, \alpha]$.

Exercise 3.15. Prove Theorem 3.24.

Exercise 3.16. Suppose $a = x_0 < x_1 < \cdots < x_N = b$ is a partition of $[a, b]$ and g is a function satisfying

$$g(x) = c_i \quad \text{for} \quad x \in (x_{i-1}, x_i).$$

Show that g is Riemann integrable, with

$$\int_a^b g(x)\, dx = \sum_{i=1}^{N} c_i(x_i - x_{i-1}).$$

Exercise 3.17 (Improper integral on an unbounded interval)**.** Show that

$$\int_1^{\infty} \tfrac{1}{x^2}\, dx = 1$$

in the following sense: For any $\varepsilon > 0$, there exists $R_0 > 0$ so that for any $R > R_0$,

$$\left|\int_1^R \tfrac{1}{x^2}\, dx - 1\right| < \varepsilon.$$

Exercise 3.18. Prove the claims made in Example 3.9.

Exercise 3.19. Prove the following claim or find a counterexample (but hopefully not both): *If f is continuous on (a, b) and bounded on (a, b), then f is uniformly continuous on (a, b).*

Exercise 3.20. Show that if f is continuous on $[a, b]$ and uniformly continuous on $[a, c]$ and $[c, b]$ for some $c \in (a, b)$, then f is uniformly continuous on $[a, b]$. Describe how to extend this to any division of $[a, b]$ into finitely many subintervals.

Exercise 3.21. Prove Theorem 3.29.

4. Foundations

In this section, our main goal will be to describe a rigorous 'construction' of the real number system. The basic idea is to build the reals from simpler pieces. We start with the natural numbers (or 'counting numbers') and integers. We then move on to the rational numbers (given by ratios of integers), before constructing the real numbers essentially as limits of sequences of rational numbers.

4.1. **Natural numbers.**

The starting point is to posit the existence of the set of *natural numbers*, denoted $\mathbb{N}$. This corresponds to our usual notion of counting numbers, so that

$$\mathbb{N} = \{0, 1, 2, 3, \dots\}.$$

Now, it is important to stop at this moment and agree that what I just wrote down is not a proper definition. For example, what do those dots mean? To be more precise, we will follow an approach due to Peano, which appeared in the 1880s. In particular, we will list several *axioms* that will completely characterize the natural numbers, and then take as our central *assumption* that such a set exists. By the way, nobody says you have to accept this assumption. In your heart, you may very well be a 'finitist'. However, rejecting the existence of the natural numbers is going to make it somewhat different to participate in most of modern mathematics.

Let's get to it. We will first list all of the axioms, and we will then briefly describe the role played by each.

- *Axiom 1.* There exists a natural number called 0.
- *Axiom 2.* If $n \in \mathbb{N}$, then n has a *successor* denoted $S(n) \in \mathbb{N}$.
- *Axiom 3.* $0 \neq S(n)$ for any $n \in \mathbb{N}$.
- *Axiom 4.* If $n, m \in \mathbb{N}$ and $S(n) = S(m)$, then $n = m$.
- *Axiom 5.* Let $P(n)$ be a statement about $n \in \mathbb{N}$. If $P(0)$ is true, and whenever $P(n)$ holds, we have that $P(S(n))$ holds as well, then $P(n)$ is true for all $n \in \mathbb{N}$.

Axiom 1 gets us started, while Axiom 2 lets us 'keep going'. In particular, we will define $1 = S(0)$, $2 = S(1)$, $3 = S(2)$, and so on. Axiom 3 prevents the natural numbers from 'wrapping around', while Axiom 4 prevents the natural numbers from 'settling' at some finite value. Finally, Axiom 5 (which is technically not an axiom but an 'axiom schema') is the important 'principle of mathematical induction'. The idea here is that the assumptions guarantee that $P(0)$, $P(1)$, $P(2)$, and so on, must all hold. This axiom also prevents the natural numbers from containing any 'extraneous' elements. For example, the set $\{0, .5, 1, 1.5, 2, \dots\}$ satisfies axioms 1–4 with the usual notion of 'successor' (namely, $S(n) = n + 1$), but not Axiom 5 (consider the statement $P(n) =$'n is not a half-integer'). An analogy to keep in mind is that of climbing a ladder: if you can get on the lowest rung of the ladder, and you can always climb one rung higher, then you can climb all the way up the ladder.

Our crucial assumption is thus:

Assumption: There exists a number system $\mathbb{N}$ satisfying Axioms 1–5.

One very convenient aspect of the natural numbers is the notion of an 'inductive definition'. For example, given $m \in \mathbb{N}$, here's how we define $m + n$ for $n \in \mathbb{N}$:

(i) We define $m + 0 = m$.

(ii) Given a definition of $m+n$, we define $m+S(n)=S(m+n)$.

Now consider the statement $P(n) =$'$m+n$ is defined'. Then (i) says that $P(0)$ is true, and (ii) says that if $P(n)$ is true, so is $P(S(n))$. So by induction, $P(n)$ is defined for all $n \in \mathbb{N}$. (By the way, this is also how we actually define the symbol $\sum_{k=0}^{n} a_k$ for a collection of real numbers $\{a_k\}$. In particular,

$$\sum_{k=0}^{0} a_k := a_0 \quad \text{and} \quad \sum_{k=0}^{n+1} a_k := \sum_{k=0}^{n} a_k + a_{n+1}.$$

Do you see why this is a proper definition?)

Using this definition of addition, we can *prove* various 'obvious things' like

$$n+0=n \quad \text{and} \quad n+S(m)=S(n+m).$$

Try it—it's fun! In particular, these facts together show us that

$$S(n)=n+1,$$

and so we can get rid of this clunky successor notation.

Other important, seemingly obvious facts include commutativity and associativity, i.e.

$$n+m=m+n \quad \text{and} \quad (\ell+m)+n=\ell+(m+n).$$

We also have the important 'cancellation law': for $a,b,c \in \mathbb{N}$,

$$a+b=a+c \implies b=c. \tag{4.1}$$

Let's prove this one.

Proof of (4.1). We use induction on a. If $0+b=0+c$, then $b=c$. Now, suppose $a+b=a+c$ implies $b=c$ for some $a \in \mathbb{N}$ and for all $b,c \in \mathbb{N}$. Then suppose that $S(a)+b=S(a)+c$ for some $b,c \in \mathbb{N}$. Then we have

$$a+S(b)=S(a)+b=S(a)+c=a+S(c),$$

so that $S(b)=S(c)$ by the inductive hypothesis. We thus obtain $b=c$. □

Our next step is to introduce the notion of positivity and order. We say $n \in \mathbb{N}$ is *positive* if $n \neq 0$. We then say that $n \geq m$ if $n=m+a$ for some $a \in \mathbb{N}$. We have various basic properties of order, such as reflexivity ($a \geq a$), transitivity ($a \geq b$ and $b \geq c$ implies $a \geq c$), and anti-symmetry ($a \geq b$ and $b \geq a$ implies $a=b$). We also have that $a \geq b$ if and only if $a+c \geq b+c$; $a<b$ if and only if $S(n) \leq b$; and $a<b$ if and only if $b=a+d$ for some positive $d \in \mathbb{N}$. Finally, we have the trichotomy, that for any $a,b \in \mathbb{N}$ we have exactly one of the following: $a<b$, $a>b$, or $a=b$.

Multiplication is defined inductively as well. That is, $0 \times m = 0$ and $S(n) \times m = (n \times m)+m$. We again get all of the usual properties, like commutativity, distributivity, associativity, the fact that $n \times m = 0$ guarantees $m=0$ and $n=0$, the fact that $a \times c = b \times c$ implies $a=b$ (provided $c \neq 0$), and so on. We may also use the standard notation ab for $a \times b$. We also have that multiplication preserves order, i.e. if $a<b$ and c is positive, then $ac<bc$.

4.2. **Integers.** Our next step is to introduce the integers, denoted $\mathbb{Z}$. An integer is actually going to be identified with an *equivalence class* of pairs of natural numbers. In particular, we define the equivalence relation

$$(a,b)\sim(c,d) \quad \text{iff} \quad a+d=b+c.$$

Secretly, we are thinking that (a,b) corresponds the integer $a-b$, so that the usual natural numbers just correspond to pairs of the form $(a,0)$ and the negative integers correspond to pairs of the form $(0,a)$ for $a\neq 0$. There are, of course, many ways to express the same integer, e.g. $(3,5)\sim(2,4)\sim(1,3)\sim(0,2)$ (and all of these ultimately correspond to the number -2).

To make this precise, one has to define addition and multiplication and check that everything is 'well-defined' (i.e. does not depend on the particular representative of the equivalence class). The definitions are

$$(a,b)+(c,d):=(a+c,b+d) \quad \text{and} \quad (a,b)\times(c,d)=(ac+bd,ad+bc).$$

These are precisely the formulas you end up with if you assume that we already know (a,b) corresponds to the integer $a-b$, but the point is that you don't know this yet. So you need to check things like

$$(a,b)\sim(a',b') \quad \text{and} \quad (c,d)\sim(c',d') \implies (a,b)+(c,d)\sim(a',b')+(c',d').$$

The negation of the integer (a,b) is the integer (b,a). We write $(b,a)=-(a,b)$. If you know that (a,b) corresponds to $a-b$, then this is very reasonable. It follows that any integer is either zero, equal to a positive integer n (that is, it is a member of the equivalence class of $(n,0)$ for some positive n), or equal to the negation of a positive integer n. In the final case, we say we have a *negative* integer.

With this 'trichotomy' in place, we go back to using a single letter to denote an integer, say x. Then we know that $x=0$, $x=n$ for some $n\in\mathbb{N}$, or $x=-n$ for some $n\in\mathbb{N}$. We can then derive all of the basic algebraic laws for integers, like

$$x+(-x)=0 \quad \text{and} \quad x1=x,$$

along with commutativity, distributivity, associativity, and so on. We define *subtraction* by $x-y=x+(-y)$; on the right-hand side, $(-y)$ represents the negation of the integer y.

The notion of order carries over nicely to the set of integers. That is, we have that $n\geq m$ if $n=m+a$ for some $a\in\mathbb{N}$, but now n and m may be any integer. All of the usual properties of order still hold; we won't list them here.

4.3. **The rationals.** From the integers we can construct the rationals, denoted $\mathbb{Q}$. As above, a rational number is going to be identified with an *equivalence class* of pairs of integers (more precisely, an element of $\mathbb{Z}\times\mathbb{Z}\backslash\{0\}$), using the equivalence relation

$$[a,b]\approx[c,d] \quad \text{iff} \quad ad=bc.$$

Secretly, we think of $[a,b]$ representing the rational number $\frac{a}{b}$. Then the fact that $[1,2]\approx[2,4]\approx[3,6]$ (and so on) should be obvious. We define sums, products, and negation by viewing $[a,b]$ as $\frac{a}{b}$ and doing our usual rules of algebra; the result is the following:

$$[a,b]+[c,d]=[ad+bc,bd], \quad [a,b]*[c,d]=[ac,bd], \quad -[a,b]=[-a,b].$$

One can check that these operations are well-defined (i.e. independent of the representative of equivalence class). Now, just as we had to find the natural numbers inside the integers, we need to find the integers inside the rational numbers. In this case, we identify the rational $[a,1]$ with the integer a. (If you remember that we are thinking of $[a,b]$ as $\frac{a}{b}$, then this is reasonable).

The rational numbers have one additional operation, namely, that of *reciprocals.* In particular, if $x=[a,b]$ with $a\neq 0$, then we set $x^{-1}=[b,a]$. To see that this is well-defined, one needs to check that if $[a,b]\approx[c,d]$, then $[b,a]\approx[d,c]$.

Using the definitions above, we can define all of the usual algebraic rules for adding/multiplying rational numbers. We also obtain that $xx^{-1}=1$ for $x\in\mathbb{Q}\backslash\{0\}$. We can also define the *quotient* of two rational numbers, namely

$$\tfrac{x}{y}=x\times y^{-1}\quad\text{for}\quad x\in\mathbb{Q},\quad y\in\mathbb{Q}\backslash\{0\}.$$

We now extend the notion of positivity and order to the rationals. In particular, x is positive if $x=[a,b]$ for some $a,b>0$, and x is negative if $x=-y$ for some positive y. It follows that every rational is either positive, negative, or equals zero. We then say $x>y$ if $x-y$ is positive, and so on. This leads to all of the usual properties of order (like transitivity, preservation of order under addition or positive multiplication, and so on).

Finally, we can define the absolute value $|x|$ of a rational $x\in\mathbb{Q}$ and the distance $|x-y|$ between two rationals in the usual ways.

4.4. **The real numbers.** Finally, we can 'construct' the real numbers, denoted $\mathbb{R}$. There are two standard approaches. One (using equivalence classes of Cauchy sequences of rational numbers) is due to Cantor; the other (using subsets of the rationals known as 'cuts') is due to Dedekind. We will follow Cantor's approach.

We first recall that a *sequence* of rational numbers is a function mapping $\mathbb{N}$ to $\mathbb{Q}$. We use the notation $\{a_n\}$. We say that a sequence is *Cauchy* if for any rational $\varepsilon>0$, there exists $N\in\mathbb{N}$ so that

$$n,m\geq N\implies |a_n-a_m|<\varepsilon.$$

This is the same definition we used before, except we restrict to *rational* $\varepsilon>0$ (since we don't know what general real numbers are yet).

We now define an equivalence relation on the class of sequences of rational numbers as follows:

$$\{a_n\}\simeq\{b_n\}\quad\text{if}\quad\lim_{n\to\infty}|a_n-b_n|=0.$$

Here we use the usual notion of limit, but again with rational ε. That is, the above means that for any rational $\varepsilon>0$, there exists $N\in\mathbb{N}$ so that

$$n\geq N\implies |a_n-b_n|<\varepsilon.$$

You should check that the relation above does in fact define an equivalence relation.

We now define a real number to be an *equivalence class* of Cauchy sequences of rational numbers.

Example 4.1. Go way back to Example 1.4, where we constructed a sequence of rationals of the form $x_n=\frac{a_n}{b_n}$. This sequence is a Cauchy sequence, but did not converge to any rational number. Indeed, we showed that any putative limit L would have to satisfy $L^2=2$, which has no rational solution. The equivalence class of $\{x_n\}$ is the real number that we would then call $\sqrt{2}$.

Since sums and products of Cauchy sequences are again Cauchy sequences (and equivalence of Cauchy sequences is preserved under these operations), we can make sense of addition and multiplication of real numbers. We can also find the rationals (and hence the integers and natural numbers) 'hiding' inside this definition of reals by simply identifying a rational number q with equivalence class of the Cauchy sequence $\{q, q, q, \dots\}$. We can define the negation of a real number by prescribing $-\{x_n\} = \{-x_n\}$. In this way, we find that all of the usual algebraic laws carry over to the real numbers. The *reciprocal* operation is a bit more subtle, but the crucial fact is this: if x is a nonzero real number, then it has a representative sequence a_n that is bounded away from zero. In this case, we can show that a_n^{-1} is again Cauchy, and we can take x^{-1} to be the real number represented by the sequence a_n^{-1}.

We incorporate positivity and order into the reals as follows. We say x is positive if it has a representative sequence that is bounded below by a positive rational c. Negative reals are the negations of positive reals. We get the usual trichotomy: every real is either positive, negative, or zero.

At this point, we can prove some interesting things about real numbers. For example, we have the *Archimedean property*:

Lemma 4.1 (Archimedean property)**.** *For any $x > 0$ and $\varepsilon > 0$, there exists $M \in \mathbb{N}$ so that $M\varepsilon > x$.*

The name for this property comes from the fact that this was used in the 'exhaustion' type arguments of Archimedes (presented at the very beginning of these notes).

To establish Lemma 4.1, we will rely on one additional lemma about the natural numbers.

Lemma 4.2. *For any integers $A \geq 1$ and $B \geq 1$, there exists $n \in \mathbb{N}$ such that $An > B$.*

Proof. We will use induction on B. First suppose $B = 0$ and $A \geq 1$. Then choosing $n = 1$, we obtain $An > B$. Next, suppose that for all $A \geq 1$, there exists n such that $An > B$. Now choose any $A \geq 1$, and choose n such that $An > B$. Then $A(n+1) = An + A \geq An + 1 > B + 1$. This completes the inductive step and hence the proof of the lemma. □

Proof of Lemma 4.2. Fix $x > 0$ and $\varepsilon > 0$, and suppose towards a contradiction that for any $M \in \mathbb{N}$, we have $x \geq M\varepsilon$. This means that for any $M \in \mathbb{N}$, we may find a Cauchy sequence q_n^M representing x so that $q_n^M \geq M\varepsilon$ for all n. By extracting the tail of the sequence, we may also assume that

$$|q_n^M - q_\ell^M| < \tfrac{1}{M} \quad \text{for all} \quad n, \ell.$$

Now consider the 'diagonal' sequence $\{q_M^M\}$. We will show that this is a Cauchy sequence of rationals. Indeed, for any M, N, n, we may write

$$\begin{aligned} |q_M^M - q_N^N| &= |q_M^M - q_n^M + q_n^M - q_n^N + q_n^N - q_N^N| \\ &\leq \tfrac{1}{M} + |q_n^M - q_n^N| + \tfrac{1}{N}. \end{aligned}$$

Now given a rational $\eta > 0$, we may choose N_0 so that $\frac{1}{N_0} < \eta$ (by the previous lemma). On the other hand, for fixed $N, M \geq N_0$, we have that $\{q_n^M\}_n$ and $\{q_n^N\}_n$ are equivalent Cauchy sequences (since they both represent x). In particular, $\lim_{n\to\infty}[q_n^N - q_n^M] = 0$, and hence we may choose n sufficiently large so that

$$|q_n^M - q_n^N| < \eta.$$

Altogether, we found N_0 so that $N, M \geq N_0$ guarantees

$$|q_M^M - q_N^N| < 3\eta,$$

showing that $\{q_M^M\}$ is Cauchy.

Now, using the fact that q_M^M is Cauchy, we may find N_0 so that

$$q_M^M \leq q_{N_0}^{N_0} + 1 \quad \text{for all} \quad M \geq N_0.$$

On the other hand, we have $q_M^M \geq M\varepsilon$, and hence we deduce

$$M\varepsilon \leq q_{N_0}^{N_0} + 1 \quad \text{for all} \quad M \geq N_0.$$

Writing $q_{N_0}^{N_0} = p/r$ for some positive integers p, r, this implies

$$Mr\varepsilon \leq p + r \quad \text{for all} \quad M \geq N_0.$$

As $Mr\varepsilon > 0$, this implies (choosing a representative sequence for ε) that we may find a rational number a/b (with $a, b > 0$) so that

$$Mra \leq b(p + r) \quad \text{for all} \quad M \geq N_0.$$

However, applying Lemma 4.2 (with $A = ra$ and $B = b(p + r)$, we may find M (which, without loss of generality, can be taken greater than N_0) so that

$$Mra > b(p + r),$$

which yields a contradiction. □

Using the Archimedean property, we also derive the important fact that:

Lemma 4.3 (Density of $\mathbb{Q}$)**.** *The rational numbers are* dense *in* $\mathbb{R}$.

This lemma means that for any $x < y$, there exists $q \in \mathbb{Q}$ so that $x < q < y$. (The irrational numbers, i.e. $\mathbb{R}\backslash\mathbb{Q}$, are dense in $\mathbb{R}$ as well.)

Proof. Let us just consider the case $0 < x < y$. By the Archimedean principle, we may choose $b \in \mathbb{N}$ so that $b(y - x) > 1$. Then $by - bx > 1$, and hence there must exist $a \in \mathbb{N}$ so that $bx < a < by$ (see below). In particular, $x < \frac{a}{b} < y$.

To argue for the existence of a, note that $\{n \in \mathbb{N} : n \geq by\}$ is nonempty (this is a consequence of Lemma 4.2). Then by the well-ordering principle (see Exercise 4.1), there is a smallest integer n_* such that $n_* \geq by$. Setting $a = n_* - 1$, we get $bx < a < by$, as desired. □

Finally, let us discuss the crucial property of $\mathbb{R}$ that we have been after for so long, namely, that of *completeness.*

Proposition 4.4. *The real numbers,* $\mathbb{R}$*, are complete.*

Proof. Let $\{x_n\}$ be a Cauchy sequence of real numbers. For each n, let $\{y_m^n\}$ be a representative Cauchy sequence of rationals for x_n. By extracting the tail of each sequence, we may assume that

$$|y_\ell^n - y_m^n| < \tfrac{1}{n} \quad \text{for all} \quad m, \ell \geq 1.$$

Now let us define the 'diagonal' sequence

$$z_n = y_n^n.$$

Let's show that $\{z_n\}$ is a Cauchy sequence: Let $\varepsilon > 0$ and (using the fact that $\{x_n\}$ is Cauchy) choose N so that

$$n, m \geq N \implies |x_n - x_m| < \varepsilon.$$

Since $x_n - x_m$ is represented by $y_\ell^n - y_\ell^m$, we may find an ℓ so that

$$|y_\ell^n - y_\ell^m| < \varepsilon.$$

Choosing N even larger so that $N\varepsilon > 1$ (which is possible by the Archimedean property), we find that for $m, n \geq N$ we have

$$\begin{aligned} |z_n - z_m| &\leq |y_n^n - y_\ell^n| + |y_\ell^n - y_\ell^m| + |y_\ell^m - y_m^m| \\ &< \tfrac{1}{n} + \varepsilon + \tfrac{1}{m} < 3\varepsilon. \end{aligned}$$

Now, let z be the real number represented by the sequence $\{z_n\}$.

We will show that $\lim_{n\to\infty} x_n = z$. Fix $\varepsilon > 0$ and let us choose N large enough that

$$N\varepsilon > 1 \quad \text{and} \quad n, m \geq N \implies |x_n - x_m| < \varepsilon.$$

Then, as before, we may find ℓ so that

$$|y_\ell^n - y_\ell^m| < \varepsilon.$$

Now fix $n \geq N$. To show that $|x_n - z| < \varepsilon$, it suffices to show that $x_n - z$ is represented by a sequence of rationals q_m satisfying $|q_m| < \varepsilon$ for all m. In fact, $x_n - z$ is represented by

$$y_m^n - z_m = y_m^n - y_m^m,$$

and we have

$$\begin{aligned} |y_m^n - y_m^m| &\leq |y_m^n - y_\ell^n| + |y_\ell^n - y_\ell^m| + |y_\ell^m - y_m^m| \\ &< \tfrac{1}{n} + \varepsilon + \tfrac{1}{m} < 3\varepsilon \end{aligned}$$

for all m. The result follows. □

As we saw above, once we have the completeness property of real numbers, other essential properties of real numbers (like existence of suprema, and the 'nested interval property') follow.

We can quickly fill in one gap that remains, namely, the question of real exponentiation. In particular, for $x > 0$ and $\alpha \in \mathbb{R}$, the quantity $x^\alpha = \lim_{n\to\infty} x^{q_n}$, where $\{q_n\}$ is any sequence of rationals converging to α. One must, of course, check that this is actually well-defined.

4.5. **Cardinality.** The last topic we will cover in this section is that of *cardinality*, which refers to the *size* of a set. To begin, we define what it means for two sets to have equal cardinality:

Definition 4.1. Two sets X and Y have *equal cardinality* if there exists a bijection mapping X to Y.

Here a *bijection* refers to a mapping that is both *injective* and *surjective.* Injective means that

$$f(x) = f(y) \implies x = y,$$

while surjective means that

$$\text{for any} \quad y \in Y, \quad \text{there exists } x \in X \quad \text{so that} \quad f(x) = y.$$

Sometimes we say 'one-to-one' instead of injective and 'onto' instead of surjective.

We will see below that it is possible for two sets to have the same cardinality, even though one is a strict subset of the other.

The emptyset $\emptyset$ is said to have cardinality zero. For any $n \in \mathbb{N}\backslash\{0\}$, we say that a set has cardinality n if it has equal cardinality with the set

$$\{j \in \mathbb{N} : 1 \leq j \leq n\}.$$

You should convince yourself that if X has cardinality n, then it cannot have cardinality m for any $m \in \mathbb{N}\backslash\{n\}$.

We say that a set is *finite* if it has cardinality n for some $n \in \mathbb{N}$. Otherwise, the set is *infinite.*

Example 4.2. The set of natural numbers, $\mathbb{N}$, is infinite.

Proof. We can show that for any $n \in \mathbb{N}$, any function

$$f : \{j \in \mathbb{N} : 1 \leq j \leq n\} \to \mathbb{N}$$

is necessarily *bounded*; that is, there exists $M \in \mathbb{N}$ so that $f(j) \leq M$ for all $1 \leq j \leq n$. It follows that for any $n \in \mathbb{N}$, there is no surjective function from $\{j \in \mathbb{N} : 1 \leq j \leq n\}$ to $\mathbb{N}$. □

The next example shows that it is possible for two sets to have equal cardinality, even though one is a strict subset of the other. (You should convince yourself that this is not possible if either has finite cardinality.)

Example 4.3. The natural numbers $\mathbb{N}$ and the integers $\mathbb{Z}$ have equal cardinality.

Proof. We define $f : \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} \frac{n+1}{2} & n \text{ odd} \\ -\frac{n}{2} & n \text{ even.} \end{cases}$$

I will leave it to you to check that this is a bijection. □

We say a set is *countable* (or *countably infinite*) if it has equal cardinality with the natural numbers. If a set is infinite but not countably infinite, we say it is *uncountable.*

Example 4.4. The rationals, $\mathbb{Q}$, are countable.

Proof. It suffices to show that the rationals in $(0, 1)$ are countable. To this end, let us describe how to construct a *surjective* map f from $\mathbb{N}$ to $(0, 1) \cap \mathbb{Q}$. We first define $f(0) = \frac{1}{2}$. Then, supposing we have defined $f(n)$ via $f(n) = \frac{p}{q}$, we define $f(n + 1)$ according to the following rule:

- If $p < q - 1$, we set $f(n + 1) = \frac{p+1}{q}$.
- If $p = q - 1$, we set $f(n + 1) = \frac{1}{q+1}$.

Thus we have

$$f(0) = \tfrac{1}{2},\ f(1) = \tfrac{1}{3},\ f(2) = \tfrac{2}{3},\ f(3) = \tfrac{1}{4},\ f(4) = \tfrac{2}{4},\ f(5) = \tfrac{3}{4},\ f(6) = \tfrac{1}{5},$$

and so on. In particular, every rational of the form $\frac{p}{q}$ with $p, q \geq 0$ and $1 \leq p < q$ is in the range of this function. This accounts for every rational in $(0, 1)$. □

Example 4.5. The real numbers, $\mathbb{R}$, are uncountable.

Proof. It is enough to show that the real numbers in $(0, 1)$ are uncountable. We will give the classic 'diagonal' argument of Cantor. This uses the notion of 'decimal expansion' of real numbers, which involves writing a real number $x \in (0, 1)$ as

$$x = .7364826482364872\ldots,$$

with the digits possibly 'going on forever'. We can actually view this as giving a Cauchy sequence of rationals representing x of the form

$$q_n = \sum_{k=1}^{n} \frac{a_k}{10^k},$$

where $a_k \in \{0, \ldots, 9\}$. (In the example above, we have $\{a_k\} = \{7, 3, 6, 4, 8, 2, \ldots\}$.) Note that the decimal expansion is not unique because of examples like

$$.1 = .09999999\ldots$$

However, if we restrict to the use of *infinite* decimal expansions, we can restore uniqueness. We will follow this convention in what follows.

What we can show is that if we have any (countable) list of real numbers in $(0, 1)$, then we can produce a real number in $(0, 1)$ that is not contained in the list. This will show that there is no surjective map from $\mathbb{N}$ to $(0, 1) \cap \mathbb{R}$, which implies uncountability.

So, we suppose we have any list of real numbers in $(0, 1)$. For example, suppose our list begins

$$\begin{aligned} x_1 &= .2764578236834\ldots, \\ x_2 &= .3487934758934\ldots, \\ x_3 &= .3948753948573\ldots, \end{aligned}$$

and so on. We will define a new real number z by prescribing its decimal expansion. We choose the first coefficient to be a nonzero number different than the first coefficient of x_1, the second coefficient to be a nonzero number different than the second coefficient of x_2, and so on. Thus our number might begin with

$$z = .355\ldots$$

Continuing in this fashion, we choose the n^{th} coefficient to be different than the n^{th} coefficient of x_n. In this way, we can construct a real number that is guaranteed not to be on our original list! □

The examples above show that while the natural numbers and the real numbers are both infinite, they are *not* the 'same size' (that is, they do not have the same cardinality). This suggests the question: are there any cardinalities between that of $\mathbb{N}$ and that of $\mathbb{R}$? This question, which came to be known as the *continuum hypothesis*, was investigated thoroughly by Cantor, who was never able to resolve the problem. In fact, it was not until work of Gödel (1940) and Cohen (1963) that the question was settled. Strangely enough, the answer is neither 'yes' or 'no'. Instead, it turns out that either possibility is consistent with the other axioms of set theory. That is, this issue is completely 'independent' from the rest of axiomatic set theory!

4.6. **Exercises.**

Exercise 4.1. Show that if S is a non-empty subset of the natural numbers, then S has a minimal element.

5. Metric space topology

In this section, we are going to present some of the content encountered above from a slightly more general viewpoint, namely, that of 'metric spaces' and 'metric space topology'. As we will see, the notion of 'continuity' can be understood in significantly more general and abstract settings than the case of real-valued functions of one variable. Indeed, this is one instance in which a little bit of abstraction pays off quite well.

5.1. **Metric spaces.**

Informally, a metric space is any non-empty set with a notion of a 'distance' between two points. To be considered a 'distance', we need certain properties to hold. The precise definition is the following:

Definition 5.1 (Metric space)**.** A metric space (X, d) is a non-empty set X together with a function $d : X \times X \to [0, \infty)$ satisfying the following:

1. $d(x, y) = d(y, x)$ for $x, y \in X$.
2. $d(x, y) = 0$ if and only if $x = y$.
3. For any $x, y, z \in X$, we have
$$d(x, z) \leq d(x, y) + d(y, z).$$

The first condition is a symmetry condition; the second condition guarantees that distinct points are a positive distance apart; and the third condition is known as the 'triangle inequality'. The triangle inequality says that you should never be able to find a strictly more efficient route by adding an extra stop to your itinerary.

Example 5.1. The real numbers, $\mathbb{R}$, with the distance $d(x, y) = |x-y|$, is a metric space. However, this is not the only 'distance' we could assign to numbers. Another example is the 'discrete metric'

$$\tilde{d}(x, y) = \begin{cases} 1 & x \neq y \\ 0. & x = y \end{cases}$$

You should check that $\tilde{d}$ satisfies the definition of metric as well!

This 'metric space' viewpoint will also allow us to study higher dimensional Euclidean space.

Example 5.2. We define $\mathbb{R}^n$ to be the set of all ordered 'n-tuples' of the form

$$x = (x_1, \dots, x_n), \quad \text{where each} \quad x_i \in \mathbb{R}.$$

The standard 'Euclidean' distance is given by

$$d(x, y) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2}.$$

Can you verify that the triangle inequality holds?

This notion of distance is intimately linked to our notion of the 'size' or 'length' (or *norm*) of an element of $\mathbb{R}^n$. In particular, we write

$$|x| = d(x, 0) = \sqrt{x_1^2 + \cdots + x_n^2}.$$

This agrees with the usual notion of 'absolute value' in one dimension; in higher dimensions it is our definition of length.

In fact, the notion of metric space goes far beyond $\mathbb{R}^n$. Let's see a few more examples.

Example 5.3 (Function spaces)**.**

(i) Consider the set of continuous real-valued functions on $[0,1]$. Define the distance
$$d(f,g) = \sup_{x\in[0,1]} |f(x)-g(x)|.$$
This gives a metric space (known as $C([0,1])$), and we may call d the *uniform* metric. What is the meaning of the 'norm' of a function f?

(ii) Again consider the set of continuous, real-valued functions on $[0,1]$. This time, define the distance
$$\tilde{d}(f,g) = \int_0^1 |f(x)-g(x)|\,dx.$$
This also forms a metric space. Now what is the meaning of the 'norm' of a function?

We first define what *convergence* means in the setting of a metric space X. We first work in the setting of sequences. Here a 'sequence' is just a function mapping $\mathbb{N}$ to X, which we denote by x_n (rather than $x(n)$, say).

Definition 5.2 (Convergent sequence)**.** Let $\{x_n\}$ be a sequence in a metric space (X,d). We say that x_n *converges to* $x\in X$ if
$$\text{for all}\quad \varepsilon>0 \quad \text{there exists} \quad N \quad \text{such that} \quad n\geq N \implies d(x_n,x)<\varepsilon.$$

We use all the same notation as before, e.g. $x_n\to x$ as $n\to\infty$, or
$$\lim_{n\to\infty} x_n = x.$$
When $X=\mathbb{R}$ and $d(x,y)=|x-y|$, this is exactly the same definition of convergence that we had before. What would convergence mean if you were using the discrete metric on $\mathbb{R}$? (See Exercise 5.2.)

Example 5.4. Let's return to the metric spaces described in Example 5.3. Now a sequence of 'points' actually refers to a sequence of *functions* f_n, each of which is defined on $[0,1]$. Let's consider the sequence
$$f_n(x) = x^n, \quad x\in[0,1].$$
(i) If we consider the uniform metric, that is,
$$d(f,g) = \sup_{x\in[0,1]} |f(x)-g(x)|,$$
then the sequence does *not* converge. To see this, note that we have
$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} x^n = f(x) := \begin{cases} 0 & 0\leq x<1, \\ 1 & x=1, \end{cases}$$
where here 'lim' is just referring to a limit of a sequence of real numbers. This implies that the limit of the sequence f_n (in terms fo the metric d) would have to be f; however, f does *not* belong to the metric space (since it is not continuous). The conclusion is that the sequence does not converge in the uniform metric. More precisely, we say that *the sequence converges pointwise but not uniformly.* Note that this language is completely consistent with our notions of pointwise and uniform convergence in the preceding sections.

(ii) However, if we consider the second metric, that is,

$$\tilde{d}(f,g) = \int_0^1 |f(x)|\,dx,$$

then the sequence f_n converges to the zero function $f(x) \equiv 0$. Indeed,

$$d(f_n, 0) = \int_0^1 |f_n(x)|\,dx = \int_0^1 x^n\,dx = \tfrac{1}{n+1} \to 0 \quad \text{as} \quad n \to \infty.$$

The notion of a Cauchy sequence also makes perfect sense in a general metric space:

Definition 5.3 (Cauchy sequence)**.** A sequence $\{x_n\}$ in a metric space (X,d) is *Cauchy* if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ so that

$$n, m \geq N \implies d(x_n, x_m) < \varepsilon.$$

It is not hard to show that every convergent sequence is a Cauchy sequence. Whether or not Cauchy sequences converge depends on whether or not the space is 'complete'.

Definition 5.4. A metric space (X,d) is *complete* if every Cauchy sequence converges.

We saw in the preceding sections that $\mathbb{R}$ is complete. This important property also holds for $\mathbb{R}^n$.

Theorem 5.1. *The space $\mathbb{R}^n$ (with the standard Euclidean metric) is complete.*

Proof. Let $\{x_k\}$ be a Cauchy sequence in $\mathbb{R}^n$. Write the components of $x \in \mathbb{R}^n$ as

$$x = (x^1, \ldots, x^n).$$

Since

$$|x^j| \leq \sqrt{[x^1]^2 + \cdots + [x^n]^2} = \|x\| \quad \text{for each} \quad j,$$

we can deduce that each sequence of components $\{x_k^j\}$ is a Cauchy sequence in $\mathbb{R}$. Thus (by completeness of $\mathbb{R}$), for each $j = 1, \ldots, n$, we have a limit z^j. In particular, given $\varepsilon > 0$, we may choose $N_1, \ldots, N_n$ so that

$$k \geq N_j \implies |x_k^j - z^j| < \varepsilon.$$

Defining $z = (z^1, \ldots, z^j)$, we find that for

$$k \geq \max\{N_1, \ldots, N_n\},$$

we have

$$\begin{aligned} |x_k - z| &= \sqrt{|x_k^1 - z^1|^2 + \cdots + |x_k^n - z^n|^2} \\ &< \sqrt{\varepsilon^2 + \cdots + \varepsilon^2} = \sqrt{n}\varepsilon. \end{aligned}$$

The result follows. □

What about the spaces in Example 5.3? That's a harder question that we won't get into here. But I encourage you to think about it!

5.2. **Metric space topology.** We have just seen that the notion of a metric gives rise to the notion of convergence. We can also use the notion of a metric to define a notion of 'open' and 'closed' sets. (There is actually a more general concept hiding in the background here, namely, that of a *topology* or a *topological space.* However, in these notes, we will keep our attention focused on the more straightforward setting of metric spaces).

We have the following two definitions:

Definition 5.5 (Open ball)**.** Let (X, d) be a metric space. Given $x \in X$ and $r > 0$, we define the *(open) ball of radius r centered at x* by

$$B_r(x) = \{y \in X : d(x, y) < r\}.$$

Definition 5.6 (Open set)**.** Let (X, d) be a metric space and let S be a subset of X. We say that S is *open* if

$$\text{for all} \quad x \in S, \quad \text{there exists} \quad r > 0 \quad \text{such that} \quad B_r(x) \subset S.$$

I like to think of openness in terms of 'wiggle room'. That is, a set is open if there is a little bit of 'wiggle room' around each point. Exactly what that looks like depends on the metric itself. By the way, you should stop for a second and make sure that you agree that 'open balls' are actually 'open sets' according to the definition above.

As an aside, the three essential properties of 'open sets' that guarantee that this definition yields a 'topology' are the following:

- The sets $\emptyset$ and X are open.
- If $\{U_\alpha\}_{\alpha \in A}$ is *any* collection of open sets, then the union $\cup_{\alpha \in A} U_\alpha$ is open.
- If $\{U_n\}_{n=1}^N$ is any *finite* collection of open sets, then the intersection $\cap_{n=1}^N U_n$ is open.

You are asked to verify these properties in Exercise 5.5.

Let us also point out that there is a way of phrasing convergence purely in terms of open sets—see Exercise 5.6.

