Chapter 1
A Brief Historical Introduction

The first problems belonging properly to mathematical analysis arose during fifth century BCE, when Greek mathematicians became interested in the properties of various curved shapes and surfaces. The problem of squaring a circle (that is, constructing a square of the same area as a given circle with only a compass and straightedge) was well known by the second half of the century, and Hippias had already discovered a curve called the *quadratix* during an attempt at a solution. Hippocrates was also active during the second half of the fifth century BCE, and he defined the areas of several regions bound by curves ("Hippocratic lunes").

The discovery of the fundamental tool of mathematical analysis, approximating the unknown value with arbitrary precision, is due to Eudoxus (408–355 BCE). Eudoxus was one of the most original thinkers in the history of mathematics. His discoveries were immediately appreciated by Greek mathematicians; Euclid (around 300 BCE) dedicated an entire book (the fifth) of his *Elements* [3] to Eudoxus's theory of proportions of magnitudes. Eudoxus invented the method of *exhaustion* as well, and used it to prove that the volume of a pyramid is one-third that of a prism with the same base and height. This beautiful proof can be found in book XII of the *Elements* as the fifth theorem.

The method of exhaustion is based on the fact that if we take away at least one-half of a quantity, then take away at least one-half of the remainder, and continue this process, then for every given value v, sooner or later we will arrive at a value smaller than v. One variant of this principle is nowadays called the *axiom of Archimedes*, even though Archimedes admits, in his book *On the Sphere and Cylinder*, that mathematicians before him had already stated this property (and in the form above, it appeared as the first theorem in book X of Euclid's *Elements*). Book XII of the *Elements* gives many applications of the method of exhaustion. It is worth noting the first application, which states that *the ratio of the area of two circles is the same as the ratio of the areas of the squares whose sides are the circles' diameters*. The proof uses the fact that the ratio between the areas of two similar polygons is the same as the ratio of the squares of the corresponding sides (and this fact is proved by Euclid

M. Laczkovich, V.T. Sós, *Real Analysis*, Undergraduate Texts in Mathematics, DOI 10.1007/978-1-4939-2766-1_1

in his previous books in detail). Consider a circle C. A square inscribed in the circle contains more than half of the area of the circle, since it is equal to half of the area of the square circumscribed about the circle. A regular octagon inscribed in the circle contains more than half of the remaining area of the circle (as seen in Figure 1.1). This is because the octagon is larger than the square by four isosceles triangles, and each isosceles triangle is larger than half of the corresponding slice of the circle, since it can be inscribed in a rectangle that is twice the triangle. A similar argument tells us that a regular 16-gon covers at least one-half of the circle not covered by the octagon, and so on. The method of exhaustion (the axiom of Archimedes) then tells us that we can inscribe a regular polygon in the circle C such that the areas differ by less than any previously fixed number.

To finish the proof, it is easiest to introduce modern notation. Consider two circles, C_1 and C_2, and let a_i and d_i denote the areas and the diameters of the circle C_i $(i = 1,2)$. We want to show that $a_1/a_2 = d_1^2/d_2^2$. Suppose that this is not true. Then a_1/a_2 is either larger or smaller than d_1^2/d_2^2. It suffices to consider the case $a_1/a_2 > d_1^2/d_2^2$, since in the second case, $a_2/a_1 > d_2^2/d_1^2$, so we can interchange the roles of the two circles to return to the first case. Now if $a_1/a_2 > d_1^2/d_2^2$, then the value

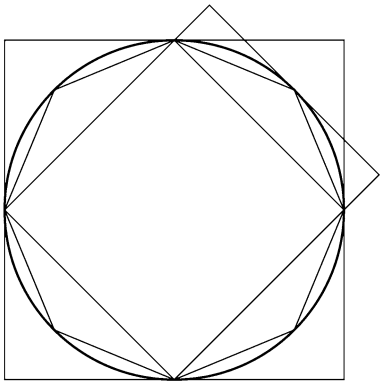

Fig. 1.1

$$\delta = \frac{a_1}{a_2} - \frac{d_1^2}{d_2^2}$$

is positive. Inscribe a regular polygon P_1 into C_1 that differs from the area of C_1 by less than $\delta \cdot a_2$. If a regular polygon P_2 similar to P_1 is inscribed in C_2, then the ratio of the areas of P_1 and P_2 is equal to the ratio of the squares of the corresponding sides, which in turn is equal to d_1^2/d_2^2 (which was shown precisely by Euclid). If the area of P_i is p_i $(i = 1,2)$, then