Once we have a notion of 'open sets', we also get a notion of 'closed sets'. But be careful: 'closed' is *not* the same thing as 'not open'.

Definition 5.7. Let (X, d) be a metric space. A set S is *closed* if its *complement* S^c is open. Here

$$S^c := \{x \in X : x \notin S\}.$$

Really, don't conflate 'closed' with 'not open' or vice versa. Remember, $\emptyset$ and X are always both open, but they are complements of one another. So they are always both closed as well.

Let's see a few examples.

Example 5.5. Consider $\mathbb{R}$ with the standard metric. An open ball is an interval of the form $(x - r, x + r)$ for some $x \in \mathbb{R}$ and $r > 0$. Intervals of the form (a, b) are open, but intervals of the form $[a, b)$ or $[a, b]$ are not. Semi-infinite intervals like (a, ∞) are open. Intervals of the form $[a, b]$ are closed, while intervals of the form $[a, b)$ or $(a, b]$ are neither open nor closed. How about an interval of the form $[a, \infty)$? Is it closed?

Example 5.6. Consider $\mathbb{R}$ with the discrete metric. Then for any $x \in \mathbb{R}$,

$$B_r(x) = \begin{cases} \{x\} & \text{if } r \leq 1, \\ \mathbb{R} & \text{if } r > 1. \end{cases}$$

This actually shows that in this topology, every single point $\{x\}$ is open. But then since arbitrary unions of open sets are open, we get that *every* set is open! That means that every set is closed as well.

We often think about closed sets in terms of sequences. In particular, we have the following:

Lemma 5.2. *Let (X, d) be a metric space and $S \subset X$. Then S is closed if and only if whenever $\{x_k\}$ is a sequence of elements in S that converge to x, we have $x \in S$.*

Proof. I will show you the $\Longleftarrow$ direction. You can work out the $\Longrightarrow$ direction yourself!

We argue by contrapositive. We suppose S is not closed. This means that S^c is not open. This in turn implies that there exists $x_* \in S^c$ so that for any n, $B_{1/n}(x_*) \not\subset S^c$. In other words, for any n, there exists $x_n \in S$ with

$$d(x_n, x_*) < \tfrac{1}{n}.$$

But now the sequence $\{x_n\}$ is contained in S but converges to x_*, which does not belong to S. Done! □

Example 5.7. Consider the set of rationals $\mathbb{Q}$ as a subset of $\mathbb{R}$ (with the standard metric). Is this set open, closed, both, or neither?

Well, the set is not closed, because I can find a sequence of rational numbers converging to $\sqrt{2}$, which is irrational. The set is not open, either, because for any $\varepsilon > 0$ and any $q \in \mathbb{Q}$, we can find an irrational x in $(q - \varepsilon, q + \varepsilon)$ (so there is no 'wiggle room').

Example 5.8. Let's go back to the function space example of $C([0, 1])$. Consider the set S of polynomials (restricted to $[0, 1]$). Is this set open, closed, both, or neither? Actually, we will see this set is a lot like the previous example.

Indeed, the set is not closed, because I can find a sequence of polynomials P_n converging uniformly to a continuous function that is *not* a polynomial. Indeed, we can just consider the Taylor polynomial approximations to some function like $\sin x$.

The set is not open, either, because for any $\varepsilon > 0$ and any polynomial P, I can find a continuous function f that is *not* a polynomial with $d(P, f) < \varepsilon$. In particular, I just need to take $f(x) = P(x) + \varepsilon \sin(x)$, say.

In fact, just like the rationals in $\mathbb{R}$, the polynomials are *dense* in $C([0, 1])$. This means that for *any* continuous function f on $[0, 1]$, we can find a sequence of polynomials P_n converging uniformly to f on $[0, 1]$. (This is a result due to Weierstrass in 1885.)

5.3. **Continuity.** Now that we have seen some of the basics of topology, we can turn to several important concepts: *continuity, connectedness, and compactness.* These are actually 'purely topological concepts', that they only depend on the notion of an open set. However, we will focus on giving 'sequential' or 'functional' versions of these, which are a little bit more intuitive to work with and end up being equivalent in the metric space setting anyway.

First, we give a definition of continuity in terms of sequences that should look pretty familiar.

Definition 5.8 (Continuity, I)**.** Let X and Y be metric spaces and $f : X \to Y$. We say that f is *continuous at* $x_0 \in X$ if

$$\text{whenever} \quad \lim_{n\to\infty} x_n = x_0, \quad \text{we have} \quad \lim_{n\to\infty} f(x_n) = f(x_0).$$

We say that f is *continuous on* X if f is continuous at every point $x_0 \in X$.

Note that when we write $\lim x_n = x_0$, we are referring to the metric on X, while $\lim f(x_n) = f(x_0)$ involves the metric on Y. When $X = Y = \mathbb{R}$ with the usual metric, this recovers our notion of continuity.

Here's a more interesting example to test your understanding:

Example 5.9. Let X be the metric space of continuous functions on $[0, 1]$ with metric

$$d(f, g) = \sup_{x\in[0,1]} |f(x) - g(x)|.$$

Define the function $J : X \to \mathbb{R}$ by

$$J(f) = \int_0^1 f(x)\, dx.$$

Then J is continuous. Indeed, $f_n \to f$ means the sequence f_n converges *uniformly* to f on $[0, 1]$. Thus for any $\varepsilon > 0$, there exists N so that

$$n \geq N \implies |f_n(x) - f(x)| < \varepsilon \quad \text{for all} \quad x \in [0, 1].$$

In this case, we have

$$|J(f_n) - J(f)| = \left| \int_0^1 f_n(x) - f(x)\, dx \right| \leq \int_0^1 |f_n(x) - f(x)|\, dx < \varepsilon$$

for any $n \geq N$.

An equivalent definition of continuity that involves the metrics a bit more explicitly is the following:

Definition 5.9 (Continuity, II)**.** Let (X, d) and $(Y, \tilde{d})$ be metric spaces and $f : X \to Y$. We say that f is *continuous at* $x_0 \in X$ if

$$\text{for all} \quad \varepsilon > 0 \quad \text{there exists} \quad \delta > 0 \quad \text{so that} \quad x \in B_\delta(x_0) \implies f(x) \in B_\varepsilon(f(x_0)).$$

That is,

$$d(x, x_0) < \delta \implies \tilde{d}(f(x), f(x_0)) < \varepsilon.$$

Note that $B_\delta(x_0)$ is a ball in the (X, d) metric, while $B_\varepsilon(f(x_0))$ is a ball in the $(Y, \tilde{d})$ metric.

At this point, you should stop and convince yourself that Definition 5.8 and Definition 5.9 are equivalent. Otherwise, we would have two competing notions of 'continuity', which would be a serious problem.

To test out your understanding, consider the following example:

Example 5.10. Let $X = Y = \mathbb{R}$. Give X the discrete metric and Y the standard metric. It follows that *every function* $f : X \to Y$ is continuous.

We can also continue from Definition 5.9 to define a notion of *uniform* continuity:

Definition 5.10 (Uniformly Continuous)**.** Let (X, d) and $(Y, \tilde{d})$ be metric spaces and $f : X \to Y$. We say that f is *uniformly continuous* if for all $\varepsilon > 0$, there exists $\delta > 0$ so that for all $x_0 \in X$,

$$x \in B_\delta(x_0) \implies f(x) \in B_\varepsilon(f(x_0)).$$

With these definitions of continuity in place, we can establish some basic results about continuous functions that parallel those we proved above. For example:

Proposition 5.3. *Let X, Y, Z be metric spaces. Suppose $f : X \to Y$ is continuous and $g : Y \to Z$ is continuous. Then $g \circ f$ is continuous.*

Proof. The proof is actually identical to the proof of Proposition 2.6. You just need to replace all instances of $|\cdot - \cdot|$ with $d(\cdot, \cdot)$, where d is the appropriate metric. □

5.4. **Connectedness and compactness.** We next introduce two properties of sets known as *connectedness* and *compactness*. More precisely, we will define 'path-connectness' and 'sequential compactness' (which turn out to be equivalent in the setting of metric spaces).

Definition 5.11 (Path-connected set)**.** Let X be a metric space. A nonempty set $S \subset X$ is *path-connected* if for any $x_0, x_1 \in S$, there exists a continuous function $f : [0, 1] \to S$ so that $f(0) = x_0$ and $f(1) = x_1$. (We call such a function a *path* between x_0 and x_1).

Example 5.11. Suppose we equip $\mathbb{R}$ with the discrete metric d. Then the only connected sets are singletons $\{x\}$. Indeed, the only functions $\gamma : [0, 1] \to (\mathbb{R}, d)$ that are continuous are constant functions. Can you see why?

Example 5.12. Intervals in $\mathbb{R}$ (with the standard metric) are connected. For example, the interval $[a, b]$ is connected because given $x_0, x_1 \in [a, b]$ we can define the path

$$f(\theta) = (1 - \theta)x_0 + \theta x_1 \in [a, b].$$

In fact, this same construction shows that in $\mathbb{R}^n$, any 'n-dimensional interval' of the form

$$I_1 \times I_2 \times \cdots \times I_n$$

(where each I_j is an interval) is connected. So, in $2d$ you get rectangles, in $3d$ you get rectangular prisms, and so on.

More generally, we have the following result about connected subsets of $\mathbb{R}$:

Lemma 5.4. *Suppose $S \subset \mathbb{R}$ is connected and $y_0 < y_1$ are elements of S. Then $[y_0, y_1] \subset S$.*

Proof. By connectedness, there exists a continuous function $\gamma : [0, 1] \to S$ with $\gamma(0) = y_0$ and $\gamma(1) = y_1$. By the Intermediate Value Theorem, γ must attain every value in $[y_0, y_1]$. □

In general, we have the following result concerning continuous functions on connected sets:

Theorem 5.5. *Suppose S is a connected subset of a metric space (X, d) and $f : X \to Y$ is a continuous function, with $(Y, \tilde{d})$ another metric space. Then $f(S)$ is a connected subset of Y.*

Proof. Take any $y_0, y_1 \in f(S)$. By definition, $y_0 = f(x_0)$ and $y_1 = f(x_1)$ for some $x_0, x_1 \in S$. As S is connected, we may find a continuous path $\gamma : [0,1] \to S$ so that $\gamma(0) = x_0$ and $\gamma(1) = x_1$. Then, using Proposition 5.3, we find that

$$f \circ \gamma : [0,1] \to f(S)$$

is a continuous path with

$$f \circ \gamma(0) = f(x_0) = y_0 \quad \text{and} \quad f \circ \gamma(1) = f(x_1) = y_1.$$

Thus $f(S)$ is connected. □

As a consequence of this general result, we can obtain a higher-dimensional version of the intermediate value theorem. Note that this is only new in dimensions $n \geq 2$.

Corollary 5.6 (Intermediate value theorem for $\mathbb{R}^n$)**.** *If $S \subset \mathbb{R}^n$ is connected and $f : S \to \mathbb{R}$ is continuous, then f satisfies the intermediate value property. More precisely, suppose that*

$$f(x_0) = y_0 \quad \textit{and} \quad f(x_1) = y_1 \quad \textit{for some} \quad x_0, x_1 \in S.$$

Then for any c between y_0 and y_1, there exists $x_* \in S$ so that $f(x_*) = c$.

Proof. From the theorem above, the image set $f(S)$ is a connected subset of $\mathbb{R}$. By assumption, this set contains y_0 and y_1. Thus, the lemma above guarantees that the image $f(S)$ contains every c between y_0 and y_1. □

Remark 5.7. Our presentation here is a bit 'backwards'. Usually, connectedness is introduced as a purely topological property. This definition of connectedness is then used to establish something like Lemma 5.4, which is subsequently used to prove the Intermediate Value Theorem (even the one-dimensional version). Instead, we essentially proved the Intermediate Value Theorem 'by hand' in dimension $n = 1$, and then used it to derive the higher dimensional analogues.

The notion of 'compactness' takes more effort to get used to, but it is an extremely important concept. The definition that follows might remind you of a result we discussed previously, namely, the Bolzano–Weierstrass Theorem 3.29.

Definition 5.12 (Compact set)**.** Let X be a metric space. A set $S \subset X$ is *(sequentially) compact* if every sequence $\{x_n\}$ in S has a subsequence $\{x_{n_k}\}$ that converges to a limit in S.

In everyday language, we use 'compact' to mean something like 'able to fit neatly into a small space'. The definition above can be connected to this more intuitive notion by means of the following result concerning compact sets in $\mathbb{R}^n$. This is a version of the 'Heine–Borel' Theorem of the late 1800s.

Theorem 5.8. *Consider $\mathbb{R}^n$ with the standard metric. Then $S \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof. $\Longrightarrow$: Suppose S is (sequentially) compact.

We first show S is closed. We suppose $\{x_k\}$ is a sequence in S converging to x_*. By compactness, we know that x_k has a subsequence that converges to a limit in S. However, this subsequential limit must also be x_*, and so we obtain $x_* \in S$.

We next show S is bounded. If not, then we may find a sequence $\{x_k\}$ in S such that $\|x_k\| > k$. However, this sequence can have no convergent subsequence!

$\Longleftarrow$: Suppose S is closed and bounded, and let $\{x_k\}$ be a sequence in S. Denote the components of x_k by x_k^j for $j = 1, \dots, n$. Because S is bounded, each component sequence $\{x_k^j\}$ is bounded. Applying the Bolzano–Weierstrass Theorem (Theorem 3.29), we can therefore find a subsequence along which x_k^1 converges. We can then take a subsequence of this subsequence so that x_k^2 also converges. Continuing in this fashion, we can find a subsequence of the original sequence along which x_k^j converges for each $j = 1, \dots, n$. It follows that $\{x_k\}$ converges along this subsequence, and because S is closed the limit must belong to S. □

With this theorem in hand, we have a simple description of compact sets in $\mathbb{R}^n$. Unfortunately, this characterization of compact sets is not valid in all settings. In particular, in more general spaces we can have closed, bounded sets that fail to be compact.

We next consider how continuous functions behave on compact sets.

Lemma 5.9. *Suppose $f : X \to Y$ is a continuous function between two metric spaces and X is compact. Then $f(X)$ is compact.*

Proof. Let $\{y_n\}$ be a sequence in $f(X)$. Then each $y_n = f(x_n)$ for some $x_n = X$. By compactness, the sequence $\{x_n\}$ has a convergent subsequence $x_{n_k} \to x_* \in X$. By continuity, $f(x_{n_k}) \to f(x_*) \in f(X)$. □

Using this, lemma, we can prove a more general version of the Extreme Value Theorem.

Theorem 5.10 (Extreme Value Theorem)**.** *Suppose $f : X \to \mathbb{R}$ is a continuous function and X is compact. Then*

(i) *f is bounded, and*
(ii) *f attains its maximal and minimal values.*

Proof. We first show that f is bounded. This means that there exists $M > 0$ so that $|f(x)| \leq M$ for all $x \in X$. We argue by contradiction. If f were not bounded, we could find $\{x_n\} \subset X$ such that $|f(x_n)| > n$ for each n. However, this would give a sequence in $f(X)$ that has no convergent subsequence, contradicting that $f(X)$ is compact.

We can therefore define $M = \sup_{x \in X} f(x)$ and $m = \inf_{x \in X} f(x)$. We will show that there exists x_* so that $f(x_*) = M$; attainment of the value m is similar. Using the definition of supremum, we see that for any $n \geq 1$, the number $M - \frac{1}{n}$ is *not* an upper bound for the set

$$\{f(x) : x \in X\}.$$

Therefore, for each $n \geq 1$, we may find $x_n \in X$ so that

$$M - \tfrac{1}{n} < f(x) \leq M.$$

By compactness of X, there exists a subsequence x_{n_k} converging to some $x_* \in X$. Noting that $f(x_{n_k}) \to M$ by construction and appealing to continuity, we derive that $f(x_*) = M$, as desired. □

Finally, we prove an analogue of Theorem 3.28 (which showed that continuous functions on $[a, b]$ are automatically uniformly continuous).

Theorem 5.11. *Suppose $f : X \to Y$ is a continuous function between two metric spaces and X is compact. Then f is* uniformly *continuous.*

Proof. Suppose f is continuous but not uniformly continuous. Then, carefully negating the definition of uniform continuity, we may find $\varepsilon_0 > 0$ and two sequences $\{x_n\}$ and $\{y_n\}$ in X such that

$$d(x_n, y_n) < \tfrac{1}{n} \quad \text{and} \quad \tilde{d}(f(x_n), f(y_n)) > \varepsilon_0.$$

Now, using compactness of X (twice), it is possible to find a subsequence along which $x_{n_k} \to x_* \in X$ and $y_{n_k} \to y_* \in X$. In fact, we must have $x_* = y_*$, since

$$d(x_*, y_*) \leq d(x_*, x_{n_k}) + d(x_{n_k}, y_{n_k}) + d(y_{n_k}, y_*) \to 0 \quad \text{as} \quad k \to \infty.$$

We now claim that f fails to be continuous at x_*, which yields a contradiction. To see this, first note that by the triangle inequality and by construction, we have

$$\tilde{d}(f(y_{n_k}), f(x_*)) + \tilde{d}(f(x_{n_k}), f(x_*)) \geq \tilde{d}(f(y_{n_k}), f(x_{n_k})) > \varepsilon_0$$

for all k. Using the pigeonhole principle, this implies that there is a subsequence z_k of either y_{n_k} or x_{n_k} along which

$$\tilde{d}(f(z_k), f(x_*)) > \tfrac{1}{2}\varepsilon_0.$$

However, as we are guaranteed that $z_k \to x_*$, this implies that f fails to be continuous at x_*, as desired. □

5.5. **Exercises.**

Exercise 5.1. Fix $1 \leq p < \infty$ and define

$$d(x, y) = \Big(|y_1 - x_1|^p + \cdots + |y_n - x_n|^p\Big)^{\frac{1}{p}} \quad \text{for} \quad x, y \in \mathbb{R}^n.$$

Show that d is a metric.

Exercise 5.2. Let $X = \mathbb{R}$ and $d(x, y)$ be the discrete metric, that is,

$$d(x, y) = \begin{cases} 1 & x \neq y, \\ 0 & x = y. \end{cases}$$

Show that a sequence x_n converges if and only if it is eventually constant (that is, there exists $c \in \mathbb{R}$ and $N \in \mathbb{N}$ so that $n \geq N \implies x_n = c$).

Exercise 5.3. Verify that an 'open ball' in a metric space is actually an open set.

Exercise 5.4. Show that the set of continuous functions on $[0, 1]$ with the metric

$$d(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|$$

is a complete metric space.

Exercise 5.5. Verify the properties of 'open sets' listed after the definition of 'open set'.

Exercise 5.6. . Let $\{x_n\}$ be a sequence in a metric space (X, d). Show that x_n converges to x if and only if for any open set U containing x, there exists $N \in \mathbb{N}$ so that

$$n \geq N \implies x_n \in U.$$

Exercise 5.7. Let (X, d) be a metric space. Show that for any distinct pair of points $x, y \in X$, there exist open sets S, T so that $x \in S$, $y \in T$, and $S \cap T = \emptyset$. (This shows that every metric space is a '*Hausdorff*' space.)

Exercise 5.8. Work out the $\implies$ direction of Lemma 5.2.

Exercise 5.9. Show that the definitions of continuity in Definition 5.8 and Definition 5.9 are equivalent.

Second Semester Content

6. Euclidean n-space and linear transformations

6.1. **Euclidean n-space.**

We recall the definition of Euclidean n-space, namely

$$\mathbb{R}^n = \{(x_1, \dots, x_n) : \text{each } x_j \in \mathbb{R}\}.$$

In the previous section, we considered this space in the context of 'metric spaces'. In this section, we will instead focus on the viewpoint of $\mathbb{R}^n$ as a 'vector space', specifically, an 'inner product space'. As we will see, the inner product structure on $\mathbb{R}^n$ gives rise to the same metric space structure that we considered previously.

The *vector space* structure of $\mathbb{R}^n$ refers to the fact that we can (i) add elements of $\mathbb{R}^n$ and (ii) multiply elements of $\mathbb{R}^n$ by scalars (i.e. real numbers). In particular,

$$\text{if} \quad x = (x_1, \dots, x_n), \quad y = (y_1, \dots, y_n), \quad \text{and} \quad c \in \mathbb{R},$$

then we can define

$$x + y = (x_1 + y_1, \dots, x_n + y_n) \in \mathbb{R}^n \quad \text{and} \quad cx = (cx_1, \dots, cx_n) \in \mathbb{R}^n. \tag{6.1}$$

To be precise, the notation above (writing x, y horizontally) should be used when we are just thinking of x and y as points in $\mathbb{R}^n$, rather than 'vectors'. When we wish to emphasize the vector space structure, we will write elements of $\mathbb{R}^n$ as *column* vectors, as follows:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

As you can see, it is much bulkier to type columns, and so often we will use rows. Nonetheless, at times it will be important to view elements of $\mathbb{R}^n$ as column vectors (e.g. when we introduce linear transformations), so we will need to be careful about this at various points. Note that in general, I will not decorate x with an arrow (i.e. write $\vec{x}$) or write it in bold (i.e. write $\mathbf{x}$), so you will have to pay attention and use context to interpret expressions like cx in (6.1) above.

The usual algebraic rules for manipulating real numbers carry over to algebraic rules for manipulating vectors. For example, $a(x + y) = ax + ay$ whenever $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$, or $x + 0 = x$ for any $x \in \mathbb{R}^n$, where 0 denotes the zero vector $(0, \dots, 0)$ (again, without any special decoration).

There are many other examples of sets (beyond just $\mathbb{R}^n$) that have 'vector space structure'. Important examples include 'function spaces' (which we also encountered when discussing metric spaces), like the space of continuous functions, or the space of differentiable functions, and so on. Any time you are working with a vector space, there are several important concepts that you will encounter, including the notions of *subspace*, *linear combination*, *span*, *linear independence*, *basis*, and *dimension*. We briefly review these concepts here:

Example 6.1 (Linear algebra refresher)**.**

- If V is a vector space and W is a subset of V, then W is called a *subspace* if $0 \in W$, and $av + bw \in W$ whenever $a, b \in \mathbb{R}$ and $v, w \in W$.
- If $\{v_1, \dots, v_k\}$ is a collection of vectors in a vector space V, then a vector of the form

$$v = a_1 v_1 + \cdots + a_k v_k, \quad a_j \in \mathbb{R}$$

is a *linear combination* of $v_1, \dots, v_k$. The set of all linear combinations of $\{v_1, \dots, v_k\}$ is called the *span* of $\{v_1, \dots, v_k\}$. The span of a set of vectors always produces a subspace.

- A set of vectors $\{v_1, \dots, v_k\}$ is *linearly independent* if
$$a_1 v_1 + \cdots a_k v_k = 0 \implies a_1 = a_2 = \cdots = a_k = 0.$$
That is, the *only* linear combination of the vectors v_k that equals zero is the 'trivial' one.
- Suppose W is a subspace of V. A collection of vectors $\{v_1, \dots, v_k\}$ is a *basis* for W if it is linearly independent and its span equals W.
- If a vector space has a basis consisting of finitely many elements, it is called *finite dimensional.* In this case, every basis necessarily has the same number of elements (not obvious, but I won't prove it here); this number is called the *dimension* of the vector space. If a vector space has no finite basis, then it is called *infinite dimensional.*

Example 6.2. $\mathbb{R}^n$ is a finite-dimensional vector space; its dimension is n. A basis for $\mathbb{R}^n$ is given by the vectors

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad e_n = (0, \dots, 0, 1).$$

Example 6.3. The set of continuous functions on $[0, 1]$ is an infinite dimensional vector space.

Once one has played around with the 'algebra' of adding and scaling vectors, it is natural to ask whether or not there is any sense in which we can take *products* of vectors. In fact, the answer is yes, there are several types of products that one could consider. For now, we will focus on one particular choice, namely, the *inner product* or *dot product* on $\mathbb{R}^n$.

Definition 6.1. The *dot product* of two vectors $x, y \in \mathbb{R}^n$ is the real number

$$x \cdot y = x_1 y_1 + \cdots + x_n y_n,$$

which we may also write

$$x \cdot y = \sum_{j=1}^{n} x_j y_j.$$

To be clear: the dot product of two vectors is not another vector, but rather a scalar (i.e. a real number). We can quickly check some algebraic properties of this product (like $x \cdot y = y \cdot x$, and so on), as well as the fact that $x \cdot x \geq 0$ for any $x \in \mathbb{R}^n$ (in fact, $x \cdot x = 0$ if and only if $x = 0$).

What is not yet clear is what the meaning of this 'product' is. Surprisingly, this definition of dot product is all we need to make sense of all of the 'geometry' of $\mathbb{R}^n$. To see this, let us first show how we can use the dot product to define a notion of the *length* or *norm* of a vector in $\mathbb{R}^n$.

Definition 6.2. The (Euclidean) *length* (or *norm*) of $x \in \mathbb{R}^n$ is defined by

$$|x| = \sqrt{x \cdot x} = \left(\sum_{j=1}^{n} x_j^2 \right)^{\frac{1}{2}}.$$

You should check that this corresponds to your usual notion of the length of a vector in $\mathbb{R}^2$ or $\mathbb{R}^3$, say. In particular, we see that $|cx| = |c|\,|x|$ for any $c \in \mathbb{R}$ and $x \in \mathbb{R}^n$, and the only vector of length zero is the zero vector.

Once we have this notion of 'norm', we can also speak of the 'distance' between two vectors:

Definition 6.3. The (Euclidean) *distance* between x and y is defined by

$$d(x,y) = |x-y| = \left(\sum_{j=1}^n (x_j - y_j)^2\right)^{\frac{1}{2}}.$$

Again, this corresponds to our usual notion of Euclidean distance. In particular, $d(x,y) = d(y,x)$ for any $x, y \in \mathbb{R}^n$, and $d(x,y) = 0$ if and only if $x = y$.

We are going to see below that this notion of 'distance' satisfies the other property discussed in Section 5 on metric spaces, namely the triangle inequality. Similarly, we will see that our notion of 'length' satisfies a triangle inequality. The key is the following essential inequality, known as the *Cauchy–Schwarz inequality.*

Theorem 6.1 (Cauchy–Schwarz)**.** *For any $x, y \in \mathbb{R}^n$,*

$$|x \cdot y| \leq |x|\,|y|.$$

Proof. It is enough to consider the case $x \neq 0$ and $y \neq 0$. In this case, we may set

$$u = \tfrac{x}{|x|} \quad \text{and} \quad v = \tfrac{y}{|y|}, \quad \text{so that} \quad |u| = |v| = 1.$$

Then we have

$$0 \leq |u-v|^2 = (u-v)\cdot(u-v) = |u|^2 - 2u\cdot v + |v|^2 = 2 - 2u \cdot v,$$

or

$$u \cdot v \leq 1, \quad \text{which means} \quad x \cdot y \leq |x|\,|y|.$$

Running the same argument with $v = -\frac{y}{|y|}$ also yields $-u \cdot v \leq 1$, and hence the result follows. □

If we expand out the definition of the inner product and norm, the Cauchy–Schwarz inequality reads

$$\left[\sum_{j=1}^n x_j y_j\right]^2 \leq \sum_{j=1}^n x_j^2 \cdot \sum_{j=1}^n y_j^2.$$

The Cauchy–Schwarz inequality implies the important *triangle inequality*:

Lemma 6.2 (Triangle inequality)**.** *For any $x, y \in \mathbb{R}^n$, we have*

$$|x+y| \leq |x| + |y|.$$

Consequently, for any $x, y, z \in \mathbb{R}^n$,

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

Proof. It is enough to prove the first inequality (can you see why?). For this, we use Cauchy–Schwarz to obtain

$$\begin{aligned} |x+y|^2 &= (x+y)\cdot(x+y) \\ &= |x|^2 + 2x\cdot y + |y|^2 \\ &\leq |x|^2 + 2|x|\,|y| + |y|^2 = (|x|+|y|)^2, \end{aligned}$$

which yields the result. □

The Cauchy–Schwarz inequality also allows us to make sense of the notion of the *angle* between two nonero vectors.

Definition 6.4 (Angle)**.** Let x and y be two nonzero vectors in $\mathbb{R}^n$. We define the *angle* between x, y to be the unique $\theta \in [0, \pi]$ such that

$$\cos\theta = \frac{x \cdot y}{|x|\,|y|}.$$

Note that in the case $y = cx$ for some $c > 0$ (so the vectors are parallel), we have

$$x \cdot y = |x|\,|y| \implies \theta = 0,$$

and similarly if $y = cx$ for some $c < 0$ (so the vectors are antiparallel), we have

$$x \cdot y = -|x|\,|y| \implies \theta = \pi.$$

In the case $\theta = \frac{\pi}{2}$ (when the vectors are perpendicular), we have $x \cdot y = 0$. We often use the word *orthogonal* instead of perpendicular, that is,

$$x \cdot y = 0 \iff x \quad \text{and} \quad y \quad \text{are orthogonal.}$$

The computation in the proof of the triangle inequality shows that

$$x \cdot y = 0 \implies |x + y|^2 = |x|^2 + |y|^2,$$

which we know as the Pythagorean Theorem.

A simple example to help you visualize the relationship between dot products and angles is the following: fix $x = (1, 0) \in \mathbb{R}^2$ and let $y = (\cos\theta, \sin\theta) \in \mathbb{R}^2$ for some $\theta \in [0, \pi]$. Then the angle between x and y is θ.

We now have the basics of Euclidean space in place. We defined vectors in $\mathbb{R}^n$, made sense of scalar addition and multiplication, and defined one product (the dot product) that allowed us to define notions of distance, length, and angle.

Before moving on to the next topic, let us point out that the notion of an 'inner product' also extends well beyond the setting of vectors in $\mathbb{R}^n$.

Example 6.4. Let $C([-1, 1])$ be the vector space of continuous functions on $[-1, 1]$. For $f, g \in C([-1, 1])$, define the so-called L^2 *inner product*

$$\langle f, g\rangle = \int_{-1}^{1} f(x)\, g(x)\, dx.$$

This defines a *norm* and *distance* (called the L^2-norm, L^2-distance) via

$$\|f\|_2 = \sqrt{\langle f, f\rangle} = \left(\int_{-1}^{1} |f(x)|^2\, dx\right)^{\frac{1}{2}} \quad \text{and} \quad d_2(f, g) = \|f - g\|_2.$$

We have the Cauchy–Schwarz inequality $|\langle f, g\rangle| \leq \|f\|_2 \|g\|_2$, which means

$$\left|\int f(x) g(x)\, dx\right| \leq \left(\int |f(x)|^2\, dx\right)^{\frac{1}{2}} \left(\int g(x)|^2\, dx\right)^{\frac{1}{2}}.$$

This implies the triangle inequality

$$\|f + g\|_2 \leq \|f\|_2 + \|g\|_2.$$

We also have the notion of the angle between two functions, and say f and g are *orthogonal* if

$$\langle f, g\rangle = 0, \quad \text{i.e.} \quad \int_{-1}^{1} f(x) g(x)\, dx = 0.$$

For example, if f is an even function and g is an odd function, then f and g are orthogonal.

We turn now to our next topic, namely, *linear mappings* between Euclidean spaces.

6.2. **Linear mappings and matrices.** We define a linear mapping (or 'linear transformation') between $\mathbb{R}^n$ and $\mathbb{R}^m$ as follows:

Definition 6.5 (Linear Mapping)**.** A function $L : \mathbb{R}^n \to \mathbb{R}^m$ is *linear* if

$$L(ax + by) = aL(x) + bL(y) \quad \text{for all} \quad a, b \in \mathbb{R} \quad \text{and} \quad x, y \in \mathbb{R}^n.$$

The canonical example of a linear mapping is given by matrix-vector multiplication. We define an $m \times n$ matrix to be a rectangular array of real numbers with m rows and n columns, as follows:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

We denote the rows of A by A_i, where $i = 1, \dots, m$. Note that each A_i is an element of $\mathbb{R}^n$. For example,

$$A_2 = (a_{21}, a_{22}, \dots, a_{2n}), \quad \text{and so on.}$$

We can then define the function $L_A : \mathbb{R}^n \to \mathbb{R}^m$ by imposing that $L_A(x)$ is an element of $\mathbb{R}^m$ satisfying

$$L_A(x) = (A_1 \cdot x, A_2 \cdot x, \dots, A_m \cdot x) \in \mathbb{R}^m.$$

To be clear:

$$\text{an} \quad m \times n \quad \text{matrix} \quad A \quad \text{yields a mapping} \quad L_A : \mathbb{R}^n \to \mathbb{R}^m.$$

Pay attention to the order of the m's and n's—it is very easy to mess this up.

The fact that the mapping $L_A : \mathbb{R}^n \to \mathbb{R}^m$ defined above is actually linear follows from the corresponding property for dot products and the definition of scalar multiplication and addition of vectors. In particular, we have

$$\begin{aligned} L_A(ax + by) &= (A_1 \cdot (ax + by), \dots, A_m \cdot (ax + by)) \\ &= (a(A_1 \cdot x) + b(A_2 \cdot y), \dots, a(A_m \cdot x) + b(A_m \cdot y)) \\ &= a(A_1 \cdot x, \dots, A_m \cdot x) + b(A_1 \cdot y, \dots, A_m \cdot y) \\ &= aL_A(x) + bL_A(y) \end{aligned}$$

for any $a, b \in \mathbb{R}$ and $x, y \in \mathbb{R}^m$.

The collection of $m \times n$ matrices actually forms a vector space (of dimension mn), which we write $\mathbb{R}^{m\times n}$. Given a matrix $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$, we can use different notation to express the vector $L_A(x) \in \mathbb{R}^m$. In particular, if we view $x \in \mathbb{R}^n$ as an $n \times 1$ matrix (we call it a *column vector*), then we may simply write

$$L_A(x) = Ax,$$

where the product on the right is an instance of *matrix multiplication*, and we again view the output (an element of $\mathbb{R}^m$) as an $m \times 1$ column vector. Another viewpoint (used below) is that Ax is obtained by taking a linear combination of the columns A^j of A, namely,

$$Ax = x_1 A^1 + \cdots + x_n A^n.$$

The general definition for matrix multiplication is as follows:

Definition 6.6 (Matrix multiplication)**.** Given an $m \times n$ matrix A and an $n \times p$ matrix B, we define the *matrix product* AB to be the $m \times p$ matrix whose ik^{th} entry is

$$(AB)_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, p.$$

If we write A_i for the rows of A and B^k for the columns of B, we may also write $(AB)_{ik} = A_i \cdot B^k$.

Matrix multiplication satisfies many of the familiar algebraic rules for multiplication, with one key exception, namely, that matrix multiplication is not in general commutative. That is, even if both products AB and BA make sense, you cannot expect $AB = BA$ in general.

Here is what we mean when we say that matrix multiplication is the 'canonical' example of a linear transformation:

Theorem 6.3. *Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear mapping. Then there exists a unique $m \times n$ matrix A such that*

$$f(x) = Ax \quad \textit{for all} \quad x \in \mathbb{R}^n.$$

Proof. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear mapping, then we define A so that its columns are given by $A^j = f(e_j)$ for $j = 1, \ldots, n$. Writing $x \in \mathbb{R}^n$ as

$$x = x_1 e_1 + \cdots + x_n e_n,$$

it follows from linearity that $f(x) = Ax$ for any x:

$$f(x) = x_1 f(e_1) + \cdots x_n f(e_n) = x_1 A^1 + \cdots + x_n A^n = Ax.$$

For uniqueness, we observe that if

$$Ax = f(x) = Bx \quad \text{for all} \quad x \in \mathbb{R}^n,$$

then by evaluating at the standard basis vectors for $\mathbb{R}^n$, we can obtain that each column of A equals each column of B, so that $A = B$. □

Example 6.5. Let's look at two special cases. First, the theorem implies that any linear mapping $f : \mathbb{R}^n \to \mathbb{R}$ is of the form $f(x) = a \cdot x$ for some $a \in \mathbb{R}^n$. Second, the theorem implies that any linear mapping $f : \mathbb{R} \to \mathbb{R}^n$ is of the form $f(x) = xa$ for some $a \in \mathbb{R}^n$.

We have just seen that any linear transformation between $\mathbb{R}^n$ and $\mathbb{R}^m$ essentially 'is' a matrix. Something similar is true for linear transformations between general finite-dimensional vector spaces. In particular, any linear transformation $f : V \to W$ between an n-dimensional vector space V and an m-dimensional vector space W can be represented by an $m \times n$ matrix (after choosing bases for V and W).

The situation is richer in the setting of infinite dimensional vector spaces. Let us just consider a few simple examples:

Example 6.6. Consider the vector space $C([0, 1])$. We can define linear transformations $T : C([0, 1]) \to \mathbb{R}$ and $S : C([0, 1]) \to \mathbb{R}$ by

$$T(f) = f(0) \quad \text{and} \quad S(f) = \int_0^1 f(x)\, dx.$$

I will leave it to you to check that these are linear transformations.

Example 6.7. Let V denote the set of infinitely differentiable functions. Define the linear transformation $T : V \to V$ by

$$Tf = \tfrac{d}{dx} f.$$

We state without proof the following useful proposition, as well:

Proposition 6.4. *If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ are linear transformations, then the composition $g \circ f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear transformation. In fact, if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{k\times n}$ are the matrices for f and g, then BA is the matrix for $g \circ f$, that is*

$$g \circ f(x) = BAx.$$

Given any linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ (so that $f(x) = Ax$ for some $m \times n$ matrix A), there are two associated subspaces that we may consider:

Definition 6.7. The *kernel* (or *null space*) of a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ is the subspace

$$\{x \in \mathbb{R}^n : f(x) = 0\} \subset \mathbb{R}^n.$$

The *image* (or *range*) of f is the subspace

$$\{y \in \mathbb{R}^m : y = f(x) \quad \text{for some} \quad x \in \mathbb{R}^n\} \subset \mathbb{R}^m.$$

It is a general fact of linear algebra that for any linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$, the sum of the dimension of the kernel and the dimension of the image equals n.

In later sections, we will need to make use of the *determinant* of certain $n \times n$ matrices. We use the notation $\det A$ or $|A|$ for the determinant of a matrix A. I assume you have worked with determinants before, but just to remind you, the definition of the determinant is inductive: for a 1×1 matrix A, we just have $\det A = A$. For an $n \times n$ matrix $A = (a_{jk})$, we let A_{jk} denote the $(n-1) \times (n-1)$ submatrix obtained by removing row j and column k from A. Then

$$\det A = \sum_{i=1}^{n} (-1)^{1+j} a_{1j} \det A_{1j}.$$

(This is also called the cofactor expansion along the first row; actually, any row or column may be used.) We will need a few facts about determinants, namely:

- The determinant is multilinear (viewing it as a function of the columns of the matrix) and alternating.
- The determinant of A is equal to the determinant of A^t (i.e. the *transpose* of A, obtained by swapping the rows and columns of A).
- $\det(AB) = \det(A)\det(B)$.

The determinant has a geometric interpretation: given a matrix $A \in \mathbb{R}^{n\times n}$, we consider the image of the standard basis vectors e_j. Then $|\det A|$ is the n-dimensional volume of the paralellepiped spanned by the vectors $\{Ae_j\}$.

The determinant is also connected to the question of *invertibility*. We say that an $n \times n$ matrix A is invertible if there exists a matrix B so that $AB = BA = I_n$, where I_n is the 'identity' matrix containing the standard basic vectors $\{e_j\}$ for its columns. We then have the following:

Theorem 6.5. *An $n \times n$ matrix A is invertible if and only if $\det A \neq 0$. This holds if and only if the columns of A are linearly independent.*

One can also use a formula involving the determinant (known as Cramer's rule) to solve linear systems; however, we won't discuss this here.

6.3. **Limits, continuity, and topology of $\mathbb{R}^n$.** The notions of limits and continuity for functions $f : \mathbb{R}^n \to \mathbb{R}^m$ were already described in the context of metric spaces discussed in Section 5. So, we will only briefly review the main ideas here:

First, we have the following definitions:

- An open ball in $\mathbb{R}^n$ is a set of the form
$$B_r(a) = \{x \in \mathbb{R}^n : |x - a| < r\}.$$
- A point $a \in \mathbb{R}^n$ is a *limit point* of a set $D \subset \mathbb{R}^n$ if
$$\text{for every} \quad r > 0, \quad D \cap [B_r(a) \backslash \{a\}] \neq \emptyset.$$
Note that a limit point of D does not necessarily have to belong to D.
- If $D \subset \mathbb{R}^n$, $f : D \to \mathbb{R}^m$, and $a \in \mathbb{R}^n$ is a limit point of D, then we write
$$\lim_{x \to a} f(x) = b$$
if
$$\text{for any} \quad \varepsilon > 0, \quad \text{there exists} \quad \delta > 0 \quad \text{such that}$$
$$\big[x \in D \quad \text{and} \quad 0 < |x - a| < \delta\big] \implies |f(x) - b| < \varepsilon.$$

The following lemma is also useful for establishing existence of limits in higher dimensions.

Lemma 6.6. *Suppose $D \subset \mathbb{R}^n$ and $f : D \to \mathbb{R}^m$. Write $f_1, \ldots, f_m$ for the m* component functions *of f. Then*
$$\lim_{x \to a} f(x) = b \iff \lim_{x \to a} f_i(x) = b_i \quad \textit{for} \quad i = 1, \ldots, m.$$

Proof. A great exercise in checking the definitions! □

We then have a few more definitions and basic results, which should once again mostly be review:

- Let $D \subset \mathbb{R}^n$. A function $f : D \to \mathbb{R}^m$ is *continuous* at $a \in D$ if
$$\lim_{x \to a} f(x) = f(a). \tag{6.2}$$
Note that this asserts two things: (i) the limit exists, and (ii) the limit equals $f(a)$. If a is not a limit point of D (in which case we call it an *isolated point*), we say that (6.2) is 'vacuously' true, and hence f is automatically continuous at any isolated point.
- Using Lemma 6.6, we can immediately see that a function is continuous at a if and only if each component function is continuous at a.
- As in the case of real-valued functions on $\mathbb{R}$, we have that the finite sum or product continuous functions is continuous, and the composition of two continuous functions is continuous as well. These facts are very useful for proving continuity of many familiar functions (without having to rely directly on the 'ε-δ' definition).

Finally, we discuss a few 'topological' properties of $\mathbb{R}^n$. These topics were covered in Section 5, so we will once again be somewhat brief in our presentation.

We first recall the notion of a convergent sequence. It is basically the same as in the case of real numbers:

Definition 6.8. A sequence $\{x_k\} \subset \mathbb{R}^n$ *converges to* $\ell \in \mathbb{R}^n$ if for any $\varepsilon > 0$, there exists N such that

$$k \geq N \implies |x_k - \ell| < \varepsilon.$$

You should check that a sequence converges if and only if each component sequence converges. (You should also make sure you can see how the completeness of $\mathbb{R}$ implies the completeness of $\mathbb{R}^n$.)

We also need the following:

- A set $S \subset \mathbb{R}^n$ is *open* if for any $x \in S$, there exists $r > 0$ so that $B_r(x) \subset S$. Arbitrary unions of open sets are open, as are finite intersections of open sets.
- A set $S \subset \mathbb{R}^n$ is *closed* if its complement $S^c = \{x \in \mathbb{R}^n : x \notin S\}$ is open. Closed does *not* mean the same thing as'not open'! A useful criterion for checking if a set is closed is the following: a set is closed if and only if it contains all of its limit points. This is equivalent to saying that S is closed if and only if whenever $\{x_k\}$ is a convergent sequence of elements of S, the limit also belongs to S.
- A set $K \subset \mathbb{R}^n$ is *compact* if every sequence in K has a subsequence that converges to a limit in K. We have the important *Heine–Borel Theorem* that says that $K \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.
- A set $D \subset \mathbb{R}^n$ is *connected* if for any two points $x, y \in D$, there exists a continuous function $\gamma : [0,1] \to D$ such that $\gamma(0) = x$ and $\gamma(1) = y$.

Finally, we recall some important theorems about continuous functions on compact and connected sets:

- If $K \subset \mathbb{R}^n$ is compact and $f : K \to \mathbb{R}^m$ is continuous, then the image $f(K)$ is compact in $\mathbb{R}^m$.
- If $K \subset \mathbb{R}^n$ is compact and $f : K \to \mathbb{R}$ is continuous, then f attains maximum and minimum values on K.
- If $K \subset \mathbb{R}^n$ is compact and $f : K \to \mathbb{R}$ is continuous, then f is *uniformly* continuous.
- If $D \subset \mathbb{R}^n$ is connected and $f : D \to \mathbb{R}^m$ is continuous, then $f(D) \subset \mathbb{R}^m$ is connected.

If you need to review any of this material, please refer back to Section 5!

6.4. **Exercises.**

7. Multivariable differential calculus

7.1. **The derivative in higher dimensions.** To get started, let us recall the definition of the derivative of a real-valued function defined on an open interval, say $f : I \to \mathbb{R}$. We said that f is differentiable at a point $a \in \mathbb{R}$ if the limit

$$\lim_{h\to 0} \frac{f(a+h) - f(a)}{h} \quad \text{exists,}$$

and we call this limit the derivative $f'(a)$. Rearranging this, we can express this by saying that there exists $\ell \in \mathbb{R}$ such that defining $R(h)$ via

$$f(a+h) = f(a) + \ell h + R(h), \quad \text{we have} \quad \lim_{h\to 0} \tfrac{R(h)}{h} = 0.$$

(In this case $\ell = f'(a)$.)

The definition in the higher dimensional case (i.e. where $F : \mathbb{R}^n \to \mathbb{R}^m$) is similar to the last formula. However, the parameter h now must be taken to be an element of $\mathbb{R}^n$, and the simple multiplication ℓh is replaced by a *linear transformation* $L(h)$, where $L : \mathbb{R}^n \to \mathbb{R}^m$. That is, we want to be able to write

$$F(a+h) = F(a) + L(h) + R(h), \quad \text{where} \quad \lim_{h\to 0} \frac{|R(h)|}{|h|} = 0$$

for some linear transformation L. The precise definition is the following:

Definition 7.1 (Differentiable)**.** Let $D \subset \mathbb{R}^n$ be an open set and $F : D \to \mathbb{R}^m$. Given $a \in D$, we say that f is *differentiable at* a if there exists a linear mapping $L : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{|h|\to 0} \frac{F(a+h) - F(a) - L(h)}{|h|} = 0. \tag{7.1}$$

Before proceeding, we need to deal with one subtle point. Namely, we should show that if F is differentiable, then the linear transformation appearing in the definition above is necessarily unique. Indeed, we have to check this, since we would like to give this transformation a name (and then use it to define the derivative of F).

Lemma 7.1. *Let $F : D \to \mathbb{R}^m$ be as in Definition 7.1 and $a \in D$. Suppose L_1 and L_2 are linear transformations from $\mathbb{R}^n \to \mathbb{R}^m$ that both satisfy* (7.1)*. Then*

$$L_1(v) = L_2(v) \quad \textit{for all} \quad v \in \mathbb{R}^n, \quad \textit{i.e.} \quad L_1 \equiv L_2.$$

Proof. For $h \in \mathbb{R}^n$, define $R_1(h)$ and $R_2(h)$ by setting

$$F(a+h) = F(a) + L_j(h) + R_j(h), \quad j = 1, 2. \tag{7.2}$$

Then, by the assumption (7.1), we have

$$\lim_{|h|\to 0} \frac{|R_j(h)|}{|h|} = 0, \quad j = 1, 2. \tag{7.3}$$

Rearranging (7.2), we can write

$$L_1(x) - L_2(x) = R_2(x) - R_1(x) \quad \text{for any} \quad x \in \mathbb{R}^n. \tag{7.4}$$

Now fix $v \in \mathbb{R}^n \backslash \{0\}$ and apply (7.4) with the sequence of vectors $x_n = \frac{1}{n}v$, where $n \in \mathbb{N}$. Using linearity of L_1 and L_2, this implies

$$\begin{aligned} \tfrac{1}{n}[L_1(v) - L_2(v)] &= R_2(\tfrac{1}{n}v) - R_2(\tfrac{1}{n}v) \\ \implies L_1(v) - L_2(v) &= |v| \cdot \left[\frac{R_2(\frac{1}{n}v)}{|\frac{1}{n}v|} - \frac{R_2(\frac{1}{n}v)}{|\frac{1}{n}v|} \right] \end{aligned}$$

However, by (7.3), the right-hand side tends to zero as $n \to \infty$ (since $|\frac{1}{n}v| \to 0$). Thus we conclude

$$L_1(v) = L_2(v) \quad \text{for all} \quad v \in \mathbb{R}^n \backslash \{0\}.$$

As L_1 and L_2 are both linear, we also have $L_1(0) = L_2(0) = 0$. Thus we conclude $L_1 \equiv L_2$, as desired. □

With uniqueness established, we can now make the following definition:

Definition 7.2 (Differential; derivative)**.** Suppose $D \subset \mathbb{R}^n$ and $F : D \to \mathbb{R}^m$ is differentiable at a point $a \in D$.

- We denote the (unique) linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ appearing in (7.1) by dF_a, and we call this linear transformation the *differential of* F *at* a.
- We denote the (unique) $m \times n$ matrix of dF_a by $F'(a)$ (see Theorem 6.3), and we call this the *derivative of* F *at* a.

In particular,

$$dF_a(x) = F'(a)x \quad \text{for all} \quad x \in \mathbb{R}^n,$$

where the right-hand side is the matrix-vector product.

We have now defined the notion of a derivative of a function $F : \mathbb{R}^n \to \mathbb{R}^m$, albeit in a somewhat abstract way (namely, as the matrix of some linear transformation). However, we can also recognize the entries of the matrix $F'(a)$ as more familiar objects, namely, partial derivatives of the component functions of F.

Theorem 7.2 (Entries of the derivative matrix)**.** *Suppose $D \subset \mathbb{R}^n$ and $F : D \to \mathbb{R}^m$ is differentiable at $a \in D$. Then, writing $(F^1, \dots, F^m)$ for the components of F, we have that the entries of $F'(a)$ are given by*

$$[F'(a)]_{jk} = D_k F^j(a) := \lim_{h \to 0} \tfrac{F^j(a+he_k) - F^j(a)}{h} \tag{7.5}$$

for $j = 1, \dots, m$ and $k = 1, \dots n$.

Remark 7.3. A few remarks are in order. First, the quantity

$$D_k F^j(a), \quad \text{also denoted} \quad \frac{\partial F^j}{\partial x_k}(a)$$

is called the *partial derivative* of the function F^j with respect to x_k. It is defined by the limit in (7.5), and part of the theorem is the assertion that this limit exists. In the formula (7.5), e_k refers to the k^{th} standard basis vector, and h here is just an element of $\mathbb{R}$ (not $\mathbb{R}^n$).

Proof. By assumption, we have

$$\lim_{|h| \to 0} \frac{F(a+h) - F(a) - F'(a)h}{|h|} = 0.$$

In particular, for each $j \in \{1, \ldots, m\}$, we have

$$\lim_{|h| \to 0} \frac{F^j(a+h) - F^j(a) - [F'(a)h]^j}{|h|} = 0,$$

where $[F'(a)h]^j$ denotes the j^{th} component of $F'(a)h$. In particular,

$$[F'(a)h]^j = \sum_{\ell=1}^{n} [F'(a)]_{j\ell} h^\ell,$$

where h^k denotes the k^{th} component of h.

Now fix $k \in \{1, \ldots, n\}$ and consider a sequence of the form $h_m e_k$ where $h_m \in \mathbb{R}$ satisfies $h_m \to 0$ as $m \to \infty$, and e_k is the k^{th} standard basis vector. Then we have

$$[F'(a)h_m e_k]^j = h_m \sum_{\ell=1}^{n} [F'(a)]_{j\ell} e_k^\ell = h_m [F'(a)]_{jk},$$

and so

$$\lim_{m \to \infty} \frac{F^j(a + h_m e_k) - F^j(a) - h_m [F'(a)]_{jk}}{h_m} = 0,$$

or, rearranging:

$$\lim_{m \to \infty} \frac{F^j(a + h_m e_k) - F^j(a)}{h_m} = [F'(a)]_{jk}.$$

As this holds for an arbitrary sequence $h_m \to 0$, we can conclude that

$$\lim_{h \to 0} \frac{F^j(a + h e_k) - F^j(a)}{h} = [F'(a)]_{jk},$$

which shows that the partial derivative exists and

$$D_k F^j(a) = [F'(a)]_{jk}.$$

□

Let's work through several examples to clarify some of the ideas above.

Example 7.1. Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation. Then we have

$$F(a+h) = F(a) + F(h) \quad \text{for all} \quad a, h \in \mathbb{R}^n$$

which implies that F is differentiable for all $a \in \mathbb{R}^n$, and in fact

$$dF_a = F \quad \text{for all} \quad a \in \mathbb{R}^n.$$

In particular, if we denote the matrix of F by A, then $F'(x) = A$ for all $x \in \mathbb{R}^n$. Put differently,

$$\text{the derivative of} \quad F(x) = Ax \quad \text{is given by} \quad F'(x) = A.$$

This is exactly what we expect based on the scalar case!

Example 7.2. Let $F : \mathbb{R}^2 \to \mathbb{R}^4$ be given by

$$F(x, y) = (y, x, xy, y^2 - x^2).$$

Then the partial derivatives all exist *and are continuous*, which implies that F is differentiable (see Lemma 7.4 below). At an arbitrary point (x, y), the derivative matrix is given by

$$F'(x,y) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ y & x \\ -2x & 2y \end{bmatrix}.$$

Example 7.3. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

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

All of the partial derivatives of f exist at $(x, y) = (0, 0)$. To see this, first note

$$\frac{f(h,0) - f(0,0)}{h} = \frac{1}{h}\frac{h \cdot 0}{h^2} \equiv 0, \quad \text{so that} \quad D_x f(0,0) = 0,$$

and similarly

$$D_y f(0,0) = 0.$$

Thus, if the differential $df_{(0,0)}$ exists, it must be the zero transformation; in particular, we must have

$$\lim_{|k|\to 0} \frac{f(k)}{|k|} = 0.$$

But observe that if we choose $k = (h, h)$ with $h > 0$, then

$$\frac{f(h,h)}{|(h,h)|} = \frac{1}{\sqrt{2}h}\frac{h^3}{2h^2} = \frac{1}{2\sqrt{2}} \quad \text{as} \quad h \to 0.$$

Thus f is not differentiable at $(0, 0)$.

The last example showed that just having partial derivatives is not enough to guarantee differentiability. However, if the partial derivatives are continuous, this does guarantee differentiability:

Lemma 7.4. *If the partial derivatives of F exist and are continuous at a, then F is differentiable at a.*

Proof. It is enough to consider a function $F : \mathbb{R}^n \to \mathbb{R}$ (for differentiability of $F : \mathbb{R}^n \to \mathbb{R}^m$ is equivalent to differentiability of each of its component functions).

We will prove differentiability of F at a by verifying that the differential of F at a is given by the linear transformation

$$L(h) = \sum_{i=1}^n D_i F(a) h_i$$

(which must be the case, if F is indeed differentiable). Thus our task is to prove

$$\lim_{h\to 0} \frac{F(a+h) - F(a) - \sum_{i=1}^n D_i F(a) h_i}{|h|} = 0.$$

To this end, given $h = (h_1, \dots, h_n) \in \mathbb{R}^n \backslash \{0\}$ we set

$$\tilde{h}_0 = 0, \quad \text{and} \quad \tilde{h}_i = (h_1, \dots, h_i, 0, \dots 0) \quad \text{for} \quad i = 1, \dots, n.$$

Then we may write

$$F(a+h)-F(a)=\sum_{i=1}^{n}[F(a+\tilde{h}_i)-F(a+\tilde{h}_{i-1})],$$

and so our task becomes to show

$$\lim_{|h|\to 0}\left\{\tfrac{1}{|h|}\sum_{i=1}^{n}[F(a+\tilde{h}_i)-F(a+\tilde{h}_{i-1})-D_iF(a)h_i]\right\}=0. \tag{7.6}$$

Now, the point is that we may write

$$F(a+\tilde{h}_i)-F(a+\tilde{h}_{i-1})=g_i(a_i+h_i)-g(a_i),$$

where $g_i:\mathbb{R}\to\mathbb{R}$ is defined by

$$g_i(x)=f(a_1+h_1,\dots,a_{i-1}+h_{i-1},x,a_{i+1},\dots,a_n).$$

In particular, by the Mean Value Theorem, we can write

$$g_i(a_i+h_i)-g_i(a_i)=g_i'(c_i)h_i \quad \text{for some} \quad c_i \quad \text{between} \quad a_i \quad \text{and} \quad a_i+h_i,$$

and by the definition of g_i, we can write

$$g_i'(c_i)=D_iF(b_i), \quad b_i=(a_1+h_1,\dots,a_{i-1}+h_{i-1},c_i,a_{i+1},\dots,a_n).$$

That is, we have

$$F(a+\tilde{h}_i)-F(a+\tilde{h}_{i-1})=D_iF(b_i)h_i,$$

where b_i have the property that

$$\lim_{|h|\to 0}b_i=a \quad \text{for} \quad i=1,\dots,n.$$

Returning to (7.6), the problem has now reduced to showing that

$$\lim_{|h|\to 0}\tfrac{h_i}{|h|}[D_iF(b_i)-D_iF(a)]=0 \quad \text{for each} \quad i=1,\dots,n.$$

But now we are in business, each since $\frac{h_i}{|h|}$ is bounded by 1 and (since the partial derivatives of F are continuous)

$$|h|\to 0 \implies b_i\to a \implies D_iF(b_i)\to D_iF(a).$$

Done! □

At this point, we have defined the derivative and partial derivatives and have seen a bit about the relationship between them. Next, we will introduce a few other related notions and some more special cases.

We first introduce the notion of a 'directional derivative'.

Definition 7.3. Suppose $F:\mathbb{R}^n\to\mathbb{R}^m$, $a\in\mathbb{R}^n$, and $v\in\mathbb{R}^n\backslash\{0\}$. We define the *directional derivative of F with respect to v at the point a* by

$$D_vF(a)=\lim_{h\to 0}\frac{F(a+hv)-F(a)}{h},$$

provided this limit exists. Note that here $h\in\mathbb{R}$ is a scalar.

If the function F is differentiable at a, then it has directional derivatives in every direction. In particular, the directional derivatives in the directions $e_1,\dots,e_n$ coincide with the partial derivatives $D_1F,\dots,D_nF$ (which are the columns consisting of $(D_kF^j)_{j=1}^m$, with $k=1,\dots,n$), and we can use these to compute all directional derivatives:

Lemma 7.5. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and $a \in \mathbb{R}^n$. If F is differentiable at a, then for any $v = (v_1, \ldots, v_n) \in \mathbb{R}^n \backslash \{0\}$, the directional derivative $D_v F(a)$ exists. In fact,*

$$D_v F(a) = dF_a(v) = \sum_{j=1}^{n} v_j D_j F(a).$$

Proof. For $t \in \mathbb{R}$, we have by definition of differentiability that

$$\lim_{t \to 0} \frac{F(a + tv) - F(a) - dF_a(tv)}{|tv|} = 0.$$

As

$$dF_a(tv) = t dF_a(v),$$

it follows that

$$\lim_{t \to 0} \frac{F(a + tv) - F(a)}{t} - dF_a(v) = 0,$$

which shows that $D_v F(a)$ exists and equals $dF_a(v)$. For the second equality, we recall that $D_j F(a)$ are the columns of the derivative matrix $F'(a)$, so that

$$D_v F(a) = dF_a(v) = \sum_{j=1}^{n} v_j dF_a(e_j) = \sum_{j=1}^{n} v_j [F'(a)] e_j = \sum_{j=1}^{n} v_j D_j F(a).$$

□

As with partial derivatives, it is possible to have directional derivatives in every direction and yet fail to be differentiable.

Example 7.4. Take the same example as before: let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

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

Let $v \in \mathbb{R}^2 \backslash \{0\}$ and observe that

$$f(tv) = \frac{t^3 v_1 v_2}{t^2 [v_1^2 + v_2^2]} = t f(v).$$

Thus

$$D_v f(0) = \lim_{t \to 0} \frac{f(tv) - f(0)}{t} = f(v).$$

In particular, all directional derivatives exist at $(0, 0)$, but we have already seen that f is not differentiable at $(0, 0)$.

Example 7.5 (The gradient and directional derivatives)**.** Suppose f is a scalar-valued function, i.e. $f : \mathbb{R}^n \to \mathbb{R}$. Suppose further that f is differentiable at $a \in \mathbb{R}^n$. Then the derivative $F'(a)$ is a $1 \times n$ matrix, i.e. a 'row vector', with entries given by the partial derivatives of f. We call this the *gradient vector* of f at a, denoted

$$\nabla f(a) = (D_1 f(a), \ldots, D_n f(a)) \in \mathbb{R}^n.$$

In particular, the conclusion of Lemma 7.5 may be restated

$$D_v F(a) = \nabla f(a) \cdot v,$$

giving a simple expression for directional derivatives.

You may remember from multivariable calculus that the gradient vector points in the direction of most rapid increase, and that the gradient vector vanishes at extreme values. We will cover these topics in more detail below.

Example 7.6 (Curves in $\mathbb{R}^m$)**.** We may also consider the case of a function $f : \mathbb{R} \to \mathbb{R}^m$, which we view as a curve inside $\mathbb{R}^m$ (perhaps the trajectory of some object). In this case, if f is differentiable at a, then its derivative $f'(a)$ is a $m \times 1$ matrix, i.e. a column vector, consisting of the ordinary derivatives of its m components. We interpret the vector $f'(a)$ as the *velocity vector* of the object.

What is the purpose of the derivative? One possible answer is that the derivative provides us with the best (local) *linear approximation* to a function. That is, if we suppose that $F : D \to \mathbb{R}^m$ (with $D \subset \mathbb{R}^n$) is differentiable at $a \in D$, then we can define the transformation

$$T(x) = F(a) + F'(a)[x - a].$$

This is not actually linear. Instead, it is called *affine* (a fixed translation of a linear transformation). By the definition of differentiability, $|F(x) - T(x)| \to 0$ as $x \to a$; in fact, we have the stronger statement that $|F(x) - T(x)| = o(|x - a|)$, which means we can divide by $|x - a|$ and the difference still tends to zero. The image of $\mathbb{R}^n$ under T is then a linear (actually, affine...) approximation to the image $F(D) \subset \mathbb{R}^m$.

Example 7.7. Let $D \subset \mathbb{R}^2$ be given by

$$D = \{(x, y) : x^2 + y^2 < 1\}$$

and $f : D \to \mathbb{R}^3$ be given by

$$f(x, y) = (x, y, 1 - x^2 - y^2).$$

Then the image $f(D)$ is a paraboloid in $\mathbb{R}^3$. The function f is differentiable, with

$$f'(x, y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -2x & -2y \end{bmatrix}.$$

The linear approximation to f at the point $(x, y) = (0, 0)$ is given by

$$T(x, y) = (0, 0, 1) + f'(0, 0)(x, y) = (x, y, 1).$$

As (x, y) vary in $\mathbb{R}^2$, this simply traces out a copy of the xy-plane, translated to height 1.

The linear approximation to f at the point $(x, y) = (\frac{1}{2}, \frac{1}{2})$ is

$$T(x, y) = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} x - \frac{1}{2} \\ y - \frac{1}{2} \end{bmatrix} = \begin{bmatrix} x \\ y \\ \frac{3}{2} - x - y \end{bmatrix}$$

In particular, the linear approximation to the surface $f(D)$ at $(x, y) = (\frac{1}{2}, \frac{1}{2})$ is the plane $z = \frac{3}{2} - x - y$.

To close this section, let us record one useful fact that carries over to the higher-dimensional case:

Proposition 7.6. *If $F : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at a, then F is continuous at a.*

Proof. For $h \neq 0$, write

$$F(a + h) - F(a) = |h| \cdot \frac{F(a + h) - F(a) - dF_a(h)}{|h|} + dF_a(h).$$

The first term tends to zero as $|h| \to 0$ by definition of differentiability. The second term tends to zero as well (since we may write $dF_a(h) = |h| F'(a) \frac{h}{|h|}$). □

7.2. **The higher dimensional chain rule.** Recall that if $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable functions, then the composition $f \circ g$ is differentiable, with

$$(g \circ f)'(x) = g'(f(x))f'(x).$$

In this section, we will establish a higher-dimensional analogue of this result:

Theorem 7.7 (The chain rule)**.** *Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open sets. Suppose $F : U \to \mathbb{R}^m$ is differentiable at $a \in U$ and $G : V \to \mathbb{R}^k$ is differentiable at $F(a) \in V$. Then the composition $H = G \circ F$ is differentiable at a, and we have*

$$dH_a = dG_{F(a)} \circ dF_a$$

as a composition of linear mappings. Consequently, in terms of the derivatives, we have

$$H'(a) = G'(F(a))F'(a)$$

as a product of matrices.

Proof. Define

$$\varphi(h) = \frac{F(a+h) - F(a) - dF_a(h)}{|h|}, \quad h \neq 0$$

and

$$\psi(k) = \frac{G(F(a)+k) - G(F(a)) - dG_{F(a)}(k)}{|k|}, \quad k \neq 0.$$

By definition of differentiability, we have

$$\lim_{h\to 0} \varphi(h) = 0 \quad \text{and} \quad \lim_{k \to 0} \psi(k) = 0.$$

Thus (using the definition of ψ with $k = F(a+h) - F(a)$) we have

$$\begin{aligned} &H(a+h) - H(a) \\ &\quad = G(F(a+h)) - G(F(a)) \\ &\quad = G(F(a) + [F(a+h) - F(a)]) - G(F(a)) \\ &\quad = dG_{F(a)}(F(a+h) - F(a)) + |F(a+h) - F(a)|\psi(F(a+h) - F(a)). \end{aligned}$$

Now, using the definition of φ and linearity of $dG_{F(a)}$ and dF_a, we can write

$$\begin{aligned} &H(a+h) - H(a) \\ &\quad = dG_{F(a)}(dF_a(h)) + |h|\varphi(h)) + \big| |h|\varphi(h) + dF_a(h)\big| \cdot \psi(F(a+h) - F(a)) \\ &\quad = dG_{F(a)}(dF_a(h)) + |h| dG_{F(a)}(\varphi(h)) \\ &\qquad + |h| \cdot |\varphi(h) + dF_a(\tfrac{h}{|h|})| \cdot \psi(F(a+h) - F(a)). \end{aligned}$$

Rearranging, we find that

$$\begin{aligned} &\frac{H(a+h) - H(a) - dG_{F(a)} \circ dF_a(h)}{|h|} \\ &\qquad = dG_{F(a)}(\varphi(h)) + |\varphi(h) + dF_a(\tfrac{h}{|h|})| \cdot \psi(F(a+h) - F(a)). \end{aligned}$$

We now claim that the right-hand side tends to zero as $|h| \to 0$, which will complete the proof. To see this, observe:

- The first term is a matrix multiplied by $\varphi(h)$; the latter tends to zero as $|h| \to 0$.

- The second term contains $\psi(F(a+h)-F(a))$. As F is continuous (indeed, it is differentiable), we have $F(a+h)-F(a) \to 0$ as $|h| \to 0$; then since $\psi(k) \to 0$ as $|k| \to 0$, this term tends to zero as well.

□

Example 7.8. Suppose $x : \mathbb{R} \to \mathbb{R}^m$ describes the trajectory of some particle, and $V : \mathbb{R}^m \to \mathbb{R}$ is some scalar 'potential energy' function. Then $V \circ x : \mathbb{R} \to \mathbb{R}$ gives the potential energy of the particle at each time, and the chain rule implies

$$\tfrac{d}{dt} V(x(t)) = \nabla V(x(t)) \cdot x'(t).$$

I will leave it to you to work out some other familiar examples. In particular, expanding out the matrices, you can derive formulas like the following: if $u = u(x,y)$ and $x = x(t)$, $y = y(t)$, then

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial t}.$$

Example 7.9. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be the polar coordinate mapping, i.e.

$$T(r,\theta) = (r\cos\theta, r\sin\theta),$$

and given differentiable $f : \mathbb{R}^2 \to \mathbb{R}$, let us define

$$g(r,\theta) = f \circ T(r,\theta) = f(r\cos\theta, r\sin\theta).$$

Then

$$\frac{\partial g}{\partial r} = \frac{\partial f}{\partial x}\cos\theta + \frac{\partial f}{\partial y}\sin\theta, \quad \frac{\partial g}{\partial \theta} = -\frac{\partial f}{\partial x} r\sin\theta + \frac{\partial f}{\partial y} r\cos\theta.$$

If one computes higher order derivatives (e.g. viewing $\frac{\partial f}{\partial x}$ as a function from $\mathbb{R}^2 \to \mathbb{R}$ and then computing a partial derivative of this function), one can obtain the identity

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 g}{\partial r^2} + \frac{1}{r}\frac{\partial g}{\partial r} + \frac{1}{r^2}\frac{\partial^2 g}{\partial \theta^2}.$$

The second-order differential operator on the left-hand side is important in many physical applications. It is called the *Laplacian* and is denoted by Δ (or sometimes ∇^2 in the physics literature). The formula above gives a representation of the Laplacian in polar coordinates; this is a typical application of the chain rule.

Example 7.10. Consider the one-dimensional wave equation

$$\frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial x^2}, \quad \text{where} \quad f : \mathbb{R}^2 \to \mathbb{R} \quad \text{is the unknown.}$$

To solve this equation, one can try to introduce a change of variables

$$\begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

for some unknowns A, B, C, D. We then define

$$g(u,v) = f(t(u,v), x(u,v)) = f(Au + Bv, Cu + Dv).$$

We can 'find' the second derivatives $\frac{\partial^2 f}{\partial t^2}$ and $\frac{\partial^2 f}{\partial x^2}$ by taking second derivatives of g. In particular,

$$\frac{\partial g}{\partial u} = A\frac{\partial f}{\partial t} + C\frac{\partial f}{\partial x}$$

and

$$\frac{\partial^2 g}{\partial v \partial u} = AB\frac{\partial^2 f}{\partial t^2} + (AD+BC)\frac{\partial^2 f}{\partial t \partial x} + CD\frac{\partial^2 f}{\partial x^2}.$$

Let us now choose A, B, C, D so that $AB = 1$, $CD = -1$, and $AD + BC = 0$. We can achieve this by taking

$$A = B = 1, \quad C = 1, \quad D = -1.$$

Then if f is to solve the wave equation, we find that

$$g(u,v) = f(u+v, u-v) \quad \text{must solve} \quad \frac{\partial^2 g}{\partial u \partial v} \equiv 0$$

For this latter equation, we see that g must have the form

$$g(u,v) = a(u) + b(v) \quad \text{for some functions} \quad a, b.$$

Inverting the transformation from $(t,x) \to (u,v)$, this implies that

$$f(t,x) = a(x+t) + b(x-t) \quad \text{for some functions} \quad a, b.$$

That is, f is the sum of two traveling waves, one moving in the positive x direction and one moving in the negative x direction. In general, we would be given an initial position and velocity for f (i.e. $f(0,x)$ and $D_t f(0,x)$) and use these two determine the precise functions a and b. Proceeding in this way leads to the *d'Alembert solution* to the wave equation (after Jean le Rond d'Alembert, 1717–1783).

We can use the chain rule to prove a few other useful facts (all of which are analogues of results in the one-dimensional case).

Theorem 7.8. *Let U be an open, connected subset of $\mathbb{R}^n$. A function $F : U \to \mathbb{R}^m$ is constant if and only if $F'(x) = 0$ for all $x \in U$. Consequently, if two differentiable functions $F, G : U \to \mathbb{R}^m$ satisfy $F'(x) = G'(x)$ for all $x \in U$, then $F - G$ is a constant function.*

Proof. It is enough to consider the case $m = 1$ (why?). It is straightforward to show that constant functions have zero derivative, so let us prove the converse. That is, let us suppose that $\nabla f(x) = 0$ for all $x \in U$ and show that f is constant.

We choose arbitrary $a, b \in U$ and want to show that $f(a) = f(b)$. As U is connected, we can choose a continuous function $\gamma : [0,1] \to U$ so that $\gamma(0) = a$ and $\gamma(1) = b$. Suppose for now that we also knew that γ were differentiable. Then, by the chain rule, we would have

$$\tfrac{d}{dt}[f \circ \gamma] = \nabla f(\gamma(t)) \cdot \gamma'(t) \equiv 0,$$

so that $f \circ \gamma : [0,1] \to \mathbb{R}$ must be constant. In particular,

$$f(b) = f \circ \gamma(1) = f \circ \gamma(0) = f(a),$$

as desired.

Now, we don't actually know that γ has to be differentiable. If you are happy assuming that γ is differentiable, that's fine. Otherwise, the argument can be modified as follows (but feel free to skip this part): We can first extend γ to be continuous on all of $\mathbb{R}$ (but zero outside of $[-1,2]$, say). We then take a differentiable function $K : \mathbb{R} \to \mathbb{R}$ that is positive, zero outside of $[-1,1]$, and obeys

$$\int_{-1}^{1} K(x)\,dx = 1. \tag{7.7}$$

We then define $K_n(x) = nK(nx)$ and note that by a change of variables, each K_n still satisfies (7.7). We then define

$$\gamma_n(x) = \int_{\mathbb{R}} K_n(x-y)\gamma(y)\,dy$$

(that is, we define each component of γ_n by a Riemann integral). One can then show that (i) each γ_n is differentiable, (ii) $\gamma_n \to \gamma$ uniformly on $[0,1]$, and (iii) for n sufficiently large, $\gamma_n([0,1]) \subset U$. Then we still have

$$\tfrac{d}{dt}[f\circ\gamma_n] = \nabla f(\gamma_n(t))\cdot\gamma_n'(t) \equiv 0 \quad \text{for each} \quad n,$$

so that each $f\circ\gamma_n$ is constant. However, since

$$\lim_{n\to\infty} f\circ\gamma_n(0) = f(a) \quad \text{and} \quad \lim_{n\to\infty} f\circ\gamma_n(1) = f(b),$$

we can derive that $|f(b)-f(a)| < \varepsilon$ for any $\varepsilon > 0$, which implies the result. □

As another application of the chain rule, we can prove a Mean Value Theorem for real-valued, differentiable functions on $\mathbb{R}^n$:

Theorem 7.9 (Mean Value Theorem)**.** *Let $U \subset \mathbb{R}^n$ be open. Suppose $a, b \in U$ and that the line segment L joining a to b is entirely contained in U. If $f : U \to \mathbb{R}$ is differentiable, then*

$$f(b) - f(a) = \nabla f(c)\cdot(b-a) \quad \textit{for some} \quad c \in L.$$

Proof. Let $\varphi : [0,1] \to U$ be given by

$$\varphi(t) = (1-t)a + tb, \quad \text{so that} \quad \varphi'(t) \equiv b - a.$$

Now define $g(t) = f(\varphi(t))$, so that $g : [0,1] \to \mathbb{R}$. By the one-dimensional Mean Value Theorem, there exists $\xi \in [0,1]$ so that

$$g(1) - g(0) = g'(\xi).$$

However, this means

$$f(b) - f(a) = g(1) - g(0) = g'(\xi) = \nabla f(\varphi(\xi))\cdot\varphi'(\xi) = \nabla f(c)\cdot(b-a),$$

where $c = \varphi(\xi) \in L$. □

As another application, we can establish the 'equality of mixed partial derivatives' under the assumption that the derivatives are continuous. (This is called *Clairaut's Theorem*, after Alexis Clairaut, 1713–1765.)