$$\frac{a_1}{a_2} - \delta = \frac{d_1^2}{d_2^2} = \frac{p_1}{p_2} > \frac{a_1 - \delta \cdot a_2}{a_2} = \frac{a_1}{a_2} - \delta,$$

which is a contradiction.

Today, we would express the above theorem by saying that the area of a circle is a constant times the square of the diameter of a circle. This constant was determined by Archimedes. In his work *Measurement of a Circle*, he proves that the area of a circle is equal to the area of the right triangle whose side lengths are the radius of the circle and the circumference of the circle. With modern notation (and using the theorem above), this is none other than the formula πr^2, where π is one-half of the circumference of the unit circle.

Archimedes (287–212 BCE) ranks as one of the greatest mathematicians of all time, but is without question the greatest mathematician of antiquity. Although the greater part of his works is lost, a substantial corpus has survived. In his works, he computed the areas of various regions bounded by curves (such as slices of the parabola), determined the surface area and volume of the sphere, the arc length of certain spirals, and studied paraboloids and hyperboloids obtained by rotations. Archimedes also used the method of exhaustion, but he extended the idea by approximating figures not only from the inside, but from the outside as well. Let us see how Archimedes used this method to find the area beneath the parabola. We will use modern notation again.

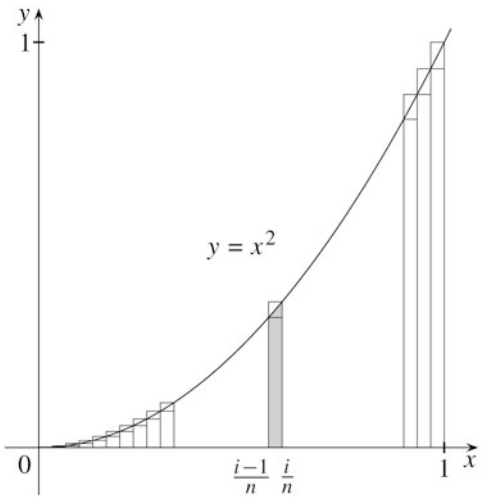

Fig. 1.2

The area of the region beneath the parabola and above $[0,1]$, as seen in Figure 1.2 will be denoted by A. The (shaded) region over the ith interval (for any n and $i \leq n$) can be approximated by rectangles from below and from above, which—with the help of Exercise 2.5(b)—gives

$$A > \frac{1}{n} \cdot \left(\left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{n-1}{n}\right)^2 \right) = \frac{(n-1)\cdot n \cdot (2n-1)}{6n^3} > \frac{1}{3} - \frac{1}{n},$$

and

$$A < \frac{1}{n} \cdot \left(\left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{n}{n}\right)^2 \right) = \frac{n \cdot (n+1) \cdot (2n+1)}{6n^3} < \frac{1}{3} + \frac{1}{n}.$$

It follows that

$$\left| A - \frac{1}{3} \right| < \frac{1}{n}. \tag{1.1}$$

This approximation does not give a precise value for A *for a specific* n. However, the approximation (1.1) *for every* n already shows that the area of A can only be $1/3$.

Indeed, if $A \neq 1/3$ were the case, that is, $|A - 1/3| = \alpha > 0$, then if $n \geq 1/\alpha$, then (1.1) would not be satisfied. Thus the only possibility is that $|A - 1/3| = 0$, so $A = 1/3$.