Theorem 7.10 (Equality of mixed partial derivatives)**.** *Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$. If the first and second derivatives of f exist and are continuous on U, then we have $D_iD_jf = D_jD_if$ for all $i, j = 1, \dots, n$.*

We begin with a lemma concerning the following quantity (sometimes called a *second difference*):

$$\Delta^2 f_a(h,k) := f(a+h+k) - f(a+h) - f(a+k) + f(a).$$

Lemma 7.11. *Let $U \subset \mathbb{R}^n$ be open and contain the parallelogram determined by the points $a, a+h, a+k$, and $a+h+k$. If $f : U \to \mathbb{R}$ and $D_hf : U \to \mathbb{R}$ are differentiable, then there exist $\alpha, \beta \in (0,1)$ so that*

$$\Delta^2 f_a(h,k) = D_kD_hf(a + \alpha h + \beta k).$$

Proof. We define the function

$$g(x) = f(x+k) - f(x),$$

which is differentiable on an open set containing the line segment joining a and $a+h$. Then, using the Mean Value Theorem (Theorem 7.9), we may write

$$\Delta^2 f_a(h,k) = g(a+h) - g(a) = \nabla g(a+\alpha h)\cdot h$$

for some $\alpha \in (0,1)$. Now, using Lemma 7.5, we write

$$\nabla g(a+\alpha h)\cdot h = D_h g(a+\alpha h) = dg_{a+\alpha h}(h).$$

By the definition of g, we can see that

$$dg_{a+\alpha h}(h) = df_{a+\alpha h+k}(h) - df_{a+\alpha h}(h),$$

which by the same reasoning as above can be rewritten

$$\begin{aligned} dg_{a+\alpha h}(h) &= D_h f(a+\alpha h+k) - D_h f(a+\alpha h) \\ &= \nabla[D_h f](a+\alpha h+\beta k)\cdot k \\ &= D_k D_h f(a+\alpha h+\beta k) \end{aligned}$$

for some $\beta \in (0,1)$. □

Proof of Theorem 7.10. By assumption, $D_j f$ and $D_i f$ are both differentiable (see Lemma 7.4). Fix $a \in U$.

By Lemma 7.11, for all h,k sufficiently small, we can find $\alpha_1,\beta_1 \in (0,1)$ so that

$$\Delta^2 f_a(he_i, ke_j) = D_{ke_j} D_{he_i} f(a+\alpha_1 he_i + \beta_1 ke_j).$$

Similarly, we can find $\alpha_2,\beta_2 \in (0,1)$ such that

$$\Delta^2 f_a(ke_j, he_i) = D_{he_i} D_{ke_j} f(a+\alpha_2 ke_j + \beta_2 he_i).$$

However, by the definition of $\Delta^2 f_a$, we can verify that $\Delta^2 f_a(X,Y) \equiv \Delta^2 f_a(Y,X)$. Noting that

$$D_{he_i} f(X) = \nabla f(X)\cdot he_i = h\nabla f(X)\cdot e_i = hD_i f(X)$$

(and similarly $D_{ke_j} f = kD_j f$), we therefore derive

$$hk\cdot D_j D_i f(a+\alpha_1 he_i + \beta_1 e_j) = hk\cdot D_i D_j (f+\alpha_2 ke_j + \beta_2 he_i).$$

As the second partial derivatives are assumed to be continuous, we can therefore send $h,k\to 0$ to deduce $D_j D_i f(a) = D_i D_j f(a)$, as desired. □

Example 7.11. For a typical looking function like

$$f(x,y) = x^2y^4 + 3x^5y^2 + xy,$$

one can readily check that $D_x D_y f = D_y D_x f$. In this case, this fact is guaranteed by Clairaut's Theorem because the function is infinitely differentiable, with all derivatives continuous.

Example 7.12. Equality of mixed partials is not guaranteed unless we check the hypotheses of Clairaut's Theorem. Indeed, consider the example

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

Then a bit of calculation shows

$$D_1 f(0,y) = -y \quad\text{and}\quad D_2 f(x,0) = x \quad\text{for all}\quad x,y.$$

In particular, we can obtain that

$$D_1D_2f(0,0) = 1 \quad \text{but} \quad D_2D_1f(0,0) = -1.$$

What went wrong?

7.3. **Taylor's formula and classification of critical points.** Our goal in this section is to establish a 'Taylor series' or 'Taylor polynomial' expansion for a function $f : \mathbb{R}^n \to \mathbb{R}$. The motivating theorem to keep in mind is the Lagrange Remainder, Theorem 2.4, which we may write in the form

$$f(a+h) = f(a) + f'(a)h + \tfrac{f''(a)}{2!}h^2 + \cdots + \tfrac{f^{(k)}(a)}{k!}h^k + \tfrac{f^{(k+1)}(c)}{(k+1)!}h^{k+1} \tag{7.8}$$

for some c between a and $a+h$. Clearly, to develop any kind of analogue of this result, it will be necessary that we better understand higher derivatives of functions of multiple variables.

Definition 7.4. Let $f : U \to \mathbb{R}$ for some open set $U \subset \mathbb{R}^n$. We say a $f \in C^k(U)$ if all iterated partial derivatives of f up to order k exist and are continuous on U. That is, if $i_1, \dots i_q \in \{1, \dots, n\}$ and $0 \le q \le k$, then

$$D_{i_1} \cdots D_{i_q} f$$

exists and is continuous on U.

Note that by Lemma 7.4 and Clairaut's Theorem (Theorem 7.10), if $f \in C^k(U)$ then the order of partial derivatives does not matter. For example,

$$D_1D_2D_3D_1f = D_1D_1D_2D_3f,$$

and so on. We use the notation $D_j^k f$ to denote

$$\underbrace{D_j \cdots D_j}_{k \text{ times}} f.$$

Similarly, if $v \in \mathbb{R}^n$, we denote the repeated directional derivative by

$$D_v^k f = \underbrace{D_v \cdots D_v}_{k \text{ times}} f.$$

Now, let us return to (7.8) and re-cast it in a form that is amenable to generalization to higher dimensions. The key is to observe that in the case $f : \mathbb{R} \to \mathbb{R}$, 'directional derivatives' are particularly simple. In particular, since the gradient $\nabla f(a)$ just equals $f'(a)$ and the 'dot product' is just multiplication, we have

$$f'(a)h = D_h f(a), \quad \text{i.e.} \quad D_h = h\tfrac{d}{dx}.$$

Thus the general term $h^k f^{(k)}(a)$ may be rewritten as $D_h^k f(a)$, and so (7.8) may be written

$$f(a+h) = \sum_{\ell=0}^{k} \frac{1}{\ell!} D_h^\ell f(a) + \tfrac{1}{(k+1)!} D_h^{k+1} f(c).$$

Viewing D_h as the directional derivative in direction $h \in \mathbb{R}^n$, we see that this formula at least *makes sense* for a function $f : \mathbb{R}^n \to \mathbb{R}$ belonging to $C^{k+1}(U)$. In fact, not only does the formula make sense, but it is still true in higher dimensions!

Theorem 7.12 (Taylor's Formula / Lagrange Remainder in Higher Dimensions)**.** *Let $U \subset \mathbb{R}^n$ and $f \in C^{k+1}(U)$. Suppose U contains the line segment L joining a and $a+h$. Then there exists $\xi \in L$ so that*

$$f(a+h) = \sum_{\ell=0}^{k} \tfrac{1}{\ell!} D_h^\ell f(a) + \tfrac{1}{(k+1)!} D_h^{k+1} f(\xi).$$

Proof. Our best bet will be to try to deduce this from the one-dimensional version. To this end, we let

$$\varphi(t) = a + th, \quad \varphi : [0,1] \to \mathbb{R}^n,$$

and set

$$g(t) = f(\varphi(t)) = f(a+th).$$

By the Lagrange Remainder Theorem in one dimension (and the fact that $g(0) = f(a)$ and $g(1) = f(a+h)$), we may obtain

$$f(a+h) = f(a) + \sum_{\ell=1}^{k} \tfrac{1}{\ell!} g^{(\ell)}(0) + \tfrac{1}{(k+1)!} g^{(k+1)}(c) \quad \text{for some} \quad c \in (0,1).$$

Then the result will follow provided we can establish the identity

$$g^{(\ell)}(t) = D_h^\ell f(a+th), \quad 1 \leq \ell \leq k+1. \tag{7.9}$$

In this case, the point $\xi \in L$ is given by $\xi = \varphi(c)$. Note that establishing (7.9) above also shows that we can actually apply the Lagrange Remainder Theorem to g (because it shows that g is $k+1$-times differentiable).

We begin by observing that by the chain rule,

$$g'(t) = \nabla f(\varphi(t)) \cdot \varphi'(t) = \nabla f(a+th) \cdot h = D_h f(a+th),$$

giving (7.9) in the case $\ell = 1$. Suppose now that (7.9) holds up to level ℓ. Then, writing

$$F = D_h^\ell f, \quad \text{so that} \quad g^{(\ell)} = F \circ \varphi,$$

we have

$$\begin{aligned} g^{(\ell+1)}(t) = \tfrac{d}{dt}[g^{(\ell)}] &= \nabla F(\varphi(t)) \cdot \varphi'(t) \\ &= \nabla F(a+th) \cdot h = D_h F(a+th) = D_h^{\ell+1} f(a+th), \end{aligned}$$

which yields the result. □

This higher-dimensional version of the Lagrange Remainder Theorem is a bit unsatisfactory as written. Instead of seeing multiple applications of the directional derivative, we may instead prefer to see a theorem that involves only the partial derivatives of the function f. To do this, we recall that

$$D_h f(a) = \nabla f(a) \cdot h = \sum_{i=1}^{n} h_i D_i f(a).$$

We can write this as an 'operator identity', namely

$$D_h = \sum_{i=1}^{n} h_i D_i = h_1 D_1 + \cdots + h_n D_n.$$

Thus

$$D_h^\ell = (h_1 D_1 + \cdots + h_n D_n)^\ell. \tag{7.10}$$

To proceed, we can expand out this final expression using the so-called *multinomial formula*

$$(c_1 + \cdots + c_n)^\ell = \sum_{|\alpha|=\ell} \binom{\ell}{\alpha} c^\alpha,$$

which contains some new notation we need to explain:

- Here $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a *multiindex*, which is a vector whose entries are nonnegative integers. The quantity $|\alpha|$ equals $\alpha_1 + \cdots + \alpha_n$. The sum above is meant to be taken over all multiindices α with $|\alpha| = \ell$. Given n and ℓ, can you figure out how many such multiindices there are?
- The *multinomial coefficient* $\binom{\ell}{\alpha}$ is given by
$$\binom{\ell}{\alpha} = \frac{\ell!}{\alpha!} := \frac{\ell!}{\alpha_1! \cdots \alpha_n!}, \quad \text{where} \quad |\alpha| = \ell.$$
More generally we have the binomial coefficient
$$\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha - \beta)!},$$
where we recall that for a multiindex $\alpha = (\alpha_1, \ldots, \alpha_n)$, the factorial is given by $\alpha! = \alpha_1! \cdots \alpha_n!$.
- The notation c^α means
$$c^\alpha = c_1^{\alpha_1} \cdots c_n^{\alpha_n}.$$

Example 7.13. This may be new notation, so let's look at a concrete example, namely, the expansion of

$$(c_1 + c_2 + c_3)^3.$$

This will be a sum of terms of the form

$$c_1^{\alpha_1} c_2^{\alpha_2} c_3^{\alpha_3},$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 3$. In particular, there are 10 possible multiindices, given by

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

and the corresponding coefficients are

$$1, 1, 1, 3, 3, 3, 3, 3, 3, 6,$$

respectively.

Continuing from (7.10), we can write

$$\begin{aligned} D_h^\ell f(a) &= \sum_{|\alpha|=\ell} \binom{\ell}{\alpha} h_1^{\alpha_1} \cdots h_n^{\alpha_n} D_1^{\alpha_1} \cdots D_n^{\alpha_n} f \\ &= \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} h^\alpha D^\alpha f, \end{aligned}$$

where we have introduced some fancy new compact notation in the final line, namely,

$$\alpha = (\alpha_1, \ldots, \alpha_n) \implies D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}.$$

With this notation, we have the following re-statement of Taylor's formula: for $h \in \mathbb{R}^n$,

$$f(a + h) = P_k(h) + R_k(h), \tag{7.11}$$

where

$$P_k(h) = \sum_{\ell=0}^{k} \sum_{|\alpha|=\ell} \tfrac{1}{\alpha!} h^\alpha D^\alpha f(a) = \sum_{|\alpha|\leq k} \tfrac{1}{\alpha!} h^\alpha D^\alpha f(a)$$

is the degree k *Taylor polynomial* and the remainder $R_k(h)$ is given by $\frac{1}{(k+1)!} D_h^{k+1}(\xi)$ for some ξ on the line segment joining a and $a+h$.

Example 7.14. Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = e^{x_1 + \cdots + x_n}$. The Taylor polynomial for f can be obtained by simply writing the one-dimensional Taylor polynomial and using $x_1 + \cdots + x_n$ for the variable. However, we can also compute it as follows: as $D^\alpha f(0) = 1$ for an arbitrary multiindex α, we have

$$P_k(x) = \sum_{\ell=0}^{k} \sum_{|\alpha|=\ell} \frac{1}{\alpha!} x^\alpha = \sum_{\ell=0}^{k} \frac{1}{\ell!} \sum_{|\alpha|=\ell} \binom{\ell}{\alpha} x^\alpha = \sum_{\ell=0}^{k} \frac{1}{\ell!} [x_1 + \cdots + x_n]^\ell,$$

where in the final step we have used the multinomial formula.

We have the following theorem (whose proof we skip) concerning the size of the error term $R_k(h)$ and the uniqueness of Taylor polynomials.

Theorem 7.13. *If $f \in C^{k+1}$ and $R_k(h)$ is the k^{th} degree remainder for f at a, then*

$$\lim_{h\to 0} \frac{R_k(h)}{|h|^k} = 0.$$

In fact, if Q is any polynomial such that

$$\lim_{x\to a} \frac{f(x) - Q(x-a)}{|x-a|^k} = 0,$$

then Q is the k^{th} degree Taylor polynomial of f at a.

This theorem is useful because it gives us a way to determine Taylor polynomials without necessarily having to compute a lot of derivatives:

Example 7.15. Let us determine the third degree Taylor polynomial of $f(x) = e^x \sin x$ at $x = 0$. We write

$$e^x = P(x) + R(x) \quad \text{and} \quad \sin(x) = \tilde{P}(x) + \tilde{R}(x),$$

where

$$P(x) = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{6}x^3 \quad \text{and} \quad \tilde{P}(x) = x - \tfrac{1}{6}x^3.$$

We multiply these two expressions and only keep up to the cubic terms. This yields

$$e^x \sin x = x + x^2 + \tfrac{1}{3}x^3 + R^*(x),$$

where

$$R^*(x) = -\tfrac{1}{2}x^5 - \tfrac{1}{36}x^6 + R(x)\tilde{P}(x) + \tilde{R}(x)P(x) + R(x)\tilde{R}(x).$$

Since we have (by the previous theorem)

$$\lim_{x\to 0} \tfrac{R^*(x)}{x^3} = 0,$$

it follows (again by the previous theorem) that $x + x^2 + \frac{1}{3}x^3$ is the degree three Taylor polynomial for $e^x \sin x$.

With the basic theory of Taylor polynomials in place, we turn to our key application, namely, the classification of critical points.

Definition 7.5. Suppose $U \subset \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is differentiable. We call $a \in U$ a *critical point* of f if $\nabla f(a) = 0$.

The significance of critical points lies in the fact that if f has a maximum or minimum at a point a, then a must be a critical point for f. For example, the function $g(x) = f(x, a_2, \ldots, a_n)$ would have an extreme point at $x = a_1$ and hence we would have $g'(a_1) = 0$ (a fact we know well from calculus in one dimension). But $g'(a_1) = D_1 f(a)$. Similarly, all partial derivatives of f would have to vanish.

Suppose now that $f \in C^3(U)$ has a critical point at $a \in U$. Then the Taylor Formula (or Lagrange Remainder Theorem) for f takes the form

$$f(a + h) = f(a) + q(h) + R(h),$$

where

$$q(h) = \tfrac{1}{2} D_h^2 f(a) \quad \text{and} \quad \lim_{h \to 0} \frac{R(h)}{|h|^2} = 0. \tag{7.12}$$

Let's open up the definition of $q(h)$ to see what it really looks like. This is important, so let us state it as a lemma.

Lemma 7.14. *With f, q as above, we may express*

$$q(x) = \tfrac{1}{2} x \cdot Hx \quad \textit{for} \quad x \in \mathbb{R}^n,$$

where H is the $n \times n$ Hessian matrix *with entries given by*

$$H_{jk} = D_j D_k f(a).$$

Proof. This is ultimately just a rephrasing of computations we have done before, but let's see it anyway. We have that

$$\begin{aligned} q(x) &= \tfrac{1}{2}(x_1 D_1 + \cdots + x_n D_n)^2 f(a) \\ &= \tfrac{1}{2} \sum_{j=1}^n \sum_{k=1}^n D_j D_k f(a) x_j x_k \\ &= \tfrac{1}{2} \sum_{k=1}^n x_k [Hx]_k = \tfrac{1}{2} x \cdot Hx. \end{aligned}$$

□

This lemma shows that the quantity $q(h)$ appearing in the Taylor expansion for f has the form of a 'quadratic form' on $\mathbb{R}^n$:

Definition 7.6. A *quadratic form* on $\mathbb{R}^n$ is a function of the form

$$F(x) = x \cdot Ax$$

for some symmetric matrix A. (Here symmetric means $A_{jk} = A_{kj}$).

In our case, the symmetry comes from the fact that $H_{jk} = D_j D_k f(a)$. Since we are working with $f \in C^3$, we are guaranteed $H_{jk} = H_{kj}$ by Clairaut's Theorem.

Now let's recall what led us here: we want to try to classify critical points (as maxima, minima, etc.). The Taylor expansion above suggests that what we really need to understand is the behavior (i.e. positivity or negativity) of the quadratic form $x \mapsto x \cdot Hx$, where H is the Hessian matrix of f at a.

Definition 7.7. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form on $\mathbb{R}^n$. We call F:

- *positive definite* if $F(x) > 0$ for all $x \neq 0$,

- *negative definite* if $F(x) < 0$ for all $x \neq 0$,
- *nondefinite* otherwise.

We can then prove the following result:

Theorem 7.15. *Suppose $f \in C^3(U)$ for some open set U containing a critical point a. Let $q(x) = \frac{1}{2}x \cdot Hx$, where H is the Hessian matrix of f at a. Then f has:*

(a) *a local minimum if q is positive definite,*
(b) *a local maximum if q is negative definite,*
(c) *neither if q is nondefinite.*

Proof. Let's prove (a) in detail. We need to prove that there exists $\delta > 0$ so that

$$0 < |h| < \delta \implies f(a+h) > f(a).$$

Using the Taylor expansion, we see that the condition $f(a+h) > f(a)$ is equivalent to the requirement $q(h) + R(h) > 0$.

Dividing by the nonnegative quantity $|h|^2$ and using the form of $q(\cdot)$, we see that it suffices to find $\delta > 0$ so that

$$0 < |h| < \delta \implies q(\tfrac{h}{|h|}) + \tfrac{R(h)}{|h|^2} > 0.$$

Now observe that for $h \neq 0$, we have

$$\tfrac{h}{|h|} \in K := \{x \in \mathbb{R}^n : |x| = 1\}.$$

As q is continuous, it attains a minimum value, say m, on the compact set K. As q is positive definite, we must have $m > 0$. On the other hand, using (7.12), we can find $\delta > 0$ so that

$$0 < |h| < \delta \implies \big|\tfrac{R(h)}{|h|^2}\big| < \tfrac{1}{2}m,$$

so that

$$0 < |h| < \delta \implies q(\tfrac{h}{|h|}) + \tfrac{R(h)}{|h|^2} > \tfrac{1}{2}m > 0,$$

as desired.

The proof of (b) is similar—you should work out the details!

Finally, for (c), we choose $h_1, h_2 \in \mathbb{R}^n$ so that $q(h_1) > 0$ and $q(h_2) < 0$. Then we can write

$$f(a + th_i) - f(a) = q(th_i) + R(th_i) = t^2[q(h_i) + |h_i|^2 \tfrac{R(th_i)}{|th_i|^2}]$$

for any $t \neq 0$. In particular, for t sufficiently small, we have

$$f(a + th_1) > f(a) \quad \text{and} \quad f(a + th_2) < f(a),$$

showing that f has neither a maximum nor a minimum at a. □

Now, Theorem 7.15 will only be useful to us if we actually have some techniques to determine whether the quadratic form q is positive or negative. This is ultimately a linear algebra problem. Let us quickly review the main we will need:

Theorem 7.16 (Diagonalization of symmetric matrices)**.** *Suppose A is a symmetric $n \times n$ matrix (that is, $A_{jk} = A_{kj}$ for all j, k). Then there exists an orthonormal basis $\{v_1, \dots, v_n\}$ and real numbers $\lambda_1, \dots, \lambda_n$ so that*

$$A = PDP^{-1}, \quad \textit{with} \quad P = [v_1 \cdots v_n] \quad \textit{and} \quad D = diag(\lambda_1, \dots, \lambda_n).$$

In particular, $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A and $v_1, \dots, v_n$ are corresponding eigenvectors.

We won't prove this here. You should have covered it in your linear algebra course!

Corollary 7.17. *The quadratic form $F(x) = x \cdot Ax$ is positive definite if and only if all of the eigenvalues of A are positive. It is negative definite if and only if all of the eigenvalues of A are negative.*

Proof. Write $A = PDP^{-1}$ as in Theorem 7.16. In fact, since the columns of P form an orthonormal set, we have $P^{-1} = P^T$ (the transpose of P). Then

$$F(x) = x \cdot Ax = x \cdot PDP^T x = [P^T x] \cdot DP^T x = \sum_{j=1}^{n} \lambda_j y_j^2,$$

where $y = P^T x$. In particular, if F is positive definite, then by choosing $x = Pe_j$ (so that $y = e_j$) we find that $\lambda_j > 0$ for each j. Conversely, if each $\lambda_j > 0$ and $x \neq 0$, then $y = P^T x \neq 0$ and hence $y_k^2 > 0$ for at least one k, so that $F(x) > 0$. A similar argument deals with the negative definite case. □

This characterization in terms of eigenvalues is very handy, since we have a systematic way to determine the eigenvalues of a matrix:

Lemma 7.18. *The eigenvalues of an $n \times n$ matrix A are the roots of the characteristic polynomial*

$$p(\lambda) := \det[A - \lambda I],$$

where I is the $n \times n$ identity matrix.

Proof. This is another standard linear algebra fact, so we will skip it. □

Let's see what all of this new technology can do for us.

Example 7.16. Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ has a critical point at a, and the quadratic form of f at a is

$$q(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 + 4x_2x_3.$$

The corresponding matrix for q is

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix},$$

which has characteristic polynomial

$$p(\lambda) = (1 - \lambda)[(1 - \lambda)^2 - 4].$$

The roots of this polynomial are $-1, 1, 3$. Thus f has neither a maximum nor a minimum at a. (This basically means that there are two directions in which the value of f moves up and one direction in which the value of f moves down.)

In the case of real-valued functions of two variables, the situation becomes relatively simple. To see this, we will use the following linear algebra lemma:

Lemma 7.19. *Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then*

$$\det A = \prod_{j=1}^{n} \lambda_j \quad \textit{and} \quad \mathrm{tr} A = \sum_{j=1}^{n} \lambda_j,$$

where $\mathrm{tr}A = \sum_{j=1}^{n} A_{jj}$ is the sum of the diagonal entries of A (called the trace *of A).*

Proof. These are more linear algebra facts. Let's prove it in the special case that A is diagonalizable (which will always be the case when A is the Hessian matrix of a function $f \in C^3$). In this case, we write $A = PDP^{-1}$, where D is the diagonal matrix that necessarily consists of the eigenvalues of A. We now need a few other linear algebra facts, namely,

$$\det(AB) = \det(A) \cdot \det(B), \quad \det(A^{-1}) = [\det A]^{-1}, \quad \text{and} \quad \operatorname{tr}(AB) = \operatorname{tr}(BA),$$

all of which you are encouraged to prove on your own if you have not seen before.

In particular, we find that

$$\det(A) = \det(PDP^{-1}) = \det(P) \cdot \det(D) \cdot \det(P^{-1}).$$

Now, since $\det(P^{-1}) = [\det(P)]^{-1}$, this implies $\det(A) = \det(D)$. But D is diagonal, so its determinant is the product of its entries, which in this case are the eigenvalues of A.

Next,

$$\operatorname{tr}(A) = \operatorname{tr}(PDP^{-1}) = \operatorname{tr}(P^{-1}PD) = \operatorname{tr}(D),$$

and this final quantity equals the sum of the diagonal entries of D, which in this case are the eigenvalues of A. □

We can now prove the following result about critical points for functions of two variables (which you might recall from your multivariable calculus course):

Theorem 7.20 (Classification of critical points, two-dimensional case)**.** *Let $U \subset \mathbb{R}^2$, and suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is a differentiable function with a critical point at $a \in U$. Define*

$$D = f_{xx}(a)f_{yy}(a) - [f_{xy}(a)]^2$$

Then:

- *If $D > 0$ and $f_{xx}(a) > 0$, then f has a local minimum at $x = a$.*
- *If $D > 0$ and $f_{xx}(a) < 0$, then f has a local maximum at $x = a$.*
- *If $D < 0$, then f has a saddle point at $x = a$.*
- *If $D = 0$, we can't say anything.*

Proof. The quantity D here is nothing but $\det(H)$ in disguise, which is essentially the matrix of the quadratic form for f.

If $D > 0$, then the product of the eigenvalues of H is positive. That means they are either both positive or both negative. If $f_{xx}(a) > 0$, then (again since $D > 0$) we must have $f_{yy}(a) > 0$. But then the trace of H is positive, and so both eigenvalues must be positive, so that f has a local minimum. If $f_{xx}(a) < 0$ then (again since $D > 0$) we must have $f_{yy}(a) < 0$, so that the trace is negative, and so both eigenvalues are negative, which implies that f has a local maximum. If $D < 0$, then the product of eigenvalues is negative, so they must have opposite signs. Then H has one positive eigenvalue and one negative eigenvalue; in the two-dimensional case, this corresponds to saddle shape around the point a. □

You'll work a few more examples in the homework problems.

In the preceding sections, we found a method to look for local maxima and minima for real-valued functions on $\mathbb{R}^n$. In particular, we look for critical points and study the quadratic form (i.e. the Hessian matrix) at these points. Using information about the eigenvalues of this matrix, we can determine whether the critical points are maxima, minima, or neither.

Our next main goal is to study the problem of *constrained* optimization. That is, we still want to look for the maximum/minimum values of a function; however, we now search for these values under some given *constraints* on the inputs. This is an important modification to the above problem: On the one hand, this type of constrained minimization problem shows up frequently in applications. On the other hand, understanding this problem requires a significant amount of new mathematics. So, off we go!

7.4. **The Implicit Function Theorem and Inverse Function Theorem.** In this section, we discuss two important theorems known as the *implicit function theorem* and *inverse function theorem.* They involve two very fundamental questions:

(i) Given a function $G(x, y)$, can we solve the equation $G(x, y) = 0$ for y as a function of x?

(ii) Given an equation $f(x) = y$, can we solve for x as a function of y?

These are the questions that will be answered by the two theorems mentioned above. The key to establishing them both will be to establish yet *another* theorem, namely, the *contraction mapping theorem* (and an upgraded version thereof). In fact, this is an important result in its own right! Although we will focus on the familiar setting of finite-dimensional Euclidean space, it is worth mentioning that the results in this section can be generalized significantly (e.g. to the 'Banach space' setting). In fact, the proofs here are presented in such a way that they carry over nearly verbatim, once you understand what Banach spaces are and what differentiation in Banach spaces means.

We begin with the following definition:

Definition 7.8. Let $U \subset \mathbb{R}^n$. A function $T : U \to \mathbb{R}^m$ is called a *contraction* if there exists $\alpha \in (0, 1)$ so that

$$|T(x) - T(y)| \leq \alpha|x - y| \quad \text{for all} \quad x, y \in U.$$

Theorem 7.21 (Contraction Mapping Principle)**.** *Suppose $U \subset \mathbb{R}^n$ is closed and $T : U \to U$ is a contraction. Then T has a unique fixed point. That is, there exists unique $x \in U$ such that $T(x) = x$.*

Proof. Let $x_0 \in U$, and define $x_k \in U$ inductively via $x_{k+1} = T(x_k)$. It follows that

$$|x_2 - x_1| = |T(x_1) - T(x_0)| \leq \alpha|x_1 - x_0|$$

for some $\alpha \in (0, 1)$; similarly,

$$|x_3 - x_2| = |T(x_2) - T(x_1)| \leq \alpha^2|x_1 - x_0|,$$

and more generally

$$|x_{k+1} - x_k| \leq \alpha^k|x_1 - x_0|.$$

Thus, for any $m > n$, we have by the triangle inequality that

$$|x_m - x_n| \leq \sum_{k=n}^{m-1} |x_{k+1} - x_k| \leq |x_1 - x_0| \sum_{k=n}^{m-1} \alpha^k \leq \tfrac{\alpha^n|x_1 - x_0|}{1-\alpha}.$$

Since $\alpha \in (0, 1)$, it follows that $\{x_n\}$ is a Cauchy sequence and hence has a limit $x \in U$. As every contraction is necessarily continuous (check!), it follows that $T(x_n) \to T(x)$. However, since $T(x_n) = x_{n+1}$, we also have that $T(x_n) \to x$. Thus $T(x) = x$. Finally, we note that if $T(x) = x$ and $T(y) = y$, then

$$|x - y| = |T(x) - T(y)| \leq \alpha|x - y|.$$

Since $\alpha < 1$, this implies $x = y$. □

We will upgrade this result considerably via Lemma 7.23 and Theorem 7.24 below. These results will really constitute the 'hard work' of this section.

We first define the following:

Definition 7.9. Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$. A function $T : U \times V \to \mathbb{R}^m$ is a *uniform contraction* if there exists $\alpha \in (0, 1)$ so that

$$|T(x, y_1) - T(x, y_2)| \leq \alpha |y_1 - y_2| \quad \text{for all} \quad x \in U, \quad y_1, y_2 \in V.$$

Then we have the following corollary to the contraction mapping principle.

Corollary 7.22. *Suppose $U \subset \mathbb{R}^n$ is open and $V \subset \mathbb{R}^m$ is closed. If $T : U \times V \to V$ is a uniform contraction, then there exists a fixed point function $g : U \to V$ such that $T(x, g(x)) = g(x)$.*

Proof. Apply the contraction mapping principle to each function F_x defined by $F_x(y) = T(x, y)$. □

The key result we will prove below is that the fixed point function g 'inherits the regularity' of the contraction T. To make all of this precise, let us first introduce a bit of notation. We will continue to look at functions $T : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$. We will write elements of $\mathbb{R}^n \times \mathbb{R}^m$ as (x, y). That is, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. If T is differentiable at a point (x, y), then its derivative will be an $m \times (n+m)$ matrix. As usual, we can make sense of its partial derivatives $D_j T(x, y)$, each of which will be a column of the matrix $T'(x, y)$. We write $D_x T(x, y)$ to denote the first n columns of this matrix (so it is an $m \times n$ matrix), and $D_y T(x, y)$ to denote the last m columns of this matrix (so it is an $m \times m$ matrix).

Throughout the proofs below, we will use the following fact several times: if A is an $n \times n$ matrix, then there exists $C > 0$ such that

$$|Ax| \leq C|x| \quad \text{for all} \quad x \in \mathbb{R}^n, \tag{7.13}$$

which is left as an exercise.

As mentioned above, we will show that the fixed point function g inherits the regularity of the contraction T. In particular, we will prove that if T is C^1, then so is g. To motivate what follows, let us see what form the derivative of g should take (assuming it is differentiable). We start by differentiating the identity $T(x, g(x)) = g(x)$, which yields

$$D_x g = D_x T(x, g(x)) + D_y T(x, g(x)) D_x g.$$

Rearranging, we find

$$(I - D_y T(x, g(x)) D_x g = D_x T(x, g(x)),$$

and hence, if $I - D_y T(x, g(x))$ is invertible, we should have

$$D_x g = (I - D_y T(x, g(x)))^{-1} D_x T(x, g(x)).$$

We will therefore need the following result:

Lemma 7.23. *Suppose $T : U \times V \to V$ is a uniform contraction (with constant α), where $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$. Let $g : U \to V$ be the fixed point function of T.*

If $T \in C^1(U \times V)$, then the $m \times m$ matrix

$$A(x) = I - D_y T(x, g(x))$$

is invertible for all $x \in U$, where I is the $m \times m$ identity matrix. Furthermore,

$$|A^{-1}(x)z| \leq \tfrac{1}{1-\alpha}|z| \quad \textit{for all} \quad x \in U,\ z \in \mathbb{R}^m. \tag{7.14}$$

Proof. The key is going to be the following bound:

$$|D_yT(x,y)z| \leq \alpha|z| \quad \text{for all} \quad x \in U,\ y \in V,\ z \in \mathbb{R}^m. \tag{7.15}$$

Let's prove this, and then see what it does for us.

Fix $x \in U$, $y \in V$, and $z \in \mathbb{R}^m \backslash \{0\}$, and let δ_n be a sequence of positive real numbers satisfying $\delta_n \to 0$. Then, by definition of differentiability ,we have

$$\lim_{n\to\infty}\left|D_yT(x,y)[\tfrac{\delta_n z}{|\delta_n z|}] - \tfrac{1}{|\delta_n z|}\big[T(x,y+\delta_n z) - T(x,y)\big]\right| = 0.$$

Now observe that by the uniform contraction property,

$$\big|\tfrac{1}{|\delta_n z|}[T(x,y+\delta_n z) - T(x,y)]\big| \leq \alpha\tfrac{|\delta_n z|}{|\delta_n z|} \leq \alpha.$$

Combining the previous two displays and using the linearity of $D_yT(x,y)$, we deduce

$$|D_yT(x,y)\tfrac{z}{|z|}| \leq \alpha, \quad \text{i.e.} \quad |D_yT(x,y)z| \leq \alpha|z|,$$

which is (7.15).

Now let's try to invert the matrix $A(x)$ defined above. The idea is that since A is of the form '$1 - R$', its inverse should be '$\frac{1}{1-R}$'. This final quantity doesn't make any sense in the present setting (since '1' and 'R' are matrices). However, we can also recognize $\frac{1}{1-R}$ as the result of summing the geometric series $\sum_{k=0}^{\infty} R^k$, and these summands actually *do* make sense, since they are just powers of a square matrix. So, here are the claims we need to prove:

(i) For any $x \in U$, we can define a linear transformation $L : \mathbb{R}^m \to \mathbb{R}^m$ by setting

$$L(z) = \sum_{k=0}^{\infty} R^k z = \lim_{n\to\infty}\sum_{k=0}^{n} R^k z, \quad \text{where} \quad R := D_yT(x,g(x)).$$

(ii) We have $L(Az) = z$ for all $z \in \mathbb{R}^m$, where $A = I - R$. Consequently, if B is the matrix for L, then $B = A^{-1}$.

We first prove (i). Given $z \in \mathbb{R}^m$, we need to show that the sequence

$$w_\ell := \sum_{k=0}^{\ell} R^k z \in \mathbb{R}^m$$

converges. To do this, we'll show it's Cauchy. In fact, this is just like the proof of the contraction mapping theorem: We first recall that $R = D_yT(x,g(x))$, so that by (7.15) we have

$$|Rz| \leq \alpha|z|, \quad |R^2 z| \leq \alpha|Rz| \leq \alpha^2|z|, \quad \text{and in general} \quad |R^k z| \leq \alpha^k|z|.$$

Thus for $\ell > j$,

$$|w_\ell - w_j| \leq \sum_{k=j+1}^{\ell} |R^k z| \leq |z| \sum_{k=j+1}^{\ell} \alpha^k \leq |z|\tfrac{\alpha^\ell}{1-\alpha}.$$

It follows that $\{w_\ell\}$ is Cauchy and hence has a limit, which we define to be $L(z)$ and denote by $\sum_{k=0}^{\infty} R^k z$.

Now that we have defined $L(\cdot)$, let's make sure it is a linear transformation. First, for $c \in \mathbb{R}$ and $\ell \geq 0$, we have

$$\sum_{k=0}^{\ell} R^k[cz] = c\sum_{k=0}^{\ell} R^k z, \quad \text{which implies} \quad L(cz) = cL(z).$$

Similarly

$$\sum_{k=0}^{\ell} R^k[z_1 + z_2] \equiv \sum_{k=0}^{\ell} R^k z_1 + \sum_{k=0}^{n} R^k z_2 \implies L(z_1 + z_2) = L(z_1) + L(z_2).$$

This finishes the proof of (i).

For (ii), we need to show that $L((I - R)z) = z$ for all $z \in \mathbb{R}^m$. To this end, note that

$$\sum_{k=0}^{\ell} R^k(1 - R)z = \sum_{k=0}^{\ell} R^k z - \sum_{k=1}^{\ell+1} R^k z = z - R^{\ell+1}z. \tag{7.16}$$

Thus, since

$$|R^{\ell+1}z| \leq \alpha^{\ell+1}|z| \to 0 \quad \text{as} \quad \ell \to \infty,$$

we deduce $L((1 - R)z) = z$ as desired. I'll leave it as an exercise to check that this implies that the matrix for L is the inverse of $(1 - R)$.

The last thing to prove is the bound

$$|A^{-1}(x)z| \leq \tfrac{1}{1-\alpha}|z|.$$

Since we just showed that the matrix for L is A^{-1}, this is equivalent to showing that

$$|L(z)| \leq \tfrac{1}{1-\alpha}|z|. \tag{7.17}$$

To see this, fix $z \in \mathbb{R}^m$ and let $L_\ell(z) = \sum_{k=0}^{\ell} R^k z$, which we know converges to $L(z)$ as $\ell \to \infty$. Then (7.16) can be rewritten

$$L_\ell(z) - RL_\ell(z) = z - R^{\ell+1}z, \quad \text{or} \quad L_\ell(z) = z + RL_\ell(z) - R^{\ell+1}z.$$

In particular,

$$|L_\ell(z)| \leq |z| + |RL_\ell(z)| + |R^{\ell+1}z| \leq |z| + \alpha|L_\ell(z)| + \alpha^{\ell+1}|z|.$$

Rearranging this gives

$$(1 - \alpha)|L_\ell(z)| \leq |z| + \alpha^{\ell+1}|z|, \quad \text{or} \quad |L_\ell(z)| \leq \tfrac{1}{1-\alpha}|z| + \tfrac{\alpha^{\ell+1}}{1-\alpha}|z|.$$

We now take the limit as $\ell \to \infty$ to obtain (7.17). □

We then have the following upgrade of the contraction mapping principle, which shows that the fixed point function inherits the regularity of the original mapping.

Theorem 7.24 (Uniform Contraction Mapping Principle)**.** *Let $T : U \times V \to V$ be a uniform contraction, where $U \subset \mathbb{R}^n$ is open and $V \subset \mathbb{R}^m$ is closed, and let $g : U \to V$ be the corresponding fixed point function.*

Let $k \in \{0, 1\}$. If $T \in C^k(U \times V)$, then $g \in C^k(U)$.

Proof. We begin with the case $k = 0$. By definition, we have

$$\begin{aligned}|g(x+h) - g(x)| &= |T(x+h, g(x+h)) - T(x, g(x))| \\ &\le |T(x+h, g(x+h)) \\ &\quad - T(x+h, g(x))| + |T(x+h, g(x)) - T(x, g(x))| \\ &\le \alpha|g(x+h) - g(x)| + |T(x+h, g(x)) - T(x, g(x))|\end{aligned}$$

for some $\alpha \in (0, 1)$. Thus

$$|g(x+h) - g(x)| \le \tfrac{1}{1-\alpha}|T(x+h, g(x)) - T(x, g(x))|.$$

Since T is assumed to be continuous we deduce

$$\lim_{h\to 0} g(x+h) = g(x),$$

yielding continuity of g.

Next, suppose $T \in C^1$. Then, using Lemma 7.23, we can define the $m\times n$ matrix

$$M(x) = [I - D_yT(x, g(x))]^{-1}D_xT(x, g(x)),$$

which depends continuously on $x \in U$ (this depends on (7.14)). We will show that $g \in C^1(U)$ by showing that $g'(x) = M(x)$. To do so, we define

$$Rg(x, h) = g(x+h) - g(x) - M(x)h \quad \text{for} \quad x \in U, \quad h \in \mathbb{R}^n,$$

and we will show that given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$0 < |h| < \delta \implies |Rg(x, h)| < \varepsilon|h|.$$

(As we saw above, you can arrive at this as the right 'guess' for $g'(x)$ by just differentiating the equation $T(x, g(x)) = g(x)$ and solving for $g'(x)$.)

To fit the computations in the margins, we need to introduce some notation. In particular, we set

$$\Delta g = \Delta g(x, h) := g(x+h) - g(x) \in \mathbb{R}^m$$

(so that $\Delta g = Rg + Mh$) and $\mathcal{E}(p, q) \in \mathbb{R}^m$ by

$$\begin{aligned}\mathcal{E}(p, q) &= \mathcal{E}((x, g(x)), (p, q)) \\ &= T(x+p, y+q) - T(x, y) - D_xT(x, y)p - D_yT(x, y)q\end{aligned}$$

where $(p, q) \in \mathbb{R}^n \times \mathbb{R}^m$.

We then have

$$\begin{aligned}\Delta g &= T(x+h, g(x+h)) - T(x, g(x)) \\ &= T(x+h, g(x) + \Delta g) - T(x, g(x)) \\ &= D_xT(x, g(x))h + D_yT(x, g(x))\Delta g + \mathcal{E}(h, \Delta g)\end{aligned}$$

Rearranging this, we get

$$\begin{aligned}\Delta g &= [I - D_yT(x, g(x)]^{-1}D_xT(x, g(x))h + [I - D_yT(x, g(x))]^{-1}\mathcal{E}(h, \Delta g) \\ &= M(x)h + [I - D_yT(x, g(x))]^{-1}\mathcal{E}(h, \Delta g),\end{aligned}$$

and so

$$Rg(x, h) = [I - D_yT(x, g(x))]^{-1}\mathcal{E}(h, \Delta g). \tag{7.18}$$

Now let $\eta > 0$ be a small parameter to be chosen below. Since $T \in C^1(U \times V)$, we can find $\mu > 0$ so that

$$|p| + |q| < \mu \implies |\mathcal{E}(p, q)| < \eta(|p| + |q|).$$

Next, since $\Delta g(x,h)$ is continuous and $\Delta g(x,0)=0$, we can find $0<\delta<\frac{1}{2}\mu$ so that

$$0<|h|<\delta \implies |\Delta g(x,h)|<\tfrac{1}{2}\mu.$$

In particular,

$$\begin{aligned} 0<|h|<\delta &\implies |h|+|\Delta g|<\mu \\ &\implies |\mathcal{E}(h,\Delta g)|<\eta(|h|+|\Delta g(x,h)|). \end{aligned}$$

Using Lemma 7.23 again and (7.18), we deduce

$$\begin{aligned} &0<|h|<\delta \\ &\quad\implies |Rg(x,h)|<\tfrac{\eta}{1-\alpha}\big[|h|+|\Delta g(x,h)|\big]<\tfrac{\eta}{1-\alpha}\big[|h|+|Rg(x,h)|+|M(x)h|\big]. \end{aligned}$$

Now, as long as η is small enough so that $\frac{\eta}{1-\alpha}<\frac{1}{2}$, we can rearrange this to deduce

$$0<|h|<\delta \implies |Rg(x,h)|<\tfrac{2\eta}{1-\alpha}\big[|h|+|M(x)h|\big].$$

But now $M(x)$ is some matrix, so that $|M(x)h|\leq C|h|$ for some $C>0$. In particular,

$$0<|h|<\delta \implies |Rg(x,h)|<\tfrac{2\eta(1+C)}{1-\alpha}|h|,$$

and by choosing η small enough depending on ε, C, α we can guarantee that the right-hand side is bounded by $\varepsilon|h|$. This completes the proof! □

Finally, we can state and prove the implicit function theorem:

Theorem 7.25 (Implicit Function Theorem)**.** *Let $\mathcal{O}\subset\mathbb{R}^n\times\mathbb{R}^m$ be open. Let $F:\mathcal{O}\to\mathbb{R}^m$ satisfy $F\in C^1(\mathcal{O})$. If $(x_0,y_0)\in\mathcal{O}$ is such that*

$$F(x_0,y_0)=0 \quad \textit{and} \quad D_yF(x_0,y_0) \quad \textit{is invertible,}$$

then there exists an open set $(x_0,y_0)\in U\times V\subset\mathcal{O}$ and $g:U\to V$ satisfying $g\in C^1(U)$ such that

$$y_0=g(x_0) \quad \textit{and} \quad F(x,g(x))=0 \quad \textit{for all} \quad x\in U.$$

Moreover, if $(x,y)\in U\times V$ and $F(x,y)=0$, then $y=g(x)$.

Proof. For convenience (and without loss of generality), we assume $(x_0,y_0)=(0,0)$. We let $L=[D_yF(0,0)]^{-1}$ and define $G:\mathcal{O}\to\mathbb{R}^m$ by

$$G(x,y)=y-LF(x,y),$$

so that in particular $G\in C^1(\mathcal{O})$. Since L is an invertible matrix, we have that

$$G(x,y)=y \quad \text{if and only if} \quad F(x,y)=0.$$

Note also that

$$G(0,0)=0 \quad \text{and} \quad D_yG(0,0)=I-[D_yF(0,0)]^{-1}D_yF(0,0)=0.$$

Thus, setting $A=D_xG(0,0)$ and using the differentiability of G at $(0,0)$, we may find $\delta_1>0$ so that

$$|x|,|y|<\delta_1 \implies |G(x,y)-Ax|<\tfrac{1}{2}(|x|+|y|).$$

Furthermore, since $D_yG(x,y)$ is continuous, there exists $\delta_2>0$ so that

$$|x|,|y|<\delta_2 \implies |D_yG(x,y)z|<\tfrac{1}{2}|z| \quad \text{for all} \quad z\in\mathbb{R}^m. \tag{7.19}$$

Now set

$$\nu=\min\{\delta_1,\delta_2\}$$

and let $0 < \mu < \nu$ be a small parameter to be determined shortly, and let

$$U = B_\mu(0) \subset \mathbb{R}^n \quad \text{and} \quad V = B_\nu(0) \subset \mathbb{R}^m.$$

It follows that if $(x, y) \in U \times \bar{V}$, then we have

$$\begin{aligned} |G(x,y)| &\leq |Ax| + \tfrac{1}{2}|x| + \tfrac{1}{2}|y| \\ &\leq [C + \tfrac{1}{2}]|x| + \tfrac{1}{2}|y| \\ &\leq [C + \tfrac{1}{2}]\mu + \tfrac{1}{2}\nu < \nu \end{aligned}$$

if μ is chosen sufficiently small. That is,

$$G : U \times \bar{V} \to \bar{V}.$$

In fact, we will now show that G is a uniform contraction. To see this, let $x \in U$ and $y_1, y_2 \in \bar{V}$, and define the function

$$a(\sigma) = G(x, (1-\sigma)y_1 + \sigma y_2).$$

Then we have that $a : [0, 1] \to \bar{V} \subset \mathbb{R}^m$. Moreover, by the chain rule, a is continuously differentiable. Thus, we can apply the fundamental theorem of calculus component by component to obtain the equality

$$a(1) - a(0) = \int_0^1 a'(\sigma)\, d\sigma.$$

However, by definition and the chain rule, this becomes

$$G(x, y_2) - G(x, y_1) = \int_0^1 D_yG(x, (1-\sigma)y_1 + \sigma y_2)[y_2 - y_1]\, d\sigma.$$

Thus, for $x \in U$ and $y_1, y_2 \in \bar{V}$, we can use (7.19) to obtain

$$\begin{aligned} |G(x, y_2) - G(x, y_1)| &\leq \int_0^1 |D_yG(x, (1-\sigma)y_1 + \sigma y_2)[y_2 - y_1]|\, d\sigma \\ &\leq \int_0^1 \tfrac{1}{2}|y_2 - y_1| d\sigma \leq \tfrac{1}{2}|y_2 - y_1|, \end{aligned}$$

showing that G is a uniform contraction. In particular, by Theorem 7.24, there exists $g : U \to \bar{V}$ with $g \in C^1(U)$ such that

$$G(x, g(x)) = g(x) \quad \text{for} \quad x \in U.$$

Moreover, given $x \in U$, we have that $g(x)$ is the *unique* point in $\bar{V}$ such that $G(x, g(x)) = g(x)$. Translating back to F, we have that $F(x, g(x)) = 0$ for all $x \in U$ with the desired regularity and uniqueness for g. □

Using the implicit function theorem, we can also prove the following:

Theorem 7.26 (Inverse Function Theorem)**.** *Let $\mathcal{O} \subset \mathbb{R}^n$ be an open set containing x_0 and let $f : \mathcal{O} \to \mathbb{R}^n$ satisfy $f \in C^1(\mathcal{O})$. If $f'(x_0)$ is invertible, then there exists an open set $x_0 \in U \subset \mathcal{O}$, an open set V containing $f(x_0)$, and a function $g : V \to U$ such that $g \in C^1(U)$, with*

$$g(f(x)) = x \quad \textit{for} \quad x \in U \quad \textit{and} \quad f(g(y)) = y \quad \textit{for} \quad y \in V.$$

Proof. Let us define the function

$$F(y,z) = y - f(z)$$

for y in a small ball V around $f(x_0)$ and z in a small ball U around x_0. Then $F \in C^1$ and we have

$$F(f(x_0), x_0) = f(x_0) - f(x_0) = 0$$

and

$$D_z F(f(x_0), x_0) = -f'(x_0) \quad \text{is invertible,}$$

showing that F satisfies the hypotheses of the implicit function theorem. In particular, there exists $g : V \to U$ with $g \in C^1(V)$ such that

$$0 = F(y, g(y)) = y - f(g(y)), \quad \text{i.e.} \quad f(g(y)) = y.$$

This constructs the desired function g. The remaining properties will be left as an exercise! □

7.5. **Lagrange multipliers.** Recall that our current goal is to study problems in 'constrained optimization'. We have just finished what probably seemed like a long and irrelevant detour towards this goal, but in fact we accomplished something very useful: we proved the implicit and inverse function theorems. They will play a role in this section.

Let us formulate the problem we will consider in this section. We will be given a function $f : \mathbb{R}^n \to \mathbb{R}$ whose maxima/minima we would like to study. However, we will no longer consider all possible $x \in \mathbb{R}^n$ as inputs to f. Instead, we will consider $x \in \mathbb{R}^n$ satisfying a *constraint*, which will be of the form

$$g(x) = 0 \quad \text{for some} \quad g \in C^1(\mathbb{R}^n).$$

Then the condition $\nabla f(x) = 0$ will no longer be the right condition to consider. Instead, as we will see, the condition we will need to consider is the statement that $\nabla f(x) = \lambda \nabla g(x)$ for some $\lambda \in \mathbb{R}$ (which we call the *Lagrange multiplier*).

Here is the key result we need about constraints like the one described above:

Theorem 7.27 (Tangent Planes)**.** *Let $g : \mathbb{R}^n \to \mathbb{R}$ satisfy $g \in C^1(\mathbb{R}^n)$. Define the set*

$$M = \{x \in \mathbb{R}^n : g(x) = 0 \quad \textit{and} \quad \nabla g(x) \neq 0\}.$$

For $a \in M$, define $T_a M$ to be the set of vectors $v \in \mathbb{R}^n$ such that there exists $\delta > 0$ and a differentiable function $\gamma : (-\delta, \delta) \to M$ such that $\gamma(0) = a$ and $\gamma'(0) = v$.

Then for each $a \in M$, $T_a M$ is an $(n-1)$-dimensional vector space (called the tangent plane *to M at a).*

Furthermore, for any $a \in M$ and $v \in T_a M$, we have $\nabla g(a) \cdot v = 0$.

Proof. Let $a \in M$. Then the fact that $\nabla g(a) \neq 0$ implies that $D_i g(a) \neq 0$ for at least one $i = 1, \dots, n$. To simplify the presentation a bit, let us suppose that we have

$$D_n g(a) \neq 0.$$

(To deal with a different value of i, we basically just need to play around with permuting indices.)

Then we can view

$$g(x) = g(x_1, \dots, x_n) = g(\hat{x}, x_n),$$

where $\hat{x} = (x_1, \dots, x_{n-1}) \in \mathbb{R}^{n-1}$, and we have

$$g(\hat{a}, a_n) = 0 \quad \text{and} \quad D_n g(\hat{a}, a_n) \neq 0 \quad \text{(i.e. `is invertible')},$$

where $a = (\hat{a}, a_n)$. By the implicit function theorem, we can find open sets U and V with $\hat{a} \in U$ and $a_n \in V$, and a function $h : U \to V$ with $h \in C^1(U)$ such that

$$a_n = h(\hat{a}) \quad \text{and} \quad g(\hat{x}, h(\hat{x})) = 0 \quad \text{for all} \quad \hat{x} \in U.$$

In fact, by the uniqueness of this function,

$$\{x \in U \times V : g(x) = 0\} = \{x \in \mathbb{R}^n : \hat{x} \in U \quad \text{and} \quad x_n = h(\hat{x})\}.$$

Now define $T_a M$ as in the statement of the theorem. We define the $n-1$ vectors $w_j \in \mathbb{R}^n$ by

$$\begin{aligned} w_1 &= (1, 0, \dots, 0, D_1 h(\hat{a})), \\ w_2 &= (0, 1, \dots, 0, D_2 h(\hat{a})), \end{aligned}$$

and so on. We claim that $T_a M = \operatorname{span}\{w_1, \dots, w_{n-1}\}$, which implies the result.

To see this, suppose $v \in T_a M$, so that there exists $\gamma : (-\delta, \delta) \to M$ with $\gamma(0) = a$ and $\gamma'(0) = v$. In particular, choosing $\delta > 0$ smaller if necessary, we may assume $\gamma(t) \in U \times V$ for all t. Thus

$$\gamma(t) \in M \implies g(\gamma(t)) = 0 \implies \gamma_n(t) = h(\hat{\gamma}(t)),$$

and so

$$\gamma_n'(t) = \nabla h(\hat{\gamma}(t)) \cdot \hat{\gamma}'(t) = \sum_{j=1}^{n-1} D_j h(\hat{\gamma}(t)) \hat{\gamma}_j'(t).$$

In particular,

$$\gamma_n'(0) = \sum_{j=1}^{n-1} \hat{\gamma}_j'(0) D_j h(\hat{a}),$$

and so

$$v = \gamma'(0) = \sum_{j=1}^{n-1} \hat{\gamma}_j'(0) w_j \in \operatorname{span}\{w_1, \dots, w_{n-1}\}.$$

(To verify this final equality, just check component by component.)

Conversely, given a vector

$$v = \sum_{i=1}^{n-1} c_j w_j \in \operatorname{span}\{w_1, \dots, w_{n-1}\},$$

we need to show that $v \in T_a M$. To see this, we define

$$\gamma(t) = (\hat{a} + tc, h(\hat{a} + tc)), \quad c = (c_1, \dots, c_{n-1}).$$

Then $\hat{a} + tc \in U$ for t small enough, so that $g(\gamma(t)) = 0$ for t small enough. Similarly since $g \in C^1$ and $\nabla g(a) \neq 0$, we have $\nabla g(\gamma(t)) \neq 0$ for t small enough. Thus $\gamma : (-\delta, \delta) \to M$ for small enough $\delta > 0$. Then

$$\gamma'(t) \equiv (c, \nabla h(\hat{a} + tc) \cdot c) \in \mathbb{R}^n.$$

However, checking component by component, we deduce that $\gamma'(0) = v$, and thus $v \in T_a M$.

Finally, take $v \in T_a M$. We need to show $\nabla g(a) \cdot v = 0$. By definition, $v = \gamma'(0)$ for some $\gamma : (-\delta, \delta) \to M$. In particular, we have

$$g \circ \gamma \equiv 0, \quad \text{so that} \quad \tfrac{d}{dt}[g \circ \gamma] = \nabla g(\gamma(t)) \cdot \gamma'(t) \equiv 0.$$

Evaluating at $t=0$ yields $\nabla g(a)\cdot v=0$, as desired. □

We then have the following result on constrained optimization.

Theorem 7.28 (Lagrange Multipliers)**.** *Suppose $g:\mathbb{R}^n\to\mathbb{R}$ satisfies $g\in C^1(\mathbb{R}^n)$, and let*

$$M=\{x\in\mathbb{R}^n : g(x)=0 \quad \textit{and} \quad \nabla g(x)\neq 0\}.$$

Suppose $f:\mathbb{R}^n\to\mathbb{R}$ is differentiable and obtains a maximum or minimum on M at some $a\in M$. Then there exists $\lambda\in\mathbb{R}$ such that

$$\nabla f(a)=\lambda\nabla g(a).$$

Proof. Let T_aM be the tangent plane at a. We claim that

$$\nabla f(a)\cdot v=0 \quad \text{for all} \quad v\in T_aM.$$

To see this, suppose $\gamma:(-\delta,\delta)\to M$ satisfies $\gamma(0)=a$ and $\gamma'(0)=v$. Then, by assumption, the function

$$\varphi(t)=f\circ\gamma(t):(-\delta,\delta)\to\mathbb{R}$$

has a maximum or minimum at $t=0$. Thus by the chain rule,

$$0=\varphi'(0)=\nabla f(\gamma(0))\cdot\gamma'(0)=\nabla f(a)\cdot v,$$

as desired.

Now, let $\{w_1,\dots,w_{n-1}\}$ be a basis for T_aM. If we extend this to a basis $\{w_1,\dots,w_n\}$ for $\mathbb{R}^n$, then it follows that

$$\nabla f(a)=c_1w_n \quad \text{and} \quad \nabla g(a)=c_2w_n \quad \text{for some} \quad c_1,c_2\in\mathbb{R}.$$

Note that $c_2\neq 0$ (since $\nabla g(a)\neq 0$), so that

$$\nabla f(a)=c_1w_n=\tfrac{c_1}{c_2}\nabla g(a),$$

as desired. □

Example 7.17. Let us find the rectangular box of volume 1000 of minimal surface area. It's not hard to guess the answer, but the point is to exhibit the Lagrange multiplier approach:

Writing x,y,z for the dimensions of the box, we are trying to minimize the function

$$f(x,y,z)=xy+xz+yz$$

subject to

$$g(x,y,z)=0, \quad \text{where} \quad g(x,y,z)=xyz-1000.$$

Then at an extreme point, we will have $\nabla f=\lambda\nabla g$, which becomes

$$y+z=\lambda yz, \quad x+z=\lambda xz, \quad x+y=\lambda xy,$$

with $xyz=1000$. We multiply the three equations by x, y, and z, respectively, and substitute $xyz=1000$ in each; we then obtain

$$xy+xz=xy+yz=xz+yz=1000\lambda$$

Thus we should take $x=y=z$, and in particular $x=y=z=10$.

The higher dimensional implicit function theorem can be used to prove a higher dimensional version of the Lagrange Multiplier Theorem (with more constraints). The result is the following:

Theorem 7.29 (Lagrange Multipliers, several constraints)**.** *Suppose $G : \mathbb{R}^n \to \mathbb{R}^m$ is a continuously differentiable function, with $m < n$. Let M denote the set of $x \in \mathbb{R}^n$ so that $G(x) = 0$ and such that*

$$\nabla G_1(x), \dots, \nabla G_m(x) \quad \textit{are linearly independent.}$$

If $f : \mathbb{R}^n \to \mathbb{R}$ attains a local maximum or minimum on M at a, then there exist $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ such that

$$\nabla f(a) = \lambda_1 \nabla G_1(a) + \cdots + \lambda_m \nabla G_m(a).$$

Proof. Let's just sketch the proof.

The first essential point is to prove an analogue of Theorem 7.27. In particular, the assumption that $\nabla G_1, \dots, \nabla G_m$ are linearly independent allows us to apply the implicit function theorem to show that for each $a \in M$, we can define a tangent plane T_aM which will now be an $n - m$ dimensional subspace of $\mathbb{R}^n$. Moreover, $\nabla G_1(a), \dots, \nabla G_m(a)$ will all be orthogonal to this plane (and in particular will form a basis for the orthogonal complement to this plane). Then, if f has a local extremum on M at a, we can show that ∇f is orthogonal to this plane as well, and hence belongs to the span of $\nabla G_1(a), \dots, \nabla G_m(a)$. This yields the result. □

Example 7.18. Let find the maximum and minimum of $f(x, y, z) = x$ on the intersection of the plane $z = 1$ and the sphere $x^2 + y^2 + z^2 = 4$. Again, the point is to illustrate the method:

We let

$$g_1(x, y, z) = z - 1 \quad \text{and} \quad g_2(x, y, z) = x^2 + y^2 + z^2 - 4.$$

Since

$$\nabla f = (1, 0, 0), \quad \nabla g_1 = (0, 0, 1), \quad \text{and} \quad \nabla g_2 = (2x, 2y, 2z),$$

we are faced with solving

$$1 = 2\lambda_1 x, \quad 0 = 2\lambda_2 y, \quad 0 = \lambda_1 + 2\lambda_2 z$$

under the two constraints. Note that $\lambda_2 = 0$ OR $y = 0$. Since $2\lambda_2 = -\lambda_1$ and $2\lambda_1 x = 1$, we must have $y = 0$. But then the constraints yield $x^2 = 3$, so $x = \pm\sqrt{3}$. Thus the maximum and minimum are $\pm\sqrt{3}$, respectively.

8. Multivariable integral calculus

8.1. **Definition of the n-dimensional integral.** The key to defining the n-dimensional integral will be a suitable definition of the n-dimensional 'volume' of sets. In particular, we want to define a quantity that recovers length in one dimension, area in two dimensions, volume in three dimensions, and so on. To do this, we basically want a definition that will be guaranteed to give us the right answer if we are trying to measure the volume of an 'n-dimensional interval' (this is the generalization of an interval in $1d$, a rectangle in $2d$, and so on), and in general we will use some sort of approximation by intervals to define our notion of volume. Here we go:

Definition 8.1. A *closed interval* in $\mathbb{R}^n$ is a set of the form

$$I = I_1 \times I_2 \times \cdots \times I_n \subset \mathbb{R}^n,$$

where each $I_k = [a_k, b_k]$ is a closed interval in $\mathbb{R}$. We define the volume of such an interval by

$$v(I) = (b_1 - a_1)(b_2 - a_2)\cdots(b_n - a_n).$$

We say two intervals I and J are *nonoverlapping* if $I^\circ \cap J^\circ = \emptyset$ (that is, their interiors are disjoint).

Definition 8.2. Let $A \subset \mathbb{R}^n$ be a bounded set (this means that there exists $M > 0$ so that $|x| < M$ for all $x \in A$). We say that A is *contented* with volume $v(A)$ if for any $\varepsilon > 0$, there exist (i) nonoverlapping closed intervals $I_1, \dots, I_m \subset A$ so that

$$\sum_{j=1}^{m} v(I_j) > v(A) - \varepsilon$$

and (ii) closed intervals $J_1, \dots, J_\ell$ so that

$$A \subset \bigcup_{j=1}^{\ell} J_j \quad \text{and} \quad \sum_{j=1}^{\ell} v(J_j) < v(A) + \varepsilon.$$

This is a pretty abstract definition, so we should spend a bit of time understanding what it actually says. In some sense, it says that a set has 'volume' (we call it *contented*) if it can be closely approximated by finitely many closed intervals, and the volume is a quantity that we can approximate arbitrarily well using closed intervals. In dimensions $n = 1, 2$, instead of 'volume' we will usually say 'length' and 'area', respectively.

Example 8.1 (Area of a triangle)**.** Let A be a triangle with base b and height a. That is, we consider A as the set

$$\{(x, y) \in \mathbb{R}^2 : x \in [0, b] \quad \text{and} \quad y \in [0, \tfrac{a}{b}x]\}.$$

We know that the area should be $\frac{1}{2}ab$. Let's see that our abstract definition above agrees with this.

We split $[0, b]$ into n equal subintervals. We then take I_k to be the rectangle with base $[(k-1)b/n, kb/n]$ and height $(k-1)a/n$ and take J_k to be the rectangle with base $[(k-1)b/n, kb/n]$ and height ka/n. The sum of the areas of the I_k is

$$\sum_{k=1}^{n} \tfrac{b}{n} \cdot \tfrac{(k-1)a}{n} = \tfrac{1}{2}ab - \tfrac{ab}{2n},$$

while the sum of the areas of the J_k is

$$\sum_{k=1}^{n} \tfrac{b}{n}\tfrac{ka}{n} = \tfrac{1}{2}ab + \tfrac{ab}{2n}$$

(you'll verify these computations in the homework). Thus, given $\varepsilon > 0$, we can choose n so that $\frac{ab}{2n} < \varepsilon$ and conclude that the area is $\frac{1}{2}ab$.

According to our definition, we simply cannot assert that every set has volume. That is, not every set is 'contented'. Here is an example:

Example 8.2. Let

$$A = \{(x, y) \in \mathbb{R}^2 : \quad x, y \in [0, 1] \cap \mathbb{Q}\}.$$

Then A contains no rectangles other than those of the form $[q, q]\times[r, r]$ for $q, r \in \mathbb{Q}$, which have volume zero. Thus for any finite collection I_j of rectangles contained in A, we have

$$\sum_{j} v(I_j) = 0.$$

On the other hand, if a finite collection of rectangles $J_1, \ldots, J_\ell$, contains A then we claim that we must have

$$[0, 1] \times [0, 1] \subset \bigcup_{j=1}^{\ell} J_j,$$

which implies that

$$\sum_{j=1}^{\ell} v(J_j) \geq 1.$$

It is therefore impossible to define a volume to the set A (because the volume $v(A)$ would have to satisfy

$$v(A) < \varepsilon \quad \text{and} \quad v(A) > 1 - \varepsilon \quad \text{for all} \quad \varepsilon > 0,$$

which is impossible for any choice of $\varepsilon \in (0, \frac{1}{2})$).

To prove the claim, we use the density of $\mathbb{Q}\times\mathbb{Q}$ in $\mathbb{R}^2$ (another exercise for you). In particular, any point $(x, y) \in [0, 1] \times [0, 1]$ can be obtained as a limit of points $(p_n, q_n) \in [0, 1] \times [0, 1]$ with $p_n, q_n \in \mathbb{Q}$. Each element of this sequence (p_n, q_n) is contained in one of the intervals $J_1, \ldots, J_\ell$. In particular, there exists J_j containing infinitely many terms (p_n, q_n) (i.e. a subsequence of the original sequence). Since this subsequence still converges to (x, y) and J_j is closed, we conclude $(x, y) \in J_j$, which yields the claim.

The previous examples show us that basic shapes are going to have the volume we expect them to, but that we cannot hope to assign a volume to every set. The issue with the second example was that it had a lot of 'boundary'. In particular, while the set itself was just a discrete set of points (ordered pairs of rationals), the closure of this set is the entire square $[0, 1] \times [0, 1]$. So the *boundary*, which is the *closure* minus the *interior*, is actually the entire square $[0, 1] \times [0, 1]$. What we will show next is that a set is contented if and only if its boundary is *negligible*, which means it has volume equal to zero.

Theorem 8.1. *A bounded set A is contented if and only it its boundary is negligible (i.e. has volume zero).*

Proof. $\Longrightarrow$: Suppose A is contented and let $\varepsilon > 0$. We choose closed intervals $I_1, \dots, I_m$ and $J_1, \dots, J_\ell$ as in Definition 8.2. Now take any closed interval R containing A, and let $\mathcal{P}$ be a partition of R (that is, a finite collection of nonoverlapping, closed intervals whose union equals R) such that each I_k and each J_j is a union of intervals of $\mathcal{P}$. We then let $R_1, \dots, R_a$ be the intervals of $\mathcal{P}$ that comprise the union

$$\bigcup_{k=1}^{m} I_k,$$

and let $R_{a+1}, \dots, R_b$ be the additional intervals contained in the union

$$\bigcup_{j=1}^{\ell} J_j.$$

It follows that the boundary of A, denoted ∂A, satisfies

$$\partial A \subset \bigcup_{k=a+1}^{b} R_k.$$

We now observe that

$$\begin{aligned} \sum_{k=a+1}^{b} v(R_k) &= \sum_{k=1}^{b} v(R_k) - \sum_{k=1}^{a} v(R_k) \\ &\leq \sum_{j=1}^{\ell} v(J_j) - \sum_{k=1}^{m} v(I_k) \\ &< [v(A) + \varepsilon] - [v(A) - \varepsilon] = 2\varepsilon. \end{aligned}$$

As $\varepsilon > 0$ was arbitrary, this implies that ∂A is contented with $v(\partial A) = 0$.

$\Longleftarrow$: We suppose ∂A is negligible. To show that A is contented, it suffices to show that for any $\varepsilon > 0$, there exist intervals $J_1, \dots, J_\ell$ containing A and nonoverlapping intervals $I_1, \dots, I_m$ contained in A such that

$$\sum_{j=1}^{\ell} v(J_j) - \sum_{k=1}^{m} v(I_k) < \varepsilon,$$

since then we obtain that $v(A)$ is the supremum of all such sums $\sum v(I_k)$ (or, the infimum of all such sums $\sum v(J_j)$).

Thus, we let R be a closed interval containing A and $\varepsilon > 0$, and we choose intervals $R_1, \dots, R_a$ covering ∂A with

$$\sum_{i=1}^{a} v(R_i) < \varepsilon.$$

We let $\mathcal{P}$ be a partition of R so that each R_i is a union of elements of $\mathcal{P}$. Then if $I_1, \dots, I_m$ are the intervals of $\mathcal{P}$ contained in A and $J_1, \dots, J_\ell$ are the intervals of $\mathcal{P}$ contained in $A \cup \bigcup_{i=1}^{a} R_i$, then we have

$$A \subset \bigcup_{j=1}^{\ell} J_j \quad \text{and} \quad \Big[\bigcup_{j=1}^{\ell} J_j\Big] \Big\backslash \Big[\bigcup_{k=1}^{m} I_k\Big] \subset \bigcup_{i=1}^{a} R_i.$$

Since all of the intervals I_k are contained in the collection $\{J_j\}$, it follows that

$$\sum_{j=1}^{\ell} v(J_j) - \sum_{k=1}^{m} v(I_k) \leq \sum_{i=1}^{a} v(R_i) < \varepsilon,$$

which yields the result. □

This result provides a useful characterization of contentedness of sets, since we can often recognize when a given set is negligible. Using this result, we can also deduce that the intersection, union, or set difference of two contented sets is again contented.

The definition of volume plays a central role in our topic of interest, namely, the n-dimensional integral. To state the definitions, we need a bit more terminology:

- If $f : \mathbb{R}^n \to \mathbb{R}$ is a nonnegative function (that is, $f(x) \geq 0$ for every $x \in \mathbb{R}^n$), we define the *ordinate set*
$$\mathcal{O}_f = \{(x, y) \in \mathbb{R}^n \times \mathbb{R} : 0 < y < f(x)\} \subset \mathbb{R}^{n+1}.$$
- We define the *positive* and *negative parts* of a function $f : \mathbb{R}^n \to \mathbb{R}$, denoted f^+ and f^-, by
$$f^+(x) = \max\{0, f(x)\} \quad \text{and} \quad f^-(x) = \max\{0, -f(x)\}.$$
- We say that f is *of bounded support* if there exists an interval $I \subset \mathbb{R}^n$ such that $f(x) = 0$ for $x \notin I$.

We can now define (Riemann) integrability for functions $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 8.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a bounded function of bounded support. We say that f is *(Riemann) integrable* if the sets $\mathcal{O}_{f^+}$ and $\mathcal{O}_{f^-}$ are both contented, and we define the integral of f by

$$\int f = v(\mathcal{O}_{f^+}) - v(\mathcal{O}_{f^-}).$$

We will see later that in the case $n = 1$, this agrees with our previous definition of Riemann integrable.

More generally, given a set $A \subset \mathbb{R}^n$, we may define the integral of f over A as follows. We let χ_A denote the *characteristic function* of A, defined by

$$\chi_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

If $f\chi_A$ is integrable, we define

$$\int_A f = \int f\chi_A.$$

If $f\chi_A$ is not integrable, we say the integral $\int_A f$ is not defined.

Using the definition alone, we can prove a few key properties of the integral:

Lemma 8.2. *If f is integrable and $f(x) \geq 0$ for all $x \in \mathbb{R}^n$, then*

$$\int f \geq 0.$$

Proof. Since $f^-(x) \equiv 0$, we have

$$\int f = v(\mathcal{O}_f) \geq 0.$$

□

Lemma 8.3. *If A is contented, then*

$$\int \chi_A = v(A).$$

Proof. Let $\varepsilon > 0$ and choose intervals $I_1, \dots, I_q$ and $J_1, \dots, J_p$ in $\mathbb{R}^n$ so that

$$\cup_{i=1}^q I_i \subset A \subset \cup_{j=1}^p J_j,$$

with

$$\sum_{i=1}^q v(I_i) > v(A) - \varepsilon \quad \text{and} \quad \sum_{j=1}^p v(J_j) < v(A) + \varepsilon.$$

Then set $I_i' = I_i \times [0,1]$ and $J_j' = J_j \times [0,1]$. It follows that

$$\cup_{i=1}^q I_i' \subset \mathcal{O}_{\chi_A} \subset \cup_{j=1}^p,$$

with

$$\sum_{i=1}^q v(I_i') > v(A) - \varepsilon \quad \text{and} \quad \sum_{j=1}^p v(J_j') < v(A) + \varepsilon.$$

It follows that $v(\mathcal{O}_{\chi_A}) = v(A)$, and hence

$$\int \chi_A = v(\mathcal{O}_{\chi_A}) = v(A),$$

as desired. □

The other fundamental properties of integrability (namely, that the set of integrable functions is a vector space and the mapping $f \mapsto \int f$ is linear) will be proven shortly.

It is a complicated question to ask for necessary and sufficient conditions for integrability. There is, however, a large class of functions that we can prove to be integrable, which includes most of the functions we are likely to encounter.

Definition 8.4 (Admissible)**.** We call f admissible if:

- f is bounded,
- f has bounded support, and
- f is continuous except on a negligible set.

Note that the set of admissible functions forms a vector space. Note also that if f is admissible, so are f^+ and f^- (these will be homework problems).

Theorem 8.4. *Admissible functions are integrable.*

Proof. Let f be an admissible function on $\mathbb{R}^n$, and let R be an interval so that $f = 0$ outside R. As $f^{\pm}$ are admissible, we may assume $f \geq 0$ without loss of generality.

We choose $M > 0$ such that $0 \leq f(x) \leq M$ for all x, and let D denote the negligible set where f is not continuous.

Now, given $\varepsilon > 0$, we choose $Q_1, \dots, Q_k$ to be closed intervals in $\mathbb{R}^n$ such that

$$D \subset \bigcup_{i=1}^{k} \operatorname{int}(Q_i) \quad \text{and} \quad \sum_{i=1}^{k} v(Q_i) < \varepsilon.$$

Now set $K = R \backslash \bigcup_{i=1}^{k} \operatorname{int}(Q_i)$. Then K is compact, and hence f is uniformly continuous on K. In particular, there exists $\delta > 0$ so that for all $x, y \in K$,

$$|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

We now let $\mathcal{P}$ be a partition of R so that each Q_i is a union of intervals of $\mathcal{P}$ and such that each interval of $\mathcal{P}$ has diameter $< \delta$. We then write $R_1, \dots, R_q$ for the intervals of $\mathcal{P}$ contained in K and set

$$a_i = \inf_{R_i} f, \quad b_i = \sup_{R_i} f,$$

so that $b_i - a_i < \varepsilon$ for all i. We then set

$$\begin{aligned} I_i &= R_i \times [0, a_i], \quad i = 1, \dots, q, \\ J_i &= R_i \times [0, b_i], \quad i = 1, \dots, q, \\ J_{q+i} &= Q_i \times [0, M], \quad i = 1, \dots, k. \end{aligned}$$

Finally, set $\mathcal{O}^* = \mathcal{O}_f \cup R \times \{0\}$. Then, since $\mathcal{O}_f = \mathcal{O}^* \backslash [R \times \{0\}]$ and $R \times \{0\}$ is contented, it suffices to show that $\mathcal{O}^*$ is contented.

To this end, we note that $I_1, \dots, I_q$ are nonoverlapping intervals contained in $\mathcal{O}^*$, while $J_1, \dots, J_{q+k}$ is a collection of intervals contaning $\mathcal{O}^*$. Moreover,

$$\begin{aligned} \sum_{j=1}^{q+k} v(J_j) - \sum_{i=1}^{q} v(I_i) &= \sum_{i=1}^{k} v(J_{q+i}) + \sum_{i=1}^{q} (b_i - a_i) v(R_i) \\ &< M \sum_{i=1}^{k} v(Q_i) + \varepsilon \sum_{i=1}^{q} v(R_i) \\ &\leq (M + v(R))\varepsilon. \end{aligned}$$

As $\varepsilon > 0$ was arbitrary, it follows that $\mathcal{O}^*$ is contented, as was needed to show. □

This result has a useful corollary:

Corollary 8.5. *If $f : A \to \mathbb{R}$ is nonnegative and continuous and $A \subset \mathbb{R}^n$ is contented, then the graph*

$$G_f = \{(x, f(x)) : x \in A\} \subset \mathbb{R}^{n+1}$$

is negligible.

Proof. We first extend f to zero outside of A. It follows that f is admissible, since it can be only be discontinuous on ∂A, which (as the boundary of a contented set) is negligible. Then, by the theorem, $\mathcal{O}_f$ is contented, and so its boundary is negligible. As G_f is a subset of $\partial \mathcal{O}_f$, we deduce that G_f is negligible. □

Note that the proof just given also demonstrated that the ordinate set $\mathcal{O}_f$ of a continuous, nonnegative function $f : A \to \mathbb{R}$ (on a contented set $A \subset \mathbb{R}^n$) is a contented set in $\mathbb{R}^{n+1}$. That is, we can always make sense of the volume of the region under the graph of a continuous function over a contented set.

We also note that if f is admissible and $A \subset \mathbb{R}^n$ is contented, then $f\chi_A$ is contented. Indeed, writing D for the negligible set where f is not continuous, we have that $f\chi_A$ is continuous off of the negligible set $D \cup \partial A$. We also have the following intuitive result:

Proposition 8.6. *Let f, g be integrable functions with $f(x) \leq g(x)$ for all $x \in \mathbb{R}^n$. Then*

$$\int f \leq \int g.$$

Consequently, if $|f(x)| \leq M$ for all $x \in \mathbb{R}^n$ and A is contented, then

$$\left|\int_A f\right| \leq M \cdot v(A).$$

In particular, $\int_A f = 0$ if A is negligible.

Proof. The condition $f \leq g$ guarantees that $f^+ \leq g^+$ and $f^- \geq g^-$. This shows that

$$\mathcal{O}_{f^+} \subset \mathcal{O}_{g^+} \quad \text{and} \quad \mathcal{O}_{g^-} \subset \mathcal{O}_{f^-},$$

which implies the first inequality.

Now we note that if $|f(x)| \leq M$, then $-M \leq f(x) \leq M$, so that

$$-M \cdot v(A) = \int_A (-M) \leq \int_A f \leq \int_A M = M \cdot v(A),$$

giving the second inequality. In particular, if $v(A) = 0$, we obtain $\int_A f = 0$. □

8.2. **Step functions, Riemann sums.** We are still missing some basic properties of the integral, like linearity. In what follows, it will be useful to have a 'step function criterion' for integrability, similar to the one we established in the one-dimensional setting. We begin with the following definition.

Definition 8.5 (Step function)**.** A function $h : \mathbb{R}^n \to \mathbb{R}$ is called a step function if there exist nonoverlapping intervals $I_1, \ldots, I_p$ and $a_1, \ldots, a_p \in \mathbb{R}$ such that

$$h = \sum_{i=1}^{p} a_i \chi_i, \quad \text{where} \quad \chi_i = \chi_{I_i}.$$

In this definition, we don't care whether the intervals are built out of closed, open, or half-open intervals in $\mathbb{R}$.

Step functions are integrable. Indeed, they are admissible!

Theorem 8.7. *If $h = \sum_{i=1}^{p} a_i \chi_i$ is a step function, then h is integrable, with*

$$\int h = \sum_{i=1}^{p} a_i v(I_i).$$

Proof. First note that h is continuous except possibly on the negligible set $\cup_{i=1}^{p} \partial I_i$. Thus h is admissible, and hence integrable.

Let's compute the integral. To simplify matters, let us take each $a_i > 0$. Then we have that

$$A := \bigcup_{i=1}^{p} I_i \times (0, a_i] \subset \mathcal{O}_h,$$

with $v(A) = \sum_{i=1}^{p} a_i v(I_i)$. On the other hand,

$$\mathcal{O}_h \subset A \cup \left[\left(\bigcup_{i=1}^{p} \partial I_i \right) \times [0, a_1 + \cdots + a_p] \right],$$

with the latter being the union of A with a negligible set. Thus we derive that

$$\int h = v(\mathcal{O}_h) = \sum_{i=1}^{p} a_i v(I_i).$$

□

Next, we establish linearity for the integral when restricted to step functions.

Theorem 8.8. *If h and k are step functions and $c \in \mathbb{R}$, then ch and $h+k$ are step functions, and we have*

$$\int ch = c \int h, \quad \int (h+k) = \int h + \int k.$$

Proof. The claims about ch are straightforward, so let us turn to the claims for $h+k$. We focus on the case that $h = a\chi_I$ and $k = b\chi_J$ (i.e. each of h and k just involves the characteristic function of a single set). The general case follows by induction on the number of intervals.

If I and J have disjoint interiors, then the result follows from the previous theorem. Otherwise, we have that $I_0 := I \cap J$ is an interval, and we can write

$$I \backslash I_0 = I_1' \cup \cdots \cup I_q', \quad J \backslash I_0 = I_1'' \cup \cdots \cup I_p'',$$

where the intervals I_j' and I_j'' are disjoint. We can then write

$$h + k = a\chi_I + b\chi_J = (a+b)\chi_{I_0} + \sum_{i=1}^{q} a\chi_{I_i'} + \sum_{i=1}^{p} b\chi_{I_i''},$$

expressing $h+k$ as a step function. By the previous theorem, we have

$$\begin{aligned} \int (h+k) &= (a+b)v(I_0) + a\sum_{i=1}^{q} v(I_i') + b\sum_{i=1}^{p} v(I_i'') \\ &= a[v(I_0) + \sum_{i=1}^{q} v(I_i')] + b[v(I_0) + \sum_{i=1}^{p} v(I_i'')] \\ &= av(I) + bv(J) = \int h + \int k, \end{aligned}$$

as desired. □

Here is our 'step function' criterion for integrability, which looks just like the $1d$ version:

Theorem 8.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be bounded with bounded support. Then f is integrable if and only if for any $\varepsilon > 0$, there exist step functions h and k such that*

$$h \leq f \leq k \quad \text{and} \quad \int (k - h) < \varepsilon.$$

Proof. The proof is similar to the $1d$ case. First suppose the criterion holds. Then, given $\varepsilon > 0$, we choose the corresponding step functions h, k and define

$$S = \{(x, x_{n+1}) \in \mathbb{R}^{n+1} : h(x) \leq x_{n+1} \leq k(x)\} \cup \partial\mathcal{O}_k \cup \partial\mathcal{O}_h.$$

Then $v(S) = \int (k-h) < \varepsilon$. However, we have

$$\hat{G}_f := \partial\mathcal{O}_f \backslash [\mathbb{R}^n \times \{0\}] \subset S,$$

and hence (since $\varepsilon > 0$ was arbitrary) we obtain that $\hat{G}_f$ is negligible.

Now, if $Q \subset \mathbb{R}^n$ is a rectangle containing the support of f, then we have that

$$\partial\mathcal{O}_{f^{\pm}} \subset [Q \times \{0\}] \cup \hat{G}_f,$$

so that $\partial\mathcal{O}_{f^{\pm}}$ are both negligible. But this in turn implies that $\mathcal{O}_{f^{\pm}}$ are both contented, and so f is integrable.

Now suppose f is integrable. Without loss of generality, we may assume $f \geq 0$ (since f^+ and f^- are both integrable). In this case, $\mathcal{O}_f$ is contented, and hence given $\varepsilon > 0$, we may find nonoverlapping intervals $I_1, \ldots, I_q$ contained in $\mathcal{O}_f$ and intervals $J_1, \ldots, J_p$ with

$$\mathcal{O}_f \subset \bigcup_{j=1}^{p} J_j$$

and

$$v(\mathcal{O}_f) - \varepsilon < \sum_{i=1}^{q} v(I_i) \leq \sum_{j=1}^{p} v(J_j) < v(\mathcal{O}_f) + \varepsilon.$$

We now define h and k as follows: if the vertical line in $\mathbb{R}^{n+1}$ through $x \in \mathbb{R}^n$ intersects I_i, we set

$$h(x) = \max\{y \in \mathbb{R} : (x, y) \in \bigcup_{i=1}^{q} I_i\},$$

and otherwise we set $h(x) = 0$. Similarly, if the vertical line in $\mathbb{R}^{n+1}$ through $x \in \mathbb{R}^n$ intersects J_j, we set

$$k(x) = \max\{y \in \mathbb{R} : (x, z) \in \bigcup_{j=1}^{p} J_j \quad \text{if} \quad z \in [0, y]\}$$

(and $k(x) = 0$ outside of the support of f). Then h and k are step functions with $h \leq f \leq k$, and since

$$\mathcal{O}_h \supset \bigcup_{j=1}^{q} I_i \quad \text{and} \quad \mathcal{O}_k \subset \bigcup_{j=1}^{p} J_j,$$

we obtain

$$\begin{aligned} \int (k-h) &= \int k - \int h \\ &= v(\mathcal{O}_k) - v(\mathcal{O}_h) \\ &< v(\mathcal{O}_f) + \varepsilon - [v(\mathcal{O}_f) - \varepsilon] = 2\varepsilon. \end{aligned}$$

This completes the proof. □

With our 'step function criterion' in place, we can at last prove linearity of the integral:

Theorem 8.10 (Linearity of the integral)**.** *If f_1, f_2 are integrable and $a_1, a_2 \in \mathbb{R}$, then $a_1 f_1 + a_2 f_2$ is integrable, with*

$$\int (a_1 f_1 + a_2 f_2) = a_1 \int f_1 + a_2 \int f_2.$$

Proof. Let's take the simplest case $a_1 > 0$ and $a_2 > 0$. Given $\varepsilon > 0$, we take step functions h_i, k_i so that $h_i \leq f_i \leq k_i$ and $\int [k_i - h_i] < \varepsilon$. Then we set $h = a_1 h_1 + a_2 h_2$ and $k = a_1 k_1 + a_2 k_2$ (which are both step functions), and we have (by Theorem 8.8)

$$h \leq a_1 f_1 + a_2 f_2 \leq k,$$

with

$$\int [k - h] = a_1 \int (k_1 - h_1) + a_2 \int (k_2 - h_2) < (a_1 + a_2)\varepsilon.$$

It follows that f is integrable. We can also check that both

$$\int a_1 f_1 + a_2 f_2 \quad \text{and} \quad a_1 \int f_1 + a_2 \int f_2$$

lie in the interval $[\int h, \int k]$, and hence are within $(a_1 + a_2)\varepsilon$ of one another. As $\varepsilon > 0$ was arbitrary, the result follows. □

As a corollary, we have a few other natural results about the integral:

Corollary 8.11.

(i) *Suppose A, B are contented with $A \cap B$ negligible, and f is integrable. Then*

$$\int_{A \cup B} f = \int_A f + \int_B f.$$

(ii) *Suppose f and g are integrable functions that are equal except on a negligible set D. Then*

$$\int f = \int g.$$

Proof. (i) In the case $A \cap B = \emptyset$, we have $\chi_{A \cup B} = \chi_A + \chi_B$, and so

$$\int_{A \cup B} f = \int f \chi_{A \cup B} = \int [f \chi_A + f \chi_B] = \int f \chi_A + \int f \chi_B = \int_A f + \int_B f.$$

In the case $A \cap B$ is negligible, we first observe $\int_{A \cap B} f = 0$. Then, using the special case just derived, we obtain

$$\begin{aligned} \int_{A \cup B} f &= \int_{A \setminus B} f + \int_{A \cap B} f + \int_{B \setminus A} f \\ &= \left[\int_{A \setminus B} f + \int_{A \cap B} f \right] + \left[\int_{B \setminus A} f + \int_{A \cap B} f \right] \\ &= \int_A f + \int_B f. \end{aligned}$$

(ii) We have that $f - g$ is zero except for on the set D, so

$$\int f - g = \int [f - g] \chi_D = \int_D [f - g] = 0,$$

where in the last step we use that D is negligible. By linearity, this implies $\int f = \int g$. □

We can now give the 'Riemann sum' formulation of integrability. We begin by introducing a few terms:

Definition 8.6.

- A *partition* of the interval $Q \subset \mathbb{R}^n$ is a collection $\mathcal{P} = \{Q_1, \ldots, Q_k\}$ is closed intervals with disjoint interiors such that $Q = \cup_{i=1}^k Q_i$.
- The *mesh* of $\mathcal{P}$ is the maximum of the diameters of the Q_i.
- A *selection* for $\mathcal{P}$ is a set $\mathcal{S} = \{x_1, \ldots, x_k\}$ with $x_i \in Q_i$ for each i.
- Given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is zero outside Q, a partition $\mathcal{P}$ of Q, and a selection $\mathcal{S}$, we define the corresponding *Riemann sum* for f by

$$R(f, \mathcal{P}, \mathcal{S}) = \sum_{i=1}^{k} f(x_i) v(Q_i).$$

Note that $R(f, \mathcal{P}, \mathcal{S})$ is the integral of the step function $h = \sum_{i=1}^k f(x_i) \chi_{Q_i}$.

Theorem 8.12. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is bounded and vanishes outside Q. Then f is integrable, with $\int f = I$, if and only if for any $\varepsilon > 0$, there exists $\delta > 0$ such that if $\mathcal{P}$ is a partition of Q with mesh $< \delta$ and $\mathcal{S}$ is any selection for $\mathcal{P}$, then we have*

$$|R(f, \mathcal{P}, \mathcal{S}) - I| < \varepsilon.$$

Proof. First suppose f is integrable and let $\varepsilon > 0$, and choose corresponding step functions h, k. We may assume that there is a partition $\mathcal{P}_0 = \{Q_1, \ldots, Q_s\}$ such that

$$h = \sum_{i=1}^{s} a_i \chi_i \quad \text{and} \quad k = \sum_{i=1}^{s} b_i \tilde{\chi}_i,$$

where $\chi_i, \tilde{\chi}_i$ are characteristic functions of intervals whose closure equals Q_i. Now set

$$A = Q \backslash \bigcup_{i=1}^{s} \text{int}(Q_i), \quad \text{so that} \quad v(A) = 0.$$

In particular, we can find $\delta > 0$ so that if $\mathcal{P}$ is a partition of Q with mesh $< \delta$, then the sum of the volumes of the intervals $P_1, \ldots, P_c$ that intersect A is less than ε. Write $P_{c+1}, \ldots, P_\ell$ for the remaining intervals in $\mathcal{P}$ (which lie in the interior of the Q_i).

Now let $\mathcal{S} = \{x_1, \ldots, x_\ell\}$ be any selection for $\mathcal{P}$. Then we have

$$h(x_i) \leq f(x_i) \leq k(x_i) \quad \text{for} \quad i = c+1, \ldots, \ell,$$

so that

$$\sum_{i=c+1}^{\ell} f(x_i) v(P_i) \quad \text{and} \quad \sum_{i=c+1}^{\ell} \int_{P_i} f$$

both belong to the interval $[\sum_{i=c+1}^{\ell} \int_{P_i} h, \sum_{i=c+1+}^{\ell} \int_{P_i} k]$, and hence

$$\left| \sum_{i=c+1}^{\ell} f(x_i) v(P_i) - \sum_{i=c+1}^{\ell} \int_{P_i} f \right| < \varepsilon.$$

On the other hand, we have that both

$$\sum_{i=1}^{c} f(x_i) v(P_i) \quad \text{and} \quad \sum_{i=1}^{c} \int_{P_i} f$$

lie in the interval $[-\|f\|\varepsilon, +\|f\|\varepsilon]$, where $\|f\|$ denotes the max of $|f|$. Thus

$$\left|\sum_{i=1}^{c} f(x_i)v(P_i) - \sum_{i=1}^{c} \int_{P_i} f\right| \leq 2\|f\|\varepsilon.$$

By the triangle inequality, we derive that

$$|R(f,\mathcal{P},\mathcal{S}) - I| = \left|\sum_{i=1}^{\ell} f(x_i)v(P_i) - \sum_{i=1}^{\ell} \int_{P_i} f\right| \leq (1 + 2\|f\|)\varepsilon,$$

as desired.

Now suppose that the Riemann sum condition holds. We let $\varepsilon > 0$ and choose a partition $\mathcal{P} = \{P_1, \ldots, P_a\}$ of Q such that for any selection $\mathcal{S}$,

$$|I - R(f,\mathcal{P},\mathcal{S})| < \varepsilon.$$

We then let $Q_1, \ldots, Q_a$ be disjoint intervals with $\overline{Q_i} = P_i$ for each $i = 1, \ldots, a$, and let χ_i be the characteristic function of Q_i. We let

$$m_i = \inf_{P_i} f \quad \text{and} \quad M_i = \inf_{P_i} f$$

and define the step functions

$$h = \sum_{i=1}^{a} m_i \chi_i, \quad k = \sum_{i=1}^{a} M_i \chi_i, \quad \text{which satisfy} \quad h \leq f \leq k.$$

Next, we let $\mathcal{S}' = \{x_1', \ldots, x_a'\}$ and $\mathcal{S}'' = \{x_1'', \ldots, x_a''\}$ be selections for $\mathcal{P}$ such that

$$|m_i - f(x_i')| < \varepsilon \quad \text{and} \quad |M_i - f(x_i'')| < \varepsilon$$

for each $i = 1, \ldots, a$. Then we have

$$\begin{aligned}\left|R(f,\mathcal{P},\mathcal{S}') - \int h\right| &= \left|\sum_{i=1}^{a} [f(x_i') - m_i]v(Q_i)\right| \\ &\leq \varepsilon \sum_{i=1}^{a} v(Q_i) < \varepsilon v(Q),\end{aligned}$$

and similarly

$$\left|R(f,\mathcal{P},\mathcal{S}'') - \int k\right| < \varepsilon v(Q).$$

We then obtain that

$$\begin{aligned}\int (k - h) \leq &\left|\int k - R(f,\mathcal{P},\mathcal{S}'')\right| + \left|R(f,\mathcal{P},\mathcal{S}'') - I\right| \\ &+ |I - R(f,\mathcal{P},\mathcal{S}')| + \left|R(f,\mathcal{P},\mathcal{S}') - \int h\right| \\ \leq &\, [2 + 2v(Q)]\varepsilon.\end{aligned}$$

Thus it follows from Theorem 8.9 that f is integrable. □

Using this Riemann sum criterion, one can show that if $f : \mathbb{R}^n \to \mathbb{R}$ is an integrable function that vanishes outside of Q, and $\{\mathcal{P}_k\}_{k=1}^{\infty}$ is a sequence of partitions of Q with the mesh of $\mathcal{P}_k$ converging to zero, then for any sequence of selections $\mathcal{S}_k$ corresponding to $\mathcal{P}_k$, we have

$$\int f = \lim_{k\to\infty} R(f,\mathcal{P}_k,\mathcal{S}_k).$$

This shows that integration is some sort of limiting process. Of course, as you already know, we don't actually use limits of Riemann sums to compute integrals. In the next section, we will discuss two important and practical results that play a key role in the actual computation of integrals over $\mathbb{R}^n$.

8.3. **Fubini's Theorem, change of variables.** In this section we prove two results that are essential for the actual computation of integrals. The first is Fubini's Theorem, which tell us that under the right conditions, we can compute an integral over $\mathbb{R}^n$ by iterating n one-dimensional integrations. The second is the change of variables formula, which is the higher-dimensional analogue of 'u-substitution' and hence is also an extremely important tool for integration.

We begin with Fubini's Theorem:

Theorem 8.13 (Fubini's Theorem)**.** *Let $f:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ be an integrable function such that for each $x\in\mathbb{R}^m$, the function*

$$f_x:\mathbb{R}^n\to\mathbb{R} \quad \textit{given by} \quad f_x(y)=f(x,y)$$

is integrable. Given contented sets $A\subset\mathbb{R}^m$ and $B\subset\mathbb{R}^n$, define $F:\mathbb{R}^m\to\mathbb{R}$ by

$$F(x)=\int_B f_x = \int_B f(x,y)\,dy.$$

Then F is integrable, with

$$\int_{A\times B} f = \int_A F.$$

Remark 8.14. Using the usual notation, we can write this as

$$\iint_{A\times B} f = \int_A \left[\int_B f(x,y)\,dy\right] dx,$$

and since the roles of x and y can be reversed, we obtain

$$\int_A \left[\int_B f(x,y)\,dy\right] dx = \int_{A\times B} f = \int_B \left[\int_A f(x,y)\,dx\right] dy$$

under suitable hypotheses on f. In particular, this theorem provides the rigorous justification for reversing the order of integration, and for computing higher dimensional integrals as iterated one-dimensional integrals.

Proof. Without loss of generality, we may assume $f(x,y)=0$ for $(x,y)\notin A\times B$.

Now, first suppose that χ is the characteristic function of an interval $I\times J\subset\mathbb{R}^m\times\mathbb{R}^n$. Then note that

$$\begin{aligned}\int\left[\int \chi(x,y)\,dy\right]dx &= \int_I\left[\int_J dy\right]dx \\ &= \int_I v(J) = v(I)v(J) = v(I\times J) = \int\chi,\end{aligned}$$

which is the conclusion of Fubini's theorem in this special case. In particular, by linearity of the integral, we deduce that Fubini's theorem holds for an arbitrary step function.

Now, let $\varepsilon > 0$ and choose corresponding step functions $h \leq f \leq k$ with $\int k - h < \varepsilon$. Defining $h_x(y) = h(x,y)$ and $k_x(y) = k(x,y)$, we have that $h_x \leq f_x \leq k_x$ for each x. Thus, if we let

$$H(x) = \int h_x \quad \text{and} \quad K(x) = \int k_x,$$

then H, K are step functions satisfying $H \leq F \leq K$ and (by Fubini's Theorem for step functions)

$$\int_{\mathbb{R}^m} K - H = \int_{\mathbb{R}^{m+n}} (k-h) < \varepsilon.$$

It follows (by the step function criterion) that F is integrable on $\mathbb{R}^m$, with

$$\int_{\mathbb{R}^m} H \leq \int_{\mathbb{R}^m} F \leq \int_{\mathbb{R}^m} K.$$

Furthermore, we see that both $\int_{\mathbb{R}^m \times \mathbb{R}^n} f$ and $\int_{\mathbb{R}^m} F$ lie between $\int_{\mathbb{R}^m \times \mathbb{R}^n} h = \int_{\mathbb{R}^m} H$ and $\int_{\mathbb{R}^m \times \mathbb{R}^n} k = \int_{\mathbb{R}^m} K$, and these last two integrals differ by $< \varepsilon$. Thus

$$\left| \int_{\mathbb{R}^m \times \mathbb{R}^n} f - \int_{\mathbb{R}^m} F \right| < \varepsilon.$$

As $\varepsilon > 0$ was arbitrary, the result follows. □

A few typical applications are the following (which we state but do not prove):

- Cavalieri's Principle refers to the following: If A is a contented subset of $\mathbb{R}^{n+1}$ with $A \subset R \times [a,b]$ for intervals $R, [a,b]$, and
$$A(t) = \{x \in \mathbb{R}^n : (x,t) \in A\}$$
is contented for each $t \in [a,b]$, then
$$v(A) = \int_a^b v(A(t))\, dt.$$
That is, we can compute volumes by adding up 'slices' of one lower dimension. (Cavalieri was computing volumes this way in the 1600s.)
- If $A \subset \mathbb{R}^n$ is contented and $f_1 \leq f_2$ are continuous functions on A, then
$$C = \{(x,y) : x \in A \quad \text{and} \quad f_1(x) \leq y \leq f_2(x)\}$$
is a contented set. If $g : C \to \mathbb{R}$ is continuous, then
$$\int_C g = \int_A \left[\int_{f_1(x)}^{f_2(x)} g(x,y)\, dy \right] dx.$$
- If $Q = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ is an interval in $\mathbb{R}^3$ and f is an integrable function, then
$$\int_Q f = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} f(x,y,z)\, dz\, dy\, dx.$$
Of course, we may integrate in whichever order we like.

You will work out some examples on the homework.

We turn next to the *change of variables* formula, which is essentially our higher-dimensional analogue of 'u-substitution'. Let us begin with a definition and the careful statement of the result.

Definition 8.7. We say that $T:\mathbb{R}^n\to\mathbb{R}^n$ is C^1-invertible on U if it is in $C^1(U)$ and injective on U, with the inverse map $T^{-1}:T(U)\to U$ belonging to $C^1(T(U))$.

Theorem 8.15 (Change of variables formula)**.** *Let Q be an interval in $\mathbb{R}^n$. Suppose $T:\mathbb{R}^n\to\mathbb{R}^n$ is a mapping that is C^1-invertible on the interior of Q. If $f:\mathbb{R}^n\to\mathbb{R}$ is integrable and $f\circ T:\mathbb{R}^n\to\mathbb{R}$ is also integrable, then*

$$\int_{T(Q)} f = \int_Q (f\circ T)|\det T'|.$$

Remark 8.16. Because $T\in C^1$, we have that $T'(x)$ is a continuous function, and hence $|\det T'(x)|$ is continuous as well. Then, since $f\circ T$ is integrable, we have that $(f\circ T)|\det T'|$ is integrable as well.

Remark 8.17. One calls $|\det T'|$ the *Jacobian determinant* of T. If we imagine trying to compute the integral

$$\int_A f(u)\,du$$

by writing $u=T(x)$ and trying to view the domain A in the form $T(Q)$ for suitable Q, then we could denote $T'(x)=\frac{du}{dx}$ and denote the determinant by $|\frac{du}{dx}|$. With this notation the change of variables formula looks like

$$\int_{T(Q)} f(u)\,du = \int_Q f(u(x))\,|\tfrac{du}{dx}|dx,$$

which 'looks reasonable' and lines up with the familiar 'u-substitution' from one-dimensional calculus. Actually, in the $1d$ setting, you typically try to recognize a given integral in the form appearing on the right-hand side, and then use the 'u-substitution' to convert it into the integral on the left. For example, when faced with

$$\int_0^2 \cos(x^2)\,2x\,dx,$$

we set $u(x)=x^2$ (so $\frac{du}{dx}=2x$) and obtain

$$\int_0^2 \cos(x^2)\,2x\,dx = \int_0^4 \cos(u)\,du.$$

Before we begin the proof of Theorem 8.15, let us at least attempt to understand why it should be true. The integral $\int_{T(Q)} f$ might be approximately computed by splitting up Q into a bunch of small intervals Q_i and summing the integrals $\int_{T(Q_i)} f$, which in turn can be approximated by $v(T(Q_i))f(T(x_i))$ for some selection of $x_i\in Q_i$. Now, if T is differentiable and Q_i is a very small interval, then (up to a translation) we can approximate the curved region $T(Q_i)$ by the image of Q_i under the differential of T; in particular we expect $v(T(Q_i))\approx v(dF_{x_i}(Q_i))$. Using linear algebra, we can show that for a linear transformation L and a rectangle R, we have $v(L(R))=|\det L|\,v(R)$. Thus we will have $v(dF_{x_i}(Q_i))=|\det T'(x_i)|\,v(Q_i)$, and hence

$$\begin{aligned}\int_{T(Q)} f &\approx \sum_i f(T(x_i))v(T(Q_i))\\ &\approx \sum_i f(T(x_i))|\det T'(x_i)|v(Q_i) \approx \int_Q (f\circ T)|\det T'|,\end{aligned}$$

which is exactly what the change of variables formula makes precise.

Proof of Theorem 8.15. We will prove the result under the slightly stronger assumption that T is C^1-invertible in a neighborhood of Q (rather than the interior). One can then upgrade to the more general result with an additional argument. I'll leave this as an exercise for the interested student.

Let $\eta > 0$. We will begin by finding a suitable Riemann sum approximation to $\int_Q (f \circ T)|\det T'|$. In particular, we choose $\delta > 0$ so that if $\mathcal{P} = \{Q_1, \dots, Q_k\}$ is a partition of Q with mesh $< \delta_1$, then

$$\left| R - \int_Q (f \circ T)|\det T'| \right| < \eta,$$

where

$$R = \sum_{i=1}^{k} f(T(a_i))|\det T'(a_i)|v(Q_i), \quad a_i = \text{center}(Q_i).$$

Writing

$$m_i = \inf_{Q_i} f \circ T \quad \text{and} \quad M_i = \sup_{Q_i} f \circ T$$

(which are finite since $f \circ T$ is integrable and thus bounded), we obtain that

$$R \in [\alpha, \beta],$$

where

$$\alpha := \sum_{i=1}^{k} m_i |\det T'(a_i)|v(Q_i) \quad \text{and} \quad \beta := \sum_{i=1}^{k} M_i |\det T'(a_i)|v(Q_i).$$

Now, using the integrability of f, we find that by choosing $\delta > 0$ possibly even smaller, we can guarantee that

$$\sum_{i=1}^{k} [M_i - m_i] v(Q_i) < \eta$$

(since this is the difference between two Riemann sums for $\int f \circ T$). In particular,

$$\beta - \alpha = \sum_{i=1}^{k} [M_i - m_i] |\det T'(a_i)| v(Q_i) \leq C\eta,$$

where $C = \sup_Q |T'|$ (which is finite, since $T \in C^1$).

Our next goal is to find $\tilde{\alpha} \leq \tilde{\beta}$ with

$$\int_{T(Q)} f \in [\tilde{\alpha}, \tilde{\beta}] \tag{8.1}$$

and such that $|\alpha - \tilde{\alpha}|$ and $|\beta - \tilde{\beta}|$ are both small. To this end, we first claim that

$$\int_{T(Q)} f = \sum_{i=1}^{k} \int_{T(Q_i)} f.$$

This relies on the fact that each $T(Q_i)$ is contented and that the $T(Q_i)$ intersect only in their boundaries (see the homework). Thus (8.1) holds with

$$\tilde{\alpha} := \sum_{i=1}^{k} m_i v(T(Q_i)) \quad \text{and} \quad \tilde{\beta} := \sum_{i=1}^{k} M_i v(T(Q_i)).$$

We now estimate $|\tilde{\alpha}-\alpha|$ and $|\tilde{\beta}-\beta|$. The key will be to obtain the following: choosing $\delta > 0$ possibly even smaller, we have

$$\big|v(T(Q_i)) - |\det T'(a_i)|v(Q_i)\big| < \eta|\det T'(a_i)|v(Q_i) \tag{8.2}$$

for each $i = 1, \dots, k$. Let's take this for granted for the moment and complete the proof of Theorem 8.15; we will then sketch a proof of (8.2) below.

We first estimate

$$\begin{aligned}|\tilde{\beta}-\beta| &\leq \sum_{i=1}^{k} |M_i| \cdot \big|v(T(Q_i)) - |\det T'(a_i)|v(Q_i)\big| \\ &\leq \tilde{C}\sum_{i=1}^{k} \eta|\det T'(a_i)|v(Q_i) \leq C\tilde{C}v(Q)\cdot\eta,\end{aligned}$$

where $\tilde{C} = \sup_Q |f \circ T|$. Similarly,

$$|\tilde{\alpha}-\alpha| \leq C\tilde{C}v(Q)\cdot\eta.$$

Thus we find

$$R \in [\alpha,\beta] \quad \text{and} \quad \int_{T(Q)} f \in [\tilde{\alpha},\tilde{\beta}].$$

These (and the fact that $\beta - \alpha \leq C\eta$) imply

$$R - \int_{T(Q)} f \leq \beta - \tilde{\alpha} \leq [\beta-\alpha] + [\alpha - \tilde{\alpha}] \leq [C + C\tilde{C}v(Q)]\eta,$$

and similarly

$$\int_{T(Q)} f - R \leq \tilde{\beta} - \alpha \leq [\tilde{\beta}-\beta] + [\beta-\alpha] \leq [C + C\tilde{C}v(Q)]\eta.$$

Thus

$$\left|R - \int_{T(Q)} f\right| \leq [C + C\tilde{C}v(Q)]\eta,$$

which (by the triangle inequality) finally yields

$$\left|\int_Q (f\circ T)\det T' - \int_{T(Q)} f\right| \leq [1 + C + C\tilde{C}v(Q)]\eta.$$

As $\eta > 0$ was arbitrary, this implies the result, other than (8.2). □

Sketch of the proof of (8.2). We need to prove the following: Suppose T is C^1-invertible on some set U. Given $\eta > 0$, there exists $\delta > 0$ such that if $Q \subset U$ is an interval of diameter $< \delta$ and $a = \text{center}(Q)$, then

$$\big|v(T(Q)) - |\det T'(a)|v(Q)\big| < \eta|\det T'(a)|v(Q),$$

or equivalently

$$\left|\frac{v(T(Q))}{|\det T'(a)|v(Q)} - 1\right| < \eta. \tag{8.3}$$

To simplify matters slightly, I'm going to prove this under the additional assumption that Q is a cube (rather than a general interval). To motivate the proof, let's first see why we expect the result to be true in the first place. We are essentially trying to show that

$$v(T(Q)) \approx |\det T'(a)|v(Q).$$

The reason that this should be true is that

$$T(a+h) \approx T(a) + T'(a)h, \tag{8.4}$$

so if we let h vary over $C_\delta = [-\delta, \delta]^n$, then the left-hand side ranges over the set $T(Q)$, while the right hand side should look like the image of C_δ under $T'(a)$, translated by $T(a)$. But by linear algebra considerations, this latter set should have volume $|\det T'(a)|v(C_\delta) = |\det T'(a)|v(Q)$. So (8.4) is the key idea, but we need to apply it in a bit of an unexpected way. We introduce the following notation for translations:

$$\tau_x(y) = x + y.$$

Then (8.4) can be rewritten

$$T \circ \tau_a(h) \approx \tau_{T(a)} \circ dT_a(h),$$

which rearranges to

$$dT_a^{-1} \circ \tau_{T(a)}^{-1} \circ T \circ \tau_a \approx \mathrm{Id},$$

where Id is the identity mapping. To make this precise, we define the mapping

$$F = dT_a^{-1} \circ \tau_{T(a)}^{-1} \circ T \circ \tau_a,$$

and observe that

$$dT_a(F(C_\delta)) = \tau_{T(a)}^{-1}(T(Q)).$$

Thus, using the facts that (i) translation preserves volume and (ii) for L linear and U contented, $v(L(U)) = |\det L| \cdot v(U)$, we have

$$v(T(Q)) = v(dT_a(F(C_\delta))) = |\det T'(a)|v(F(C_\delta)).$$

In particular, if we can show that $v(F(C_\delta)) \approx v(C_\delta) = v(Q)$, then we will be done. This should follow from the fact that $F \approx \mathrm{Id}$. To see this, we claim that given $\varepsilon > 0$, there exists $\delta > 0$ so that if $|x - a| < \delta$, then

$$C_{(1-\varepsilon)\delta} \subset F(C_\delta) \subset C_{(1+\varepsilon)\delta}. \tag{8.5}$$

In this case, we have

$$(1-\varepsilon)^n\delta^n \leq v(F(C_\delta)) \leq (1+\varepsilon)^n\delta^n.$$

In particular,

$$(1-\varepsilon)^n \leq \frac{v(F(C_\delta))}{v(Q)} = \frac{v(T(Q))}{|\det T'(a)|v(Q)} \leq (1+\varepsilon)^n,$$

which (for $\varepsilon = \varepsilon(\eta)$ sufficiently small) yields (8.3).

So, take $x \in C_\delta$ and observe that by construction,

$$F(x) = dT_a^{-1}[T(a+x) - T(a)], \quad \text{so that} \quad F(0) = 0.$$

In particular, applying the mean value theorem to each component F^j, we can write

$$F^j(x) = \nabla F^j(c_j) \cdot x = \sum_{k=1}^n D_k F^j(c_j) x_k \quad \text{for some} \quad c_j \in C_\delta.$$

Now, by the chain rule (and the fact that the differential of a translation is the identity), we have

$$dF_x = dT_a^{-1} \circ dT_x$$

As T is C^1-invertible, it follows that $dF_x \to \mathrm{Id}$ as $x \to a$. In particular, for $\delta > 0$ sufficiently small, we have

$$|D_k F^j(c) - \delta_{jk}| < \frac{\varepsilon}{n} \quad \text{for all} \quad c \in C_\delta,$$

and thus

$$|F^j(x) - x_j| = \left| \sum_{k=1}^{n} [D_k F^j(c_j) - \delta_{jk}] x_k \right| \leq \varepsilon |x| \leq \varepsilon \delta,$$

which (since $|x_j| \leq \delta$) implies

$$(1-\varepsilon)\delta \leq F^j(x) \leq (1+\varepsilon)\delta \quad \text{for each} \quad j.$$

This implies (8.5) and completes the sketch of the proof of (8.2). □

The standard applications of the change of variables formula involve the use of different coordinate systems for computing integrals.

Example 8.3 (Polar coordinates)**.** Let $T(r,\theta) = (r\cos\theta, r\sin\theta)$ be the polar coordinates mapping. Suppose $A \subset \mathbb{R}^2$ is the region

$$A = \{(x,y) \in \mathbb{R}^2 : a^2 \leq x^2 + y^2 \leq b^2\}.$$

Then we have $A = T(Q)$, where

$$Q = \{(r,\theta) \in \mathbb{R}^2 : r \in [a,b] \quad \text{and} \quad \theta \in [0, 2\pi]\},$$

and T is C^1 invertible on the interior of Q. Then we can write

$$\int_A f = \int_{T(Q)} f = \int_Q (f \circ T) |\det T'|.$$

Since

$$T'(r,\theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} \implies |\det T'| = r,$$

we find

$$\int_A f = \int_0^{2\pi} \int_a^b f(r\cos\theta, r\sin\theta) r \, dr \, d\theta,$$

which is the familiar formula for integration in polar coordinates.

Example 8.4 (Spherical coordinates)**.** The 'spherical coordinates' mapping is given by

$$T(\rho, \varphi, \theta) = (\rho \sin\varphi \cos\theta, \rho \sin\varphi \sin\theta, \rho\cos\varphi).$$

Then (check!) we have

$$|\det T'(\rho,\varphi,\theta)| = \rho^2 \sin\varphi,$$

and so if A is the image under T of

$$Q = \{(\rho,\varphi,\theta) : \rho \in [\rho_1, \rho_2],\ \varphi \in [\varphi_1, \varphi_2],\ \theta \in [\theta_1, \theta_2]\},$$

then

$$\int_A f = \int_{\theta_1}^{\theta_2} \int_{\varphi_1}^{\varphi_2} \int_{\rho_1}^{\rho_2} f(\rho\sin\varphi\cos\theta, \rho\sin\varphi\sin\theta, \rho\cos\varphi) \rho^2 \sin\varphi \, d\rho \, d\varphi \, d\theta.$$

This is the usual formula for integration in spherical coordinates.

You will work out a few more examples in the homework.

9. Differential forms and the classic theorems of vector calculus

Our final main goal is to prove the classic theorems of vector space (like Green's Theorem, Stokes' Theorem, the Divergence Theorem, and so on). We are going to take the approach involving the integration of 'differential forms'. Essentially, the main new thing we need to understand is how to integrate functions over k-dimensional 'surface patches' inside of $\mathbb{R}^n$, and this is what differential forms will let us do. So, this is where we are headed, but beware, there is a fair bit of abstraction and machinery that we need to build up before we can get there.

9.1. **Multilinear algebra and differential forms.**

We are now going to move beyond the word of 'linear' algebra and discuss some *multilinear* algebra.

Definition 9.1. A function $M : (\mathbb{R}^n)^k \to \mathbb{R}$ is called *k-multilinear (on $\mathbb{R}^n$)* (or just *multilinear*) if it is linear in each variable separately. So, for example,

$$M(\alpha a + \beta b, a^2, \dots, a^k) = \alpha M(a, a^2, \dots, a^n) + \beta M(b, a^2, \dots, a^k),$$

where $a, b, a^2, \dots, a^n \in \mathbb{R}^k$ and $\alpha, \beta \in \mathbb{R}$, with similar formulas holding for each component.

We already understand the case $k = 1$:

Lemma 9.1. *For $i = 1, \dots, n$, define the projection functions $dx_1, \dots, dx_n$ by*

$$dx_i(a) = a_i, \quad \textit{where} \quad a = (a_1, \dots, a_n).$$

For any linear function $L : \mathbb{R}^n \to \mathbb{R}$, there exist unique $\ell_1, \dots, \ell_n$ such that

$$L = \sum_{i=1}^n \ell_i dx_i.$$

Equivalently, $L(x) = \ell \cdot x$ for all $x \in \mathbb{R}^n$, where $\ell = (\ell_1, \dots, \ell_n)$.

Proof. Take $\ell_i = L(e_i)$. I'll leave it to you to check that this works, as well as uniqueness. □

Let's see if we can establish an analogous result for *multilinear* functions.

Definition 9.2. Given

$$I = (i_1, \dots, i_k), \quad \text{where} \quad 1 \le i_r \le n \quad \text{for each} \quad r,$$

we define

$$dx_I : (\mathbb{R}^n)^k \to \mathbb{R}$$

as follows. Given $(a^1, \dots, a^k) \in (\mathbb{R}^n)^k$, we form the $n \times k$ matrix A with columns $a^1, \dots, a^k$. We then let A_I be the $k \times k$ matrix whose r^{th} row is the $(i_r)^{th}$ row of A. We then define

$$dx_I(a^1, \dots, a^k) = \det A_I.$$

Example 9.1. Let $n = 4$ and $k = 2$. Suppose

$$A = [a^1\, a^2] = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix}.$$

Then

$$\text{if} \quad I = (3, 2), \quad \text{we have} \quad dx_I(a^1, a^2) = \det \begin{bmatrix} 5 & 6 \\ 3 & 4 \end{bmatrix} = 2.$$

Note that our definition makes dx_I a k-multilinear function, since we have that the determinant function $\det : (\mathbb{R}^k)^k \to \mathbb{R}$ is k-multilinear (viewed as a function of its rows).

Similar to finding the matrix of a linear transformation, we can describe an arbitrary k-multilinear function in the following way:

Lemma 9.2. *Let $M : (\mathbb{R}^n)^k \to \mathbb{R}$ be a k-multilinear function. Given $i_1, \dots, i_k \in [1, n]$, define*

$$\alpha_{i_1,\dots,i_k} = M(e^{i_1}, \dots, e^{i_k}),$$

where e^i is the i^{th} standard basis vector in $\mathbb{R}^n$. Then

$$M(a^1, \dots, a^k) = \sum_{i_1,\dots,i_k=1}^{n} \alpha_{i_1,\dots,i_k} a^1_{i_1} a^2_{i_2} \dots a^k_{i_k},$$

where $a^1 = (a^1_1, a^1_2, \dots, a^1_n)$, and so on.

The proof is by induction on k, where $k = 1$ is equivalent to the fact that a linear map $L : \mathbb{R}^n \to \mathbb{R}$ is of the form $L(a) = \alpha \cdot a$ for some $\alpha \in \mathbb{R}^n$. The proof is not particularly illuminating, so let's skip it and instead work out the next simplest case:

Example 9.2. Suppose $M : (\mathbb{R}^3)^2 \to \mathbb{R}$ is 2-multilinear on $\mathbb{R}^3$. Then we define

$$\alpha_{i,j} = M(e^i, e^j), \quad 1 \le i, j \le 3.$$

Then

$$M(a^1, a^2) = \sum_{i=1}^{3} \sum_{j=1}^{3} \alpha_{i,j} a^1_i a^2_j$$

In particular, M is just determined by the 3^2 numbers $\alpha_{i,j}$.

Lemma 9.2 is the best one can hope for if one considers arbitrary multilinear functions. So far, we have not succeeded in writing a multilinear function as a linear combination of the functions dx_I. In fact, we cannot write *arbitrary* multilinear functions in terms of the dx_I. This is because the dx_I have a certain special property (preserved under linear combinations), namely that of being *alternating*:

Definition 9.3. A multilinear function $M : (\mathbb{R}^n)^k \to \mathbb{R}$ is *alternating* if whenever some pair of the vectors $a^1, \dots, a^k$ is equal, we have

$$M(a^1, \dots, a^k) = 0.$$

We denote the set of alternating k-multlinear functions on $\mathbb{R}^n$ by $\Lambda^k(\mathbb{R}^n)$.

The fact that the functions dx_I are alternating follows from the corresponding property of the determinant; namely, the determinant of a matrix that has two equal rows is guaranteed to be zero. To see this, consider Example 9.1 again, but consider evaluating $dx_I(a^1, a^1)$.

By linearity, the property of being alternating is equivalent to the property that M changes sign if two inputs are exchanged, e.g.

$$M(a, b) = -M(b, a)$$

in the setting of a 2-multilinear function. Indeed, in this setting the formula above is equivalent to $M(a + b, a + b) = 0$, and the general case is similar.

We can also see that if M is alternating and $a^1, \dots, a^k$ are linearly dependent, then

$$M(a^1, \dots, a^k) = 0.$$

Indeed, in this case, one of the vectors may be written as a linear combination of the others, for example

$$a^1 = \sum_{j=2}^{k} c_j a^j.$$

But in this case, we have

$$M(a^1, \dots, a^k) = \sum_{j=2}^{k} c_j M(a^j, a^2, \dots, a^k).$$

As each summand involves evaluating M on a set of vectors in which a pair is equal, we have that each summand is zero. This result shows that if $k > n$, then there are no non-trivial alternating k-multilinear functions on $\mathbb{R}^n$.

Alternating k-multilinear functions on $\mathbb{R}^n$ can indeed be written as linear combinations of the functions dx_I. In the following, we say that $I = (i_1, \dots, i_k)$ is *increasing* if $1 \le i_1 < i_2 < \cdots < i_k \le n$.

Theorem 9.3. *Let* $M \in \Lambda^k(\mathbb{R}^n)$*. Define*

$$\alpha_I = M(e^{i_1}, \dots, e^{i_k}), \quad \textit{where} \quad I = (i_1, \dots, i_k).$$

Then

$$M = \sum_{I \ \textit{increasing}} \alpha_I dx_I.$$

Proof. We first need the following identity: given I, J increasing,

$$dx_I(e^{j_1}, \dots, e^{j_k}) = \begin{cases} 1 & I = J \\ 0 & I \neq J. \end{cases}$$

To see this, first note that if $I = J$, then $dx_I(e^{j_1}, \dots, e^{j_k})$ is the determinant of the $k \times k$ identity matrix, which equals one. If $I \neq J$, then $dx_I(e^{j_1}, \dots, e^{j_k})$ is the determinant of a matrix with one row equal to zero, and hence equals zero. Let's just illustrate this with a specific example. Suppose $n = 4$ and $k = 2$, and $I = (1, 2)$, but $J = (1, 3)$. Then we form the matrix

$$[e^{j_1} \ e^{j_2}] = [e^1 \ e^3] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

and so (since $I = (1, 2)$) we take the first and second rows and compute

$$dx_I[e^1 \ e^3] = \det \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = 0.$$

(If instead $I = J = (1, 3)$ then we get the determinant of the 2×2 identity matrix.)

Now, let us define $\tilde{M} : (\mathbb{R}^n)^k \to \mathbb{R}$ by

$$\tilde{M}(a^1, \dots, a^k) = \sum_{I \text{ increasing}} \alpha_I dx_I(a^1, \dots, a^k),$$

which is manifestly k-multilinear and alternating. To show that $M = \tilde{M}$, it is enough to check that they agree on each $(e^{j_1}, \dots, e^{j_k})$, with $J(j_1, \dots, j_k)$ increasing. In fact, by the identity above,

$$\begin{aligned}\tilde{M}(e^{j_1}, \dots, e^{j_k}) &= \sum_{I \text{ increasing}} \alpha_I dx_I(e^{j_1}, \dots, e^{j_k}) \\ &= \alpha_J = M(e^{j_1}, \dots, e^{j_k}).\end{aligned}$$

□

A corollary to the result above is that the *only* alternating n-multilinear function D on $\mathbb{R}^n$ satisfying

$$D(e^1, \dots, e^n) = 1$$

is the function $D(a^1, \dots, a^n) = \det(A)$ (i.e. the determinant function), where A is the matrix with columns $a^1, \dots, a^n$.

With this bit of multilinear algebra under our belts, we can now define the notion of a differential form.

Definition 9.4 (Differential k-form)**.** Let $U \subset \mathbb{R}^n$. A *differential k-form* on U is a function

$$\alpha : U \to \Lambda^k(\mathbb{R}^n).$$

For notational purposes, we often denote $\alpha(x)$ by α_x. By Theorem 9.3, we may write

$$\alpha(x) = \sum_{I \text{ increasing}} a_I(x) dx_I$$

for suitable coefficients $a_I : U \to \mathbb{R}$.

We say that α is *continuous* if each a_I is continuous. Similarly we say $\alpha \in C^k(U)$ if each $a_I \in C^k(U)$.

Example 9.3. A differential 1-form on $\mathbb{R}^n$ is of the form

$$\alpha(x) = a_1(x)dx_1 + \cdots + a_n(x)dx_n.$$

We know this already; it is the content of Lemma 9.1. This can be written $\alpha(x) = a(x) \cdot x$ for some $a : \mathbb{R}^n \to \mathbb{R}$.

Example 9.4. A differential 2-form on $\mathbb{R}^3$ is of the form

$$\alpha(x) = a_{(1,2)}(x)dx_{(1,2)} + a_{(1,3)}(x)dx_{(1,3)} + a_{(2,3)}(x)dx_{(2,3)}.$$

For example, switching notation from $\alpha(x)$ to $\alpha|_x$, if

$$\alpha|_x = f(x)dx_{(1,2)} + g(x)dx_{(1,3)},$$

then for $x \in U$ and $b, c \in \mathbb{R}^3$,

$$\alpha|_x(b, c) = f(x) \det \begin{bmatrix} b_1 & c_1 \\ b_2 & c_2 \end{bmatrix} + g(x) \det \begin{bmatrix} b_1 & c_1 \\ b_3 & c_3 \end{bmatrix}$$

Example 9.5. Recall the notation for the differential of a function $a : U \subset \mathbb{R}^n \to \mathbb{R}$. In particular, the differential is the function $da : U \to \mathbb{R}^n$ given by

$$da|_x = (\tfrac{\partial a}{\partial x_1}\big|_x, \dots, \tfrac{\partial a}{\partial x_n}\big|_x).$$

We can then identify $da|_x$ with the differential 1-form on U with coefficients given by $\frac{\partial a}{\partial x_i}|_x$. Thus, in the notation above, we would write

$$da = \tfrac{\partial a}{\partial x_1}dx_1 + \cdots + \tfrac{\partial a}{\partial x_n}dx_n = \sum_{i=1}^{n} \tfrac{\partial a}{\partial x_i}dx_i.$$

This is a very reasonable-looking formula. In fact, you may have seen formulas like this before, although possibly without the proper context of how this formula should be interpreted.

By the way, what if $a(x) = x_i$ for some $i = 1, \dots, n$? Then the differential is $da \equiv (0, \dots, 1, \dots, 0)$, with the 1 in the i^{th} position. Then the formula we just wrote reduces to $dx_i = dx_i$, where the d on the left stands for the differential, x_i on the left stands for the function $x \mapsto x_i$, and the dx_i on the right means one of our standard 1-forms from above. Pretty nifty! This identity shows that the notation we are using is consistent.

Note that if we take the view that a function $a : U \to \mathbb{R}^n$ is a "0-form" on $\mathbb{R}^n$, then we obtain that the differential da of a 0-form a is a 1-form. Later, we will define the differential of an arbitrary C^1 differential form, and we will see that if α is a k-form, then $d\alpha$ is a $k+1$-form.

We are now going to introduce the notion of the *product* (also called the *wedge product* or *exterior product*) of differential forms; along the way, we will also find a new way of expressing the forms dx_I.

Given $\alpha \in \Lambda^k(\mathbb{R}^n)$ and $i \in \{1, \dots, n\}$, let's first try to define the notion of

$$\alpha \wedge dx_m \in \Lambda^{k+1}(\mathbb{R}^n), \quad m \in \{1, \dots, n\}.$$

Writing $\alpha = \sum a_I dx_I$ (the sum being over increasing I), we would like to impose linearity in the sense that

$$\Big[\sum a_I dx_I\Big] \wedge dx_m = \sum a_I [dx_I \wedge dx_m].$$

Thus it is sufficient to define $dx_I \wedge dx_m$ for an arbitrary k-tuple $I = (i_1, \dots, i_k)$. The most natural way to do this is simply to take

$$dx_I \wedge dx_m = dx_{(I,m)} \in \Lambda^{k+1}(\mathbb{R}^n).$$

To be clear, if $I = (i_1, \dots, i_k)$, then we write (I, m) to denote $(i_1, \dots, i_k, m)$.

The construction above can be extended further. In particular, if $I = (i_1, \dots, i_k)$ and $J = (j_1, \dots, j_\ell)$, then we can define

$$dx_I \wedge dx_J \in \Lambda^{k+\ell}(\mathbb{R}^n) \quad \text{by} \quad dx_I \wedge dx_J = dx_{(I,J)},$$

where $(I, J) = (i_1, \dots, i_k, j_1, \dots, j_\ell)$. According to this definition, if $I = (i_1, \dots, i_k)$, then we can equally well write

$$dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

In light of the discussion above, we can now make the following definition:

Definition 9.5. Let $\alpha \in \Lambda^k(\mathbb{R}^n)$ and $\beta \in \Lambda^\ell(\mathbb{R}^n)$. Then we define

$$\alpha \wedge \beta \in \Lambda^{k+\ell}(\mathbb{R}^n)$$

as follows: if

$$\alpha = \sum_{I \text{ increasing}} a_I dx_I \quad \text{and} \quad \beta = \sum_{J \text{ increasing}} b_J dx_J,$$

then

$$\alpha \wedge \beta = \sum_{I \text{ increasing}} \sum_{J \text{ increasing}} a_I b_J \, dx_I \wedge dx_J.$$

Remark 9.4. The definition is a little unsatisfying in the sense that we have not written $\alpha \wedge \beta$ as a sum over increasing $(k+\ell)$-tuples. However, noting that (i) if $\tilde{I}$ is obtained from I by interchanging two entries, then $dx_{\tilde{I}} = -dx_I$ and (ii) if I has any repeated entries, then $dx_I \equiv 0$, we can rewrite the sum above as a sum over increasing $(k+\ell)$-tuples. For example,

$$\begin{aligned} &dx_{(1,4,5)} \wedge dx_{(2,7)} = dx_{(1,4,5,2,7)} = -dx_{(1,2,4,5,7)}, \\ &dx_{(1,4,5)} \wedge dx_{(2,4)} = 0. \end{aligned}$$

In particular, for any pair I, J satisfying $i_k = j_\ell$ for some k, ℓ, we will have $dx_I \wedge dx_J = 0$. On the other hand, given a pair I, J such that (I, J) has all distinct entries, we can perform interchanges to obtain an increasing $(k+\ell)$-tuple L containing the same entries as (I, J); then $dx_L = \pm dx_I \wedge dx_J$ depending on whether an even or odd number of interchanges are needed.

Using the definition above, we can obtain the following general anticommutativity property: for $\alpha \in \Lambda^k(\mathbb{R}^n)$ and $\beta \in \Lambda^\ell(\mathbb{R}^n)$,

$$\beta \wedge \alpha = (-1)^{k\ell} \alpha \wedge \beta.$$

To see this, it suffices to consider $\alpha = dx_I$ and $\beta = dx_J$. Then the identity boils down to the claim that it takes $k\ell$ interchanges to turn (I, J) into (J, I). Indeed, it takes k interchanges to transform

$$(i_1, \dots, i_k, j_1, \dots, j_\ell) \quad \text{into} \quad (j_1, i_1, \dots, i_k, j_2, \dots, j_\ell),$$

and then similarly it takes k interchanges to transform

$$(j_1, i_1, \dots, i_k, j_2, \dots, j_\ell) \quad \text{into} \quad (j_1, j_2, i_1, \dots, i_k, j_3, \dots, j_\ell).$$

Since we need to move $j_1, \dots, j_\ell$ to the left, we see that altogether we require $k\ell$ interchanges.

The wedge product of differential forms is a more general case of an important product that you have already encountered in multivariable calculus.

Example 9.6. Suppose $\alpha, \beta \in \Lambda^1(\mathbb{R}^3)$. Then

$$\alpha = a_1 \, dx_1 + a_2 \, dx_2 + a_3 \, dx_3 \quad \text{and} \quad \beta = b_1 \, dx_1 + b_2 \, dx_2 + b_3 \, dx_3.$$

(In particular, we can identify α with $a \in \mathbb{R}^3$ via $\alpha(x) = a \cdot x$ for $x \in \mathbb{R}^3$, and similarly for β.) Let's compute $\alpha \wedge \beta \in \Lambda^2(\mathbb{R}^3)$. We have built this product to obey linearity, so (using anticommutativity properties) we can write:

$$\begin{aligned} \alpha \wedge \beta &= (a_1 \, dx_1 + a_2 \, dx_2 + a_3 \, dx_3) \wedge (b_1 \, dx_1 + b_2 \, dx_2 + b_3 dx_3) \\ &= (a_1 b_2 - a_2 b_1) dx_1 \wedge dx_2 + (a_1 b_3 - a_3 b_1) \, dx_1 \wedge dx_3 + (a_2 b_3 - a_3 b_2) \, dx_2 \wedge dx_3 \end{aligned}$$

The coefficients resemble those appearing in the definition of the cross product $a \times b$.

While the usefulness of differential forms may not yet be apparent, hopefully you are at least getting a handle on the 'algebra' of differential forms. In the meantime, we are going to push ahead in this direction, and continue defining operations on differential forms (namely, differentiation and integration).

Definition 9.6 (Differential of a k-form)**.** Suppose $\alpha \in C^1(U)$ is a differential k-form with coefficients a_I:

$$\alpha = \sum_{I \text{ increasing}} a_I \, dx_I, \quad a_I : U \subset \mathbb{R}^n \to \mathbb{R}.$$

We define

$$d\alpha : U \to \Lambda^{k+1}(\mathbb{R}^n)$$

by

$$d\alpha = \sum_{I \text{ increasing}} (da_I) \wedge dx_I,$$

where $da_I : U \to \mathbb{R}$ is the differential of a_I and the sum is over increasing k-tuples. In particular each $da_I : U \to \Lambda^1(\mathbb{R}^n)$, so the formula above shows that $d\alpha : U \to \Lambda^{k+1}(\mathbb{R}^n)$.

Note that the differential (viewed as a transformation from Λ^k to Λ^{k+1}) is a linear transformation, which is to be expected.

Example 9.7. In the case $k = 0$, α is just a C^1 real-valued function on $U \subset \mathbb{R}^n$ and its differential is just $d\alpha$ (the usual differential).

Example 9.8. Suppose $\alpha \in C^1(\mathbb{R}^3)$ is a differential 1-form of the form

$$\alpha = P \, dx + Q \, dy + R \, dz$$

Then

$$d\alpha = dP \wedge dx + dQ \wedge dy + dR \wedge dz.$$

Writing $dP = P_x dx + P_y dy + P_z dz$ (subscripts denoting partial derivatives) and similarly for dQ and dR, and using the commutativity properties, we derive

$$d\alpha = [R_y - Q_z] dy \wedge dz + [P_z - R_x] dz \wedge dx + [Q_x - P_y] dx \wedge dy.$$

The coefficients here are the same as those appearing in curl of the vector field (P, Q, R).

Example 9.9. Suppose $\beta \in C^1(\mathbb{R}^3)$ is a differential 2-form of the form

$$\beta = A \, dy \wedge dz + B \, dz \wedge dx + C \, dx \wedge dy,$$

then

$$d\beta = (A_x + B_y + C_z) dx \wedge dy \wedge dz.$$

This coefficient is the divergence of the vector field (A, B, C).

At this point, it looks like applying the differential to a differential form will keep producing higher and higher order differential forms. Actually, though, the following is true:

Theorem 9.5. *Suppose $\alpha \in C^2(U)$ is a differential k-form. Then $d(d\alpha) = 0$.*

Proof. By writing $\alpha = \sum_I a_I dx_I$ and applying linearity, we see that it suffices to consider α of the form

$$\alpha = f dx_I, \quad f \in C^2(U), \quad I = \{i_1, \ldots, i_k\}.$$

In this case,

$$d\alpha = df \wedge dx_I = \sum_{j=1}^{n} \tfrac{\partial f}{\partial x_j} \; [dx_j \wedge dx_I].$$

Then

$$d(d\alpha) = \sum_{\ell=1}^{n} \sum_{j=1}^{n} \tfrac{\partial^2 f}{\partial x_\ell \partial x_j} [dx_\ell \wedge dx_j \wedge dx_I].$$

Now observe that the diagonal entries (with $\ell = j$) are zero, since $dx_\ell \wedge dx_\ell \wedge dx_I = 0$ for any I. Otherwise, the (ℓ, j) and (j, ℓ) terms cancel, since the mixed second derivatives agree but

$$dx_\ell \wedge dx_j \wedge dx_I = -dx_j \wedge dx_\ell \wedge dx_I.$$

□

There is also a product rule for differentials of forms:

Theorem 9.6 (Product rule for differentials of forms)**.** *If $\alpha \in \Lambda^k$ and $\beta \in \Lambda^\ell$ are differentiable, then*

$$d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^k [\alpha \wedge (d\beta)].$$

Proof. By linearity, it is again enough to work with

$$\alpha = f dx_I, \quad \beta = g dx_J,$$

with $f, g \in C^1$. Then we have

$$\begin{aligned} d(a \wedge b) = d[(fg) \wedge (dx_I \wedge dx_J)] &= d(fg) \wedge (dx_I \wedge dx_J) \\ &= [df\, g + f\, dg] \wedge (dx_I \wedge dx_J) \end{aligned}$$

The first term becomes

$$g\, df \wedge (dx_I \wedge dx_J) = [df \wedge dx_I] \wedge [g dx_J] = (d\alpha) \wedge \beta.$$

For the second, we have to move dg next to J; this requires k interchanges to pass through $dx_{i_1} \wedge \cdots \wedge dx_{i_k}$, and hence introduces a factor of $(-1)^k$. In particular, this term becomes

$$f\, dg \wedge (dx_I \wedge dx_J) = (-1)^k [f dx_I] \wedge [dg \wedge dx_J] = (-1)^k \alpha \wedge (d\beta).$$

The result follows. □

We now turn to our main topic concerning differential forms, namely, *integration.* Recall that we are motivated by the desire to integrate over k-dimensional surfaces in $\mathbb{R}^n$. With this in mind, we define the following:

Definition 9.7 (Surface patch)**.** A k-dimensional *surface patch* in $\mathbb{R}^n$ is a C^1 mapping $F : Q \to \mathbb{R}^n$, where $Q \subset \mathbb{R}^k$ is an interval, such that F is one-to-one on the interior of F.

Example 9.10. Let $Q = [0, 2\pi] \times [0, \pi] \subset \mathbb{R}^2$ and let $F : Q \to \mathbb{R}^3$ be given by

$$F(\theta, \varphi) = \begin{bmatrix} \sin\varphi \cos\theta \\ \sin\varphi \sin\theta \\ \cos\varphi \end{bmatrix}.$$

Then the image of F is the unit sphere in $\mathbb{R}^3$. We identify F with its image $F(Q)$.

Definition 9.8. Let α be a C^1 differential k-form on $\mathbb{R}^n$, and let $\varphi : Q \to \mathbb{R}^n$ be a k-dimensional surface patch. The *integral of α over φ* is defined by

$$\int_{\varphi(Q)} \alpha := \int_Q \alpha\big|_{\varphi(u)}(D_1\varphi(u), \dots, D_k\varphi(u)),$$

where the integral on the right-hand side is the integral over a real-valued function on the k-dimensional interval $Q \subset \mathbb{R}^k$ (with integration variable u). In the integral on the right-hand side, we may write $du_1 \cdots du_k$ to emphasize that this is an k-dimensional integral with integration variable u. Another option would be $d^k u$, or (if you don't care to specify the dimension k) just du.

Example 9.11. Let us consider a special case, namely when $\varphi : \mathbb{R}^k \to \mathbb{R}^k$ is given by $\varphi(x) = x$ and $\alpha = fdu := f\, du_1 \wedge \cdots \wedge du_k$. Then

$$\int_Q fdu = \int_Q fdu_1 \wedge \cdots \wedge du_k = \int_Q fdu_1 \wedge \cdots du_k(e_1, \dots, e_k) = \int_Q f.$$

This means that when integrating a function f, you can equally well think of integrating the k-form $fdu_1 \wedge \cdots \wedge \text{\d{u}}_k$; if we denote $du = du_1 \wedge \cdots \wedge du_k$, then we can say that we are integrating fdu.

Example 9.12. Let $Q = [0,1] \times [0,1]$ and let

$$\varphi(u,v) = (u+v, u-v, uv), \quad \alpha = xdy \wedge dz + y\, dx \wedge dz.$$

Let's compute $\int_{\varphi(Q)} \alpha$: Noting that

$$\varphi'(u,v) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ v & u \end{bmatrix},$$

we can see that

$$dy \wedge dz(D_u\varphi, D_v\varphi) = u+v, \quad dx \wedge dz(D_u\varphi, D_v\varphi) = u-v.$$

Thus

$$\int_{\varphi(Q)} \alpha = \int_Q [(u+v)^2 + (u-v)^2] = \int_0^1 \int_0^1 2(u^2+v^2)\, du\, dv = \tfrac{4}{3}.$$

Example 9.13 (Arclength integrals)**.** Let us consider the case $k = 1$, so that $\gamma : [a,b] \to \mathbb{R}^n$ simply yields a curve in $\mathbb{R}^n$. We then wish to compute the integral $\int_\gamma \omega$ for a 1-form ω. In particular, if $\omega = \sum_{i=1}^n F_i dx_i$, then

$$\begin{aligned} \int_\gamma \omega &= \int_a^b \sum_{i=1}^n F_i(\gamma(t)) dx_i(\gamma'(t))\, dt \\ &= \int_a^b \sum_{i=1}^n F_i(\gamma(t))\gamma'(t)\, dt = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\, dt. \end{aligned}$$

We can express this integral in a slightly different way, which will be important later when we discuss Stokes' Theorem. In particular, we define the *unit tangent vector* at $x \in \gamma([a,b])$ by

$$T(x) = \tfrac{\gamma'(t)}{|\gamma'(t)|}, \quad \text{where} \quad x = \gamma(t).$$

One can check that this definition is actually independent of the parametrization $\gamma(\cdot)$ (as long as the parametrizations are 'equivalent', or 'orientation-preserving'), and in fact a choice of unit tangent vector provides an orientation of the curve.

Given the unit tangent T for a curve $C \in \mathbb{R}^n$, we define the *arclength form* ds by

$$ds|_x(v) = T(x) \cdot v,$$

which defines a differential 1-form on C. One can then verify that for $\omega = \sum_{i=1}^n F_i dx_i$ as above,

$$\int_\gamma \omega = \int_\gamma F \cdot T\, ds. \tag{9.1}$$

Proof of (9.1). Let $\gamma : [a,b] \to \mathbb{R}^n$ be an (oriented) parametrization of the curve C. Then, by definition of integration of a differential form,

$$\begin{aligned}\int_\gamma F \cdot T\, ds &= \int_a^b F(\gamma(t)) \cdot \tfrac{\gamma'(t)}{|\gamma'(t)|}\, ds_{\gamma(t)}(\gamma'(t)) \\ &= \int_a^b F(\gamma(t)) \cdot \tfrac{\gamma'(t)}{|\gamma'(t)|} T(\gamma(t)) \cdot \gamma'(t) \\ &= \int_a^b F(\gamma(t)) \cdot \gamma'(t)\, \tfrac{|\gamma'(t)|^2}{|\gamma'(t)|^2} \\ &= \int_a^b F(\gamma(t)) \cdot \gamma'(t)\, dt.\end{aligned}$$

The result follows. □

We are next going to introduce the notion of the *pullback* of a differential form, which will allow us to give a cleaner formula than the one appearing in the definition.

Definition 9.9 (Pullback)**.** Let $\varphi : \mathbb{R}^m \to \mathbb{R}^n$ be a C^1 mapping, and let α be a k-form on $\mathbb{R}^n$. Then for $u \in \mathbb{R}^m$, we define

$$(\varphi^*\alpha)\big|_u : (\mathbb{R}^n)^k \to \mathbb{R},$$

by

$$(\varphi^*\alpha)\big|_u(v_1, \dots, v_k) = \alpha\big|_{\varphi(u)}(d\varphi|_u(v_1), \dots, d\varphi|_u(v_k)).$$

We call $\varphi^*\alpha$ the *pullback* of α by φ.

Just based on the definition, we can see some connection to Definition 9.8 above, but what on earth does this operation really mean? Let's see some examples:

Example 9.14. What is the pullback of a function f (i.e. a '0-form')? It is another function:

$$(\varphi^* f)\big|_u = f|_{\varphi(u)}, \quad \text{i.e.} \quad (\varphi^* f)(u) = f(\varphi(u)) = f \circ \varphi(u).$$

Thus $\varphi^* f = f \circ \varphi$.

Example 9.15. What is the pullback of the 1-form $f\, dx_i$? By the definition,

$$(\varphi^*[f\, dx_i])\big|_u(v) = [f\, dx_i]\big|_{\varphi(u)}(d\varphi|_u(v)) = f \circ \varphi(u)\, [d\varphi\big|_u(v)]_i$$

Now note that $[d\varphi\big|_u(v)]_i$ given by $d\varphi_i\big|_u \cdot v$. It follows that

$$\varphi^*[f dx_i] = [f \circ \varphi]\, d\varphi_i = (\varphi^* f)\, d\varphi_i.$$

In particular, we can expand

$$\varphi^*[dx_i] = d\varphi_i = \sum_{j=1}^{m} \frac{\partial \varphi_i}{\partial u_j}\ du_j.$$

Example 9.16. What is the pullback of $dx_i \wedge dx_j$? We claim that

$$\varphi^*[dx_i \wedge dx_j] = \varphi^* dx_i \wedge \varphi^* dx_j = d\varphi_i \wedge d\varphi_j.$$

To see this, we first write

$$\begin{aligned}(\varphi^*[dx_i \wedge dx_j])|_u(v_1, v_2) &= [dx_i \wedge dx_j]\big|_{\varphi(u)}(d\varphi|_u(v_1), d\varphi|_u(v_2)) \\ &= [dx_i \wedge dx_j](d\varphi|_u(v_1), d\varphi|_u(v_2)) \\ &= dx_{(i,j)}(d\varphi|_u(v_1), d\varphi|_u(v_2)) \\ &= \det\begin{bmatrix} [d\varphi|_u(v_1)]_i & [d\varphi|_u(v_2)]_i \\ [d\varphi|_u(v_1)]_j & [d\varphi|_u(v_2)]_j \end{bmatrix}.\end{aligned}$$

Now note that

$$[d\varphi\big|_u(v_1)]_i = d\varphi_i\big|_u \cdot v_1 = \sum_{a=1}^{m} \varphi_{ia} v_{1a},$$

where we write $\varphi_{ia} = \frac{\partial \varphi_i}{\partial u_a}\big|_u$ and v_{1a} for the a^{th} component of v_1. With similar formulas for the other terms appearing in the matrix above, we derive

$$(\varphi^*[dx_i \wedge dx_j])|_u(v_1, v_2) = \sum_{a=1}^{m}\sum_{b=1}^{m} \varphi_{ia}\varphi_{jb}[v_{1a}v_{2b} - v_{2a}v_{1b}].$$

On the other hand, keeping the same notation from above, we have

$$\varphi^* dx_i = d\varphi_i = \sum_{a=1}^{m} \varphi_{ia} du_a,$$

so that

$$\begin{aligned}[\varphi^* dx_i \wedge \varphi^* dx_j]\big|_u(v_1, v_2) &= \sum_{a=1}^{m}\sum_{b=1}^{m} \varphi_{ia}\varphi_{jb}[du_a \wedge du_b](v_1, v_2) \\ &= \sum_{a=1}^{m}\sum_{b=1}^{m} \varphi_{ia}\varphi_{jb} \det\begin{bmatrix} v_{1a} & v_{1b} \\ v_{2a} & v_{2b} \end{bmatrix} \\ &= \sum_{a=1}^{m}\sum_{b=1}^{m} \varphi_{ia}\varphi_{jb}[v_{1a}v_{2b} - v_{1b}v_{2a}].\end{aligned}$$

The claim follows.

In light of the preceding examples, we can now see that some nice formulas hold for the pullback, like

$$\varphi^*(f\alpha) = (f \circ \varphi)\varphi^*\alpha, \quad \varphi^*(\alpha \wedge \beta) = \varphi^*\alpha \wedge \varphi^*\beta, \quad \varphi^*(\alpha + \beta) = \varphi^*\alpha + \varphi^*\beta.$$

We will prove the following:

Theorem 9.7. *Let $\varphi : Q \subset \mathbb{R}^k \to \mathbb{R}^n$ be a C^1 surface patch. Let α be a differential k-form on $\mathbb{R}^n$. Then*

$$\int_{\varphi(Q)} \alpha = \int_Q \varphi^*\alpha.$$

Recall that in light of Example 9.11, the integrand on the right-hand side should be viewed as $\varphi^*\alpha|_u$ evaluated at $(e_1,\dots,e_k)$.

Proof. By additivity, it is enough to consider a k-form of the type $\alpha = f dx_I$, where $I=(i_1,\dots,i_k)$ and $f:\mathbb{R}^n\to\mathbb{R}$. In this case, we have

$$\int_{\varphi(Q)} f dx_I = \int f\circ\varphi(u)\, dx_I(D_1\varphi(u),\dots,D_k\varphi(u)).$$

Now, by definition of dx_I, we have that

$$dx_I(D_1\varphi(u),\dots,D_k\varphi(u)) = \det\Big[\frac{\partial\varphi_{i_a}}{\partial u_j}\Big|_u\Big],$$

where $a,j=1,\dots,k$ and the matrix above has the specified value in row a and column j. So

$$\int_{\varphi(Q)} f dx_I = \int_Q f\circ\varphi(u)\det\Big[\frac{\partial\varphi_{i_a}}{\partial u_j}\Big|_u\Big].$$

On the other hand, we have

$$\begin{aligned}\varphi^*(f dx_I) &= (f\circ\varphi)[\varphi^* dx_{i_1}]\wedge\cdots\wedge[\varphi^* dx_{i_k}]\\ &= (f\circ\varphi)\, d\varphi_{i_1}\wedge\cdots\wedge d\varphi_{i_k}\\ &= (f\circ\varphi)\sum_{\ell_1,\dots,\ell_k=1}^{k}\frac{\partial\varphi_{i_1}}{\partial u_{\ell_1}}\cdots\frac{\partial\varphi_{i_k}}{\partial u_{\ell_k}}\, du_{\ell_1}\wedge\cdots\wedge du_{\ell_k},\end{aligned}$$

where we have expanded each $d\varphi_{i_a}$ in terms of the differentials du_ℓ, and each derivative is evaluated at u. Now observe that the sum above can be restricted to permutations $(\ell_1,\dots,\ell_k)$ of $(1,\dots,k)$ (i.e. we cannot have $\ell_i=\ell_j$ for any $i\neq j$, for then the sum reduces to zero). Writing $L=(\ell_1,\dots,\ell_k)$ for a typical permutation and

$$M_{aj} = \frac{\partial\varphi_{i_a}}{\partial u_j}\Big|_u,$$

we obtain

$$\varphi^*(f dx_I) = (f\circ\varphi)\Big[\sum_{\text{permutations } L}\sigma(L)M_{1\ell_1}\cdots M_{k\ell_k}\Big]du_1\wedge\cdots\wedge du_k,$$

where $\sigma(L)$ is the sign of the permutation L. By linear algebra considerations, we have

$$\sum_{\text{permutations } L}\sigma(L)M_{i_1\ell_1}\cdots M_{i_k\ell_k} = \det M = \det\Big[\frac{\partial\varphi_{i_a}}{\partial u_j}\Big|_u\Big].$$

Thus we have

$$\int_Q \varphi^*(dx_I)\, du = \int f\circ\varphi(u)\det\Big[\frac{\partial\varphi_{i_a}}{\partial u_j}\Big|_u\Big]$$

where we write $du = du_1\wedge\cdots\ du_k$. The result follows. □

Let's use this result to compute the integral from Example 9.12 another way.

Example 9.17. Let Q,φ,α be as in Example 9.12, that is,

$$\varphi(u,v) = (u+v, u-v, uv), \quad \alpha = x\,dy\wedge dz + y\,dx\wedge dz.$$

Then the pullback $\varphi^*\alpha$ is given by

$$\begin{aligned}\varphi^*\alpha &= (u+v)\varphi^*(dy)\wedge\varphi^*(dz) + (u-v)\varphi^*(dx)\wedge\varphi^*(dz)\\ &= (u+v)[du-dv]\wedge[vdu+udv] + (u-v)[du+dv]\wedge[vdu+udv]\\ &= (uv+v^2+uv+v^2)du\wedge dv + (u^2-uv-uv+v^2)du\wedge dv\\ &= 2(u^2+v^2)\,du\wedge\,dv.\end{aligned}$$

Then

$$\int_{\varphi(Q)}\alpha = \int_Q \varphi^*\alpha = \int_Q 2(u^2+v^2) = \tfrac{4}{3},$$

just as we computed above.

There's one more important result we should prove about pullbacks, namely:

Theorem 9.8. *If $\varphi : \mathbb{R}^m \to \mathbb{R}^n$ is C^1 and α is a C^1 differential k-form on $\mathbb{R}^n$, then*

$$d(\varphi^*\alpha) = \varphi^*(d\alpha).$$

Proof. By induction on k: If $k=0$, then α is a C^1 function, and by definition of the differential and pulback,

$$\varphi^*(d\alpha) = \sum_{j=1}^n \frac{\partial\alpha}{\partial x_j}\bigg|_{\varphi}\varphi^*(dx_j) = \sum_{j=1}^n\sum_{k=1}^m \frac{\partial\alpha}{\partial x_j}\bigg|_{\varphi}\frac{\partial\varphi_j}{\partial u_k}du_k.$$

On the other hand, by the chain rule,

$$d(\varphi^*\alpha) = \sum_{k=1}^m \frac{\partial}{\partial u_k}\big[\alpha\circ\varphi\big]du_k = \sum_{k=1}^m\sum_{j=1}^n \frac{\partial\alpha}{\partial x_j}\bigg|_{\varphi}\frac{\partial\varphi_j}{\partial u_k}\,du_k.$$

Thus the result holds for $k=0$.

Now suppose the result holds up to $(k-1)$-forms. Take a k-form of the form

$$\alpha = f dx_I\wedge dx_\ell =: \beta\wedge dx_\ell.$$

Then, by the product rule for differential forms and Theorem 9.5

$$d\alpha = d\beta\wedge dx_\ell - \beta\wedge d[dx_\ell] = d\beta\wedge dx_\ell.$$

Thus, using the inductive hypothesis,

$$\begin{aligned}\varphi^*(d\alpha) &= \varphi^*(d\beta\wedge dx_\ell)\\ &= \varphi^*(d\beta)\wedge\varphi^*(dx_\ell)\\ &= d(\varphi^*\beta)\wedge\varphi^*(dx_\ell).\end{aligned}$$

On the other hand, by the product rule, inductive hypothesis, and Theorem 9.5,

$$\begin{aligned}d(\varphi^*\alpha) &= d(\varphi^*\beta\wedge\varphi^*dx_\ell)\\ &= d(\varphi^*\beta)\wedge\varphi^*(dx_\ell) - \varphi^*\beta\wedge d[\varphi^*dx_\ell]\\ &= d(\varphi^*\beta)\wedge\varphi^*(dx_\ell) - \varphi^*\beta\wedge\varphi^*[d(dx_\ell)]\\ &= d(\varphi^*\beta)\wedge\varphi^*(dx_\ell).\end{aligned}$$

The result follows. □

9.2. **Manifolds and surface area.** We are now going to move towards some actual applications of the theory of differential forms. In particular, we would like to see how we can compute the area of surfaces inside $\mathbb{R}^n$.

We begin with the following definition:

Definition 9.10 (Area of a surface patch)**.** Suppose $F : Q \subset \mathbb{R}^k \to \mathbb{R}^n$ is a k-dimensional surface patch. We define the *k-dimensional surface area* of $F(Q)$ by

$$a(F) = \int_Q [\det[F'(u)]^t F'(u)]^{1/2},$$

where ^{t} denotes the transpose of a matrix.

This definition is motivated by the fact that if P is a k-dimensional parallelepiped in $\mathbb{R}^n$ spanned by $a_1, \dots, a_k$, then the k-dimensional area of P is given by

$$a(P) = [\det A^t A]^{1/2},$$

where A is the $n \times k$ matrix with columns given by $a_1, \dots, a_k$. As this is essentially a fact from linear algebra, we take it for granted here. Using this together with a Riemann sum type construction, it becomes natural to define the area of a surface patch as above.

Example 9.18. Let $F : [0, 2\pi] \to \mathbb{R}^2$ be given by $F(u) = \begin{bmatrix} \cos u \\ \sin u \end{bmatrix}$. Then

$$F'(u) = \begin{bmatrix} -\sin u \\ \cos u \end{bmatrix} \implies [F'(u)]^t F'(u) = \sin^2 u + \cos^2 u = 1.$$

Then the '1-dimensional surface area' (which is the same thing as arclength) of F is

$$a(F) = \int_0^{2\pi} 1 = 2\pi,$$

giving the arclength of the circle. Similarly, one can compute the surface area of the 2-dimensional sphere in $\mathbb{R}^3$ using the spherical coordinate map.

So far, we have defined the area of a surface patch that relies on the specific choice of a function describing the surface (also called a *parametrization*). For example, our previous computation relies on specific choice $F(u) = (\cos u, \sin u)$. On the other hand, it should make sense to speak of the arclength of a curve, or the surface area of a surface, without making reference to a specific choice of parametrization. In particular, if you chose to describe the circle in a different way than I did, we should still be able to agree on the length of the curve.

With this in mind, we make a new definition:

Definition 9.11. A set $A \subset \mathbb{R}^n$ is called a *k-cell* if there exists an open set $U \subset \mathbb{R}^k$ containing the unit cube $I \subset \mathbb{R}^k$ and a one-to-one C^1 function $\varphi : U \to \mathbb{R}^n$ such that

(i) $A = \varphi(I)$, and
(ii) $\text{rank}[\varphi'(u)] = k$ for all $u \in I$.

The restriction $\varphi|_I$ is called a *parametrization* of A.

The surface area of a k-cell, as defined above, is independent of parametrization:

Theorem 9.9. *If φ and ψ are two parametrizations of the same k-cell $A \subset \mathbb{R}^n$, then*

$$\int_I [\det(\varphi')^t\varphi]^{1/2} = \int_I [\det(\psi')^t\psi]^{1/2}.$$

Sketch of proof. We set $T = \psi^{-1} \circ \varphi$. Using the fact that φ' and ψ' are full rank, we can use the inverse function theorem to prove that T is C^1 invertible. Then we have $\varphi = \psi \circ T$, and so

$$(\varphi')^t(\varphi') = [(\psi \circ T)']^t[\psi \circ T]' = (T')^t[\psi' \circ T]^t[\psi' \circ T]T'.$$

This implies

$$[\det(\varphi')^t\varphi']^{1/2} = [\det(\psi' \circ T)^t\psi' \circ T]^{1/2}|\det T'|.$$

The result now follows from the change of variables formula. □

The surfaces in $\mathbb{R}^n$ that we are interested in studying may be obtained by taking nonoverlapping unions of k-cells.

Definition 9.12 (Manifold)**.** A set M is called a *compact k-dimensional manifold* if it can be written as the union of a finite number of nonoverlapping k-cells.

Warning 9.10. *This is not the standard definition of a manifold, but it will be convenient for our purposes. The 'real' definition of a k-manifold is a set such that at any point p, one can find an open set $p \in U \subset \mathbb{R}^n$ such that $U \cap M$ is a 'k-dimensional patch'.*

We have actually encountered manifolds in disguise before. In particular, when we studied Lagrange multipliers, we dealt with sets of the form

$$M = \{x \in \mathbb{R}^n : g(x) = 0 \quad \text{and} \quad \nabla g(x) \neq 0\}$$

for suitable g. Such sets are manifolds. Moreover, the proof that showed the existence of tangent planes carries over to the present setting; in particular, a k-dimensional manifold has a k-dimensional tangent plane at each point:

Definition 9.13. Let $p \in M$. We say that v is in the tangent plane of M at p if there exists $\delta > 0$ and a differentiable function $\gamma : (-\delta, \delta) \to M$ such that $\gamma(0) = p$ and $\gamma'(0) = v$.

Similar to the proof given in the section on Lagrange multipliers, we can show that T_pM is a k-dimensional vector space for each $p \in M$.

Other examples of manifolds include familiar sets like circles or spheres, cylinders, paraboloids, hyperboloids, and so on.

Definition 9.14 (Area of a manifold)**.** Let $M = \cup_{i=1}^r A_i$ is a compact k-manifold, where A_i are k-cells. Then we define the *k-dimensional area of M* by

$$a(M) = \sum_{i=1}^r a(A_i).$$

Remark 9.11. To see that the area of M is well-defined, one must verify that if $M = \cup_{i=1}^m B_i$ is another decomposition of M into k-cells, then $\sum_{i=1}^m a(B_i) = \sum_{i=1}^r a(A_i)$. This can be done using an argument similar to the one appearing in Theorem 9.9.

Our next goal will be to define the so-called *surface area form* for a k-manifold. Before we can do this, we need a bit more language associated to manifolds:

Definition 9.15 (Coordinate patches; orientation)**.** Let M be a k-manifold.

- A *coordinate patch* is an injective, C^1 mapping $\varphi : U \to M$, where $U \subset \mathbb{R}^n$ is open, such that $d\varphi_u$ has rank k for all $u \in U$.
- An *atlas* for M is a collection of coordinate patches $\psi_i : U_i \to M$ such that $M \subset \cup_i \psi_i(U_i)$.
- An *orientation* for M is an atlas $\{\psi_i\}$ with the following property: if $\psi_i(U_i) \cap \psi_j(U_j) \neq \emptyset$, then the 'change of coordinates mapping'
$$T_{ij} = \psi_j^{-1} \circ \psi_i$$
satisfies $\det T'_{ij} > 0$ on its domain. The pair $(M, \{\psi_i\})$ is called an *oriented manifold.*
- A coordinate patch $\varphi : U \to (M, \{\psi_i\})$ is *orientation-preserving* if it overlaps positively with each ψ_i (in the sense above); it is *orientation-reversing* if it overlaps negatively with each ψ_i.

Remark 9.12. The condition that $d\varphi_u$ has rank k guarantees that $\det[(\varphi'(u))^t \varphi'(u)] \neq 0$ for each u.

We now introduce the surface area form; then we'll prove that it actually does the job it's supposed to do.

Definition 9.16 (Surface area form)**.** Let M be an oriented k-manifold. Let $x \in M$ and let $I = (i_1, \ldots, i_k)$. Choosing $\varphi : U \to M$ be an orientation-preserving patch such that $x = \phi(u)$, we define
$$n_I(x) = \frac{\det[\varphi'_I(u)]}{[\det(\varphi'(u))^t \varphi'(u)]^{1/2}},$$
where φ'_I is the $k \times k$ submatrix of φ' obtained by choosing rows $i_1, \ldots, i_k$. We then define the *surface area form* on M by
$$dA = \sum_{I \text{ increasing}} n_I \, dx_I.$$

As usual, we need to check that the definition above does not depend on the choice of coordinate patch (so that the surface area form is actually well-defined). For this (and the next theorem) we will need a lemma:

Lemma 9.13. *Suppose P is a k-dimensional parallelepiped in $\mathbb{R}^n$ spanned by $a_1, \ldots, a_k$. Writing $A = [a_1 \; a_2 \; \ldots \; a_k]$, we have*
$$a(P) = [\det(A^t A)]^{1/2} = \Big[\sum_{I \text{ increasing}} (\det A_I)^2 \Big]^{1/2},$$
where A_I is the $k \times k$ matrix obtained by choosing rows $i_1, \ldots, i_k$ from A.

Sketch of proof. We begin by quoting a theorem known as the *Binet–Cauchy product formula*: given a $k \times n$ matrix A and an $n \times k$ matrix B,
$$\det AB = \sum_{I \text{ increasing}} [\det A_I^t][\det B_I].$$
To prove this, fix A and consider the map $B \mapsto \det AB$. This is an alternating k-multilinear function on $\mathbb{R}^n$. Using Theorem 9.3 and the fact that $dx_I(b^1, \ldots, b^k) = \det B_I$, one can deduce that the coefficients α_I appearing in the represention of $B \mapsto \det AB$ are given by $\det A_I^t$, which implies the result.

In the special case $A = B$, we obtain

$$\det(A^t A) = \sum_{I \text{ increasing}} (\det A_I)^2.$$

Since $a(P) = [\det(A^t A)]^{1/2}$, the result follows. □

Proof that n_I is well-defined. We now suppose that $\psi : V \to M$ is another orientation-preserving coordinate patch with $x = \psi(v)$, and that $\varphi(U) \cap \psi(V)$ intersect. Then we set $T = \psi^{-1} \circ \varphi$ and consider increasing $J = (j_1, \dots, j_k)$. By the chain rule and the fact that $\varphi = \psi \circ T$,

$$(\psi \circ T)'_J = \psi'_J \circ T \cdot T' \implies \det \varphi'_J = \det[\psi'_J \circ T] \cdot \det[T'].$$

Thus, using $\det T' > 0$, we deduce

$$\begin{aligned} \frac{\det \varphi'_I(u)}{[\sum_{J \text{ increasing}} [\det \varphi'_J(u)]^2]^{1/2}} &= \frac{\det \psi'_J(v) \det T'(u)}{[\sum_{J \text{ increasing}} [\det \psi'_J(v) \det T'(u)]^2]^{1/2}} \\ &= \frac{\det \psi'_J(v)}{[\sum_{J \text{ increasing}} [\det \psi'_J(v)]^2]^{1/2}}, \end{aligned}$$

i.e. (by the previous lemma)

$$\frac{\det[\varphi_I(u)]}{[\det(\varphi'(u))^t \varphi'(u)]^{1/2}} = \frac{\det[\psi_I(v)]}{[\det(\psi'(v)^t \psi'(v)]^{1/2}}.$$

But this shows that the definition for n_I in Definition 9.16 using the φ coordinates agrees with the one using the ψ coordinates; in particular, n_I is well-defined. □

Example 9.19 (Outer normal vector)**.** Let M be an oriented smooth $(n-1)$-dimensional manifold in $\mathbb{R}^n$. Define $N : M \to \mathbb{R}^n$ as follows: for $x \in M$, choose an orientation-preserving coordinate patch $\varphi : U \to M$ with $x = \varphi(u)$. Then we let the i^{th} component n_i of N is defined to be $(-1)^{i-1} n_{I_i}$, where I_i is the $(n-1)$-tuple with i removed. (For example, if $n = 3$, then $n_2 = -n_{I_2}$, where $I_2 = (1, 3)$.)

The surface area form can then be written

$$dA = \sum_{i=1}^{n} (-1)^{i-1} n_i dx_{I_i}.$$

We call N a unit *normal* vector, because N is orthogonal to the vectors $\frac{\partial \varphi}{\partial u_j}$ for $j = 1, \dots, n-1$ (and these vectors form a basis for the tangent space of M). Let's verify this in the case $n = 3$, and take $j = 1$ (we use u, v instead of u_1, u_2). Then the inner product of N and $\frac{\partial \varphi}{\partial u_1}$ is given by

$$\begin{aligned} &\det \begin{bmatrix} \varphi^2_u & \varphi^2_v \\ \varphi^3_u & \varphi^3_v \end{bmatrix} \varphi^1_u - \det \begin{bmatrix} \varphi^1_u & \varphi^1_v \\ \varphi^3_u & \varphi^3_v \end{bmatrix} \varphi^2_u + \det \begin{bmatrix} \varphi^1_u & \varphi^1_v \\ \varphi^2_u & \varphi^2_v \end{bmatrix} \varphi^3_u \\ &\quad = \varphi^1_u \varphi^2_u \varphi^3_v - \varphi^1_u \varphi^2_v \varphi^3_u - \varphi^1_u \varphi^2_u \varphi^3_v + \varphi^1_v \varphi^2_u \varphi^3_u + \varphi^1_u \varphi^2_v \varphi^3_u - \varphi^1_v \varphi^2_u \varphi^3_u \\ &\quad = 0. \end{aligned}$$

The following theorem shows the role of the surface area form:

Theorem 9.14 (Role of the surface area form)**.** *Let M be an oriented k-manifold in $\mathbb{R}^n$ with surface area form dA (cf. Definition 9.16). Suppose $\varphi : Q \to M$ is* *the restriction of an orientation-preserving coordinate patch to an interval $Q \subset \mathbb{R}^k$. Then*

$$a(\varphi) = \int_{\varphi(Q)} dA.$$

Proof. Beginning with the definition of area of a surface patch, applying the lemma above and the definition of n_I, and recalling the definition of the integral of a k-form and of the form dA, we obtain

$$\begin{aligned} a(\varphi) &= \int_Q [\det \varphi'^t \, \varphi']^{1/2} \\ &= \int_Q \frac{\det \varphi'^t \varphi'}{[\det \varphi'^t \, \varphi']^{1/2}} \\ &= \int_Q \frac{\sum_{I \text{ increasing}} (\det \varphi'_I)^2}{[\det \varphi'^t \, \varphi']^{1/2}} \\ &= \int_Q \frac{\sum_{I \text{ increasing}} [\det \varphi'^t \varphi']^{1/2} n_I \circ \varphi \, \det \varphi'_I}{[\det \varphi'^t \, \varphi']^{1/2}} \\ &= \int_Q \sum_{I \text{ increasing}} n_I \circ \varphi \, dx_I(D_1\varphi, \dots, D_k\varphi) \\ &= \int_{\varphi(Q)} \sum_{I \text{ increasing}} n_I dx_I = \int_{\varphi(Q)} dA. \end{aligned}$$

□

Given a compact manifold of the form $M = \cup_{i=1}^r A_i$, where A_i are k-cells, we essentially define the area of M as the sum of the area of each A_i. In particular, we suppose that this decomposition is *oriented*, in the sense that each A_i has a parametrization $\varphi_i : Q_i \to A_i$ that extends to an orientation-preserving coordinate match for M defined on an open set containing Q_i. We then wish to define

$$a(M) = \sum_{i=1}^r \int_{\varphi_i(Q_i)} dA.$$

More generally, given any continuous k-form α on M (actually, an open set containing M), we would like to define

$$\int_M \alpha = \sum_{i=1}^r \int_{\varphi_i(Q_i)} \alpha.$$

(Then, in particular, we get the formula $a(M) = \int_M dA$). To make these definitions, we need to be sure that the result is independent of the specific parametrizations used. We have proved similar things above (you use the chain rule, the change of variables formula, and a positive Jacobian assumption), so we will not prove this here. Instead, let's try to see how we can actually use all of this abstract machinery to compute something.

Example 9.20. Let M be the unit sphere in $\mathbb{R}^3$. We parametrize M using the spherical coordinates

$$x = \sin\alpha \cos\theta, \quad y = \sin\alpha \sin\theta, \quad z = \cos\alpha,$$

i.e $\varphi(\alpha,\theta) = (\sin\alpha\cos\theta, \sin\alpha\sin\theta, \cos\alpha)^t$. Then a computation shows

$$\varphi' = \begin{bmatrix} \cos\alpha\cos\theta & -\sin\alpha\sin\theta \\ \cos\alpha\sin\theta & \sin\alpha\cos\theta \\ -\sin\alpha & 0 \end{bmatrix}, \quad \text{so that} \quad [\det\varphi'^t\varphi']^{1/2} = \sin\alpha.$$

To compute dA, we need to compute each n_I. We find

$$\begin{aligned} I = (1,2) &\implies \det\varphi'_I = \cos\alpha\sin\alpha, \\ I = (2,3) &\implies \det\varphi'_I = \sin^2\alpha\cos\theta, \\ I = (1,3) &\implies \det\varphi'_I = -\sin^2\alpha\sin\theta, \end{aligned}$$

so that

$$n_{(1,2)} = \cos\alpha = z, \quad n_{(2,3)} = \sin\alpha\cos\theta = x, \quad n_{(1,3)} = -\sin\alpha\sin\theta = -y.$$

It follows that

$$dA = x\,dy\wedge dz + y\,dz\wedge dx + z\,dx\wedge\ dy.$$

Now let $Q = \{(\theta,\alpha) : \theta\in[0,2\pi],\ \alpha\in[0,\pi]\}$. Then

$$\int_{\varphi(Q)} z\,dx\wedge\ dy = \int_Q \cos\alpha\,dx\wedge\ dy(D_\alpha\varphi, D_\theta\varphi) = \int_Q \cos^2\alpha\sin\alpha = \tfrac{4\pi}{3}.$$

Similarly,

$$\begin{aligned} \int_{\varphi(Q)} y\,dz\wedge\ dx &= \int_Q \sin^3\alpha\sin^2\theta = \tfrac{4\pi}{3}, \\ \int_{\varphi(Q)} z\,dx\wedge dy &= \int_Q \cos^2\alpha\sin\alpha = \tfrac{4\pi}{3}. \end{aligned}$$

Thus

$$a(M) = \int_{\varphi(Q)} dA = 3\cdot\tfrac{4\pi}{3} = 4\pi.$$

9.3. **Stokes' Theorem and the Classical Theorems of Vector Calculus.** We turn to our attention to a deep theorem known as *Stokes' Theorem.* Before we can even state this result precisely, we need to introduce a few new notions.

Definition 9.17. A compact oriented smooth *k-manifold with boundary* is a compact region V in an oriented k-manifold $M\subset\mathbb{R}^n$ such that the boundary ∂V is a smooth compact $(k-1)$-manifold.

The *positive* orientation on ∂V is defined as follows: Given $p\in\partial V$, we let $\Phi : U\to M$ be a coordinate patch with $p\in\Phi(U)$, $\Phi^{-1}(\partial V)\subset\mathbb{R}^k$, and $\Phi^{-1}(\text{int}V)\subset\{x\in\mathbb{R}^k : x_k > 0\}$. We choose Φ to be orientation-preserving for k even, orientation-reserving for k odd. Then $\varphi|_{U\cap\mathbb{R}^{k-1}}$ is a coordinate patch for ∂V.

If we choose patches $\Phi_1,\ldots,\Phi_m$ that cover ∂V, their restrictions $\varphi_1,\ldots,\varphi_m$ form an orientation for ∂V (which is the positive orientation, by definition).

Example 9.21. If we consider the unit circle as the boundary of the unit ball in $\mathbb{R}^2$, this construction yields the counterclockwise orientation for the circle.

Here is the theorem we will prove in the next section (modulo some facts about manifolds):

Theorem 9.15 (Stokes' Theorem). *Let V be an oriented compact smooth k-manifold with boundary in the oriented smooth k-manifold $M \subset \mathbb{R}^n$. If ∂V has the positive orientation and α is a C^1 differential $(k-1)$-form on an open set containing V, then*

$$\int_V d\alpha = \int_{\partial V} \alpha.$$

For now, let us see how many of the classical theorems from vector calculus actually follow from Theorem 9.15.

Example 9.22 (Green's Theorem in $n = 2$). In the case $n = 2$, V is a region in $\mathbb{R}^2$ and ∂V is a curve in the plane. Then 'α' should be a 1-form on $\mathbb{R}^2$, so that

$$\alpha = P\,dx + Q\,dy \quad \text{for some} \quad P, Q.$$

Then $d\alpha$ is the 2-form given by

$$d\alpha = [\tfrac{\partial Q}{\partial x} - \tfrac{\partial P}{\partial y}]\,dx \wedge dy,$$

and so Stokes' Theorem becomes the familiar Green's Theorem:

$$\int_V [\tfrac{\partial Q}{\partial x} - \tfrac{\partial P}{\partial y}]\,dx\,dy = \int_{\partial V} P\,dx + Q\,dy.$$

We next consider the divergence theorem; here we can handle the n-dimensional case.

Definition 9.18. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be C^1 (we call F a *vector field*). Denote the components by $F_1, \dots, F_n$. Then the *divergence* of F is the function $\operatorname{div} F : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\operatorname{div} F = \sum_{i=1}^n \tfrac{\partial F_i}{\partial x_i}.$$

Theorem 9.16 (Divergence Theorem). *Let F be a C^1 vector field defined on a neighborhood of a compact n-manifold with boundary $V \subset \mathbb{R}^n$. Then*

$$\int_V \mathit{div}\, F = \int_{\partial V} F \cdot N\, dA,$$

where N is the outer normal (see Example 9.19) and dA is the surface area form of the positively-oriented boundary ∂V.

Proof. We want to write the left-hand side as $\int_V d\alpha$ for a suitable $n-1$ form α, and the right-hand side as $\int_{\partial V} \alpha$, since then the result follows from Stokes' Theorem.

To this end, we set

$$\alpha := \sum_{i=1}^n (-1)^{i-1} F_i dx_{I_i}, \quad \text{where} \quad I_i = (1, \dots, i-1, i+1, \dots, n).$$

Then

$$\begin{aligned} d\alpha &= \sum_{i=1}^n (-1)^{i-1} \big[\sum_{j=1}^n \tfrac{\partial F_i}{\partial x_j} dx_j\big] \wedge dx_{I_i} \\ &= \sum_{i=1}^n (-1)^{i-1} \tfrac{\partial F_i}{\partial x_i}\, dx_i \wedge dx_{I_i} \\ &= \sum_{i=1}^n \tfrac{\partial F_i}{\partial x_i}\, dx_1 \wedge \cdots \wedge dx_n = \operatorname{div} F\, dx. \end{aligned}$$

This shows

$$\int_V \operatorname{div} F = \int_V d\alpha.$$

Now let's work on the right-hand side. Now, in the definition of $\int_{\partial V} F \cdot N \, dA$, we sum over coordinate patches for ∂V. On each such patch, the integral is defined by evaluating $F \cdot N \, dA$ on the column vectors $D\varphi := (D_1\varphi(\cdot), \dots, D_{n-1}\varphi(\cdot))$. Thus, using Example 9.19 (recalling $(-1)^{j-1}n_j = n_{I_j}$), Definition 9.16, and Lemma 9.13, we have that

$$\begin{aligned} F \cdot N(\varphi) \, dA(D\varphi) &= F \cdot N(\varphi) \sum_{j=1}^{n} n_{I_j}(\varphi) dx_{I_j}(D\varphi) \\ &= F \cdot N(\varphi) \sum_{j=1}^{n} \frac{\det[\varphi'_{I_j}]^2}{[\det(\varphi')^t\varphi']^{1/2}} \\ &= [\det(\varphi')^t\varphi']^{1/2} \sum_{i=1}^{n} (-1)^{i-1} F_i(\varphi) n_{I_i}(\varphi) \\ &= \sum_{i=1}^{n} (-1)^{i-1} F_i(\varphi) \det[\varphi'_{I_i}] \\ &= \sum_{i=1}^{n} (-1)^{i-1} F_i(\varphi) dx_{I_i}(D\varphi) = \alpha(D\varphi). \end{aligned}$$

Thus we conclude

$$\int_{\partial V} F \cdot N \, dA = \int_{\partial V} \alpha.$$

The result now follows from Stokes' Theorem. □

Next, we'll establish the version of Stokes' Theorem that you likely encountered in your multivariable calculus class.

Definition 9.19 (Curl)**.** Let $F : \mathbb{R}^3 \to \mathbb{R}^3$ be continuously differentiable. The *curl* of F, denoted $\operatorname{curl} F$ or $\nabla \times F$, is given by

$$\nabla \times F = (\tfrac{\partial F_3}{\partial x_2} - \tfrac{\partial F_2}{\partial x_3}, \tfrac{\partial F_1}{\partial x_3} - \tfrac{\partial F_3}{\partial x_1}, \tfrac{\partial F_2}{\partial x_1} - \tfrac{\partial F_1}{\partial x_2}).$$

Theorem 9.17 (Stokes' Theorem, Familiar Version)**.** *Let D be an oriented, compact 2-manifold with boundary in $\mathbb{R}^3$. Let N be the unit normal vector on D and T the unit tangent on ∂D. If F is a C^1 vector field on an open set containing D, then*

$$\int_D [\nabla \times F] \cdot N \, dA = \int_{\partial D} F \cdot T \, ds.$$

Proof. Recalling Example 9.13, we can write

$$\int_{\partial D} F \cdot T \, ds = \int_{\partial D} \omega,$$

where

$$\omega = F_1 \, dx + F_2 \, dy + F_3 \, dz.$$

By Stokes' Theorem, it then suffices to show that

$$\int_D [\nabla \times F] \cdot N \, dA = \int_D d\omega.$$

Recalling Example 9.8, we first note that

$$d\omega = [\nabla \times F]_1\, dy \wedge dz + [\nabla \times F]_2\, dz \wedge\, dx + [\nabla \times F]_3\, dx \wedge dy.$$

Thus, writing n_i for the components of N, the result follows provided we can prove that

$$n_1\, dA = dy \wedge dz, \quad n_2\, dA = dz \wedge dx, \quad n_3\, dA = dx \wedge dy$$

(at least, when restricted to vectors $(D_1\varphi, D_2\varphi)$, where $\varphi : \mathbb{R}^2 \to D$ is a coordinate patch for D). However, we have basically done this computation already (in the proof of the divergence theorem). In particular, using the definition of the surface area form and Lemma 9.13,

$$\begin{aligned} n_1(\varphi) dA(D_1\varphi, D_2\varphi) &= n_{(2,3)}(\varphi) \sum_{I \text{ increasing}} n_I(\varphi) dx_I(D_1\varphi, D_2\varphi) \\ &= \frac{\det \varphi'_{(2,3)}}{[\det \varphi'(u)^t \varphi'(u)]} \sum_{I \text{ increasing}} (\det \varphi'_I)^2 \\ &= \det \varphi'_{(2,3)} \\ &= dy \wedge dz(D_1\varphi, D_2\varphi). \end{aligned}$$

As the computation for n_2, n_3 is similar, the we obtain the result. □

9.4. **Sketch of the proof of Stokes' Theorem.** We recall the statement of Stokes' Theorem:

Theorem 9.18 (Stokes' Theorem)**.** *Let V be an oriented compact smooth k-manifold with boundary in the oriented smooth k-manifold $M \subset \mathbb{R}^n$. If ∂V has the positive orientation and α is a C^1 differential $(k-1)$-form on an open set containing V, then*

$$\int_V d\alpha = \int_{\partial V} \alpha.$$

Let us try a much more modest goal, namely, to prove Stokes' Theorem on the unit cube. We let

$$I^k = [0,1]^k \subset \mathbb{R}^k.$$

Let us take a moment to discuss what the boundary ∂I^k looks like:

Lemma 9.19. *The boundary ∂I^k of the unit cube I^k is the nonoverlapping union of $2k$ $(k-1)$-dimensional faces of the form*

$$I^{k-1}_{i,\sigma} = \{x \in I^k : x_i = \sigma\}, \quad i = 1, \dots, k, \quad \sigma \in \{0,1\}.$$

Each face $I^{k-1}_{i,\sigma}$ is the image of $I^{k-1} \subset \mathbb{R}^{k-1}$ under the map

$$e_{i,\sigma} : I^{k-1} \to \mathbb{R}^k, \quad \textit{where} \quad e_{i,\varepsilon}(x_1, \dots, x_{k-1}) = (x_1, \dots, x_{i-1}, \sigma, x_i, \dots, x_{k-1}).$$

The mapping $e_{i,\sigma}$ provides an orientation for $I^{k-1}_{i,\sigma}$.

Proof. A point is on the boundary if one of the components is 0 or 1. There are then $2k$ faces to the boundary, since for each k you have two options (0 or 1). □

Definition 9.20 (Integrals over ∂I^k and I^k)**.** Given a $(k-1)$-form α over ∂I^k, we define

$$\int_{\partial I^k} \alpha = \sum_{i=1}^k \sum_{\sigma \in \{0,1\}} (-1)^{i+\sigma} \int_{e_{i,\sigma}} \alpha.$$

To integrate k-forms over I^k, we just need to recall that if $\alpha = f dx_1 \wedge \cdots \wedge dx_k$, then we take $\int_{I^k} \alpha = \int_{I^k} f$, where the second integral is the ordinary integral of a real-valued function over a cube in $\mathbb{R}^k$.

Example 9.23 (k=2)**.** I^2 is the unit square. Its boundary consists of four unit line segments. The orientation given by $e_{i,\sigma}$ is such that we integrate counterclockwise around the square. In particular, if ω is a 1-form, then

$$\int_{\partial I^2} \omega = -\int_{e_{1,0}} \omega + \int_{e_{1,1}} \omega - \int_{e_{2,1}} \omega + \int_{e_{2,0}} \omega.$$

Example 9.24 (k=3)**.** I^3 is the unit cube in $\mathbb{R}^3$. Its boundary consists of four unit cubes. If we wish to integrate a 2-form α over the boundary, then, for example, the integral over $e_{1,1}$ (the face where $x_1 = 1$) would come with a plus sign, while the integral over $e_{1,0}$ (the face where $x_1 = 0$) would come with a minus sign.

With the preliminaries in place, we can prove the following:

Theorem 9.20 (Stokes' Theorem for the unit cube)**.** *Let α be a C^1 differential $(k-1)$-form on an open set containing I^k. Then*

$$\int_{I^k} d\alpha = \int_{\partial I^k} \alpha.$$

Proof. Since α is a $(k-1)$-form on $\mathbb{R}^k$, we may write

$$\alpha = \sum_{i=1}^{k} a_i \, dx_{I_i}, \quad \text{where} \quad I_i = (1, \dots, i-1, i+1, \dots, k).$$

We start by computing the right-hand side. In particular, we need to compute the integral over each face $e_{i,\sigma}$. Using Theorem 9.7,

$$\int_{e_{i,\sigma}} \alpha = \int_{e_{i,\sigma}} \sum_{j=1}^{k} a_j dx_{I_j} = \int_{I^{k-1}} \sum_{j=1}^{k} (a_j \circ e_{i,\sigma}) e_{i,\sigma}^*(dx_{I_j}).$$

Now, we claim that

$$e_{i,\sigma}^*(dx_{I_j}) = \begin{cases} dx_{I_i} & i = j \\ 0 & i \neq j. \end{cases}$$

Indeed,

$$e_{i,\sigma}^*(dx_\ell) = \sum_{m=1}^{k-1} \frac{\partial [e_{i,\sigma}]_\ell}{\partial x_m} dx_m.$$

In particular, dx_{I_j} contains dx_i (which occurs if $i \neq j$) then the pullback is zero (since the i^{th} component of $e_{i,\sigma}$ is constant). If instead $i = j$ then dx_{I_j} does not contain dx_i and the computation above reduces to $e_{i,\sigma}^*(dx_\ell) = dx_\ell$ for each ℓ, which implies the result.

Continuing from above, we have

$$\int_{e_{i,\sigma}} \alpha = \int_{I^{k-1}} a_i \circ e_{i,\sigma} \, dx_{I_i} = \int_{e_{i,\sigma}} a_i \, dx_{I_i},$$

so that

$$\begin{aligned}\int_{\partial I^k} \alpha &= \sum_{i=1}^{k} \sum_{\sigma\in\{0,1\}} (-1)^{i+\sigma} \int_{e_{i,\sigma}} \alpha \\ &= \sum_{i=1}^{k} (-1)^{i-1} \biggl[\int_{e_{i,1}} a_i \, dx_{I_i} - \int_{e_{i,0}} a_i \, dx_{I_i} \biggr].\end{aligned} \tag{9.2}$$

We turn to the left-hand side of the identity. We first observe that

$$d\alpha = \sum_{i=1}^{k} da_i dx_{I_i} = \sum_{i=1}^{k}\sum_{j=1}^{k} \tfrac{\partial a_i}{\partial x_j} dx_j \wedge dx_{I_i} = \biggl[\sum_{i=1}^{k} (-1)^{i-1} \tfrac{\partial a_i}{\partial x_i}\biggr] dx^k,$$

where we write dx^k to denote $dx_1 \wedge \cdots \wedge dx_k$.

We now apply Fubini's Theorem and the one-dimensional fundamental theorem of calculus to obtain

$$\begin{aligned}&\int \tfrac{\partial a_i}{\partial x_i} \, dx^k \\ &\quad = \int \biggl[\int_0^1 \tfrac{\partial a_i}{\partial x_i} dx_i \biggr] dx_{I_i} \\ &\quad = \int_{I^{k-1}} [a_i(x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_k) - a_i(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_k) \, dx_{I_i} \\ &\quad = \int_{e_{i,1}} a_i \, dx_{I_i} - \int_{e_{i,0}} a_i \, dx_{I_i}.\end{aligned}$$

Thus

$$\int_{I^k} d\alpha = \sum_{i=1}^{k} (-1)^{i-1} \biggl[\int_{e_{i,1}} a_i \, dx_{I_i} - \int_{e_{i,0}} a_i \, dx_{I_i} \biggr],$$

which agrees with the expression appearing in (9.2). □

We have now verified the conclusion of Stokes' Theorem on the unit cube. The proof also revealed that the key is simply the fundamental theorem of calculus, and indeed, Stokes' Theorem should be viewed as an extreme generalization of this fundamental result. In what follows, we will only mention the main steps that one could take to obtain Stokes' Theorem in the generality appearing in Theorem 9.18. The strategy is as follows:

- Establish Stokes' Theorem for the unit cube.
- Establish Stokes' Theorem for an oriented k-cell in a smooth k-manifold in $\mathbb{R}^n$. This relies primarily on the properties of pullback and differentials obtained previously, which allow us to pull the computation back to the unit cube.
- Establish Stokes' Theorem for regions that can be obtained by piecing together oriented k-cells (these are called *cellulated regions*). By orienting things properly, integrals over interior faces will cancel in pairs.
- Finally, extend Stokes' Theorem to the setting of k-manifold with boundary by piecing together cellulated regions.