It was a long time before the work of Archimedes was taken up again and built upon. There are several possible reasons for this: the lack of a proper system of notation, the limitations of the geometric approach, and the fact that the mathematicians of the time had an aversion to questions concerned with infinity and with movement. Whether it is for these reasons or others, analysis as a widely applicable general method or as a branch of science on its own appeared only when European mathematicians of the seventeenth century decided to describe movement and change using the language of mathematics. This description was motivated by problems that occur in everyday life and in physics. Here are some examples:

- Compute the velocity and acceleration of a free-falling object.
- Determine the trajectory of a thrown object. Determine what height the object reaches and where it lands.
- Describe other physical processes, such as the temperature of a cooling object. If we know the temperature at two points in time, can we determine the temperature at every time?
- Construct tangent lines to various curves. How, for example, do we draw the tangent line to a parabola at a given point?
- What is the shape of a suspended rope?
- Solve maximum/minimum problems such as the following: What is the cylinder with the largest volume that can be inscribed in a ball? What path between two points can be traversed in the shortest time if velocity varies based on location? (This last question is inspired by the refraction of light.)
- Find approximate solutions of equations.
- Approximate values of powers (e.g., $2^{\sqrt{3}}$) and trigonometric functions (e.g., $\sin 1^\circ$).

It turned out that these questions are strongly linked with determining area, volume, and arc length, which are also problems arising from the real world. Finally, to solve these problems, mathematicians of the seventeenth century devised a theory, called the *differential calculus*, or in today's terms, differentiation, which had three components.

The first component was the coordinate system, which is attributed to René Descartes (1596–1650), even though such a system had been used by Apollonius (262–190 BCE) when he described conic sections. However, Descartes was the first to point out that the coordinate system can help transform geometric problems into algebraic problems.

Consider the parabola, for example. By definition this is the set of points that lie equally distant from a given point and a given line. This geometric definition can

be transformed into a simple algebraic condition with the help of the coordinate system. Let the given point be $P = (0, p)$, and the given line the horizontal line $y = -p$, where p is a fixed positive number. The distance of the point (x, y) from the point P is $\sqrt{x^2 + (y-p)^2}$, while the distance of the point from the line is $|y + p|$. Thus the point (x, y) is on the parabola if and only if $\sqrt{x^2 + (y-p)^2} = |y + p|$. Squaring both sides gives

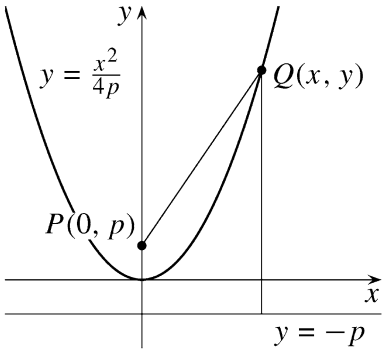

Fig. 1.3

$$x^2 + y^2 - 2py + p^2 = y^2 + 2py + p^2,$$

which we can rearrange to give us that $x^2 = 4py$ or $y = x^2/(4p)$. Thus we get the equation of the parabola, an algebraic condition that describes the points of the parabola precisely: a point (x, y) lies on the parabola if and only if $y = x^2/(4p)$.

The second component of the differential calculus is the concept of a variable quantity. Mathematicians of the seventeenth century considered quantities appearing in physical problems to be variables depending on time, whose values change from one moment to another. They extended this idea to geometric problems as well. Thus every curve was thought of as the path of a continuously moving point. This concept does not interpret the equation $y = x^2/(4p)$ as one in which y depends on x, but as a relation between x and y both depending on time as the point (x, y) traverses the parabola.

The third and most important component of the differential calculus was the notion of a differential of a variable quantity. The essence of this concept is the intuitive notion that every change is the result of the sum of "infinitesimally small" changes. Thus time itself is made up of infinitesimally small time intervals. The differential of the variable quantity x is the infinitesimally small value by which x changes during an infinitesimally small time interval. The differential of x is denoted by dx. Thus the value of x after an infinitesimally small time interval changes to $x + dx$.

How did calculus work? We illustrate the thinking of the wielders of calculus with the help of some simple examples.

The key to solving maximum/minimum problems was the fact that if the variable y reaches its highest value at a point, then $dy = 0$ there (since when a thrown object reaches the highest point along its path, it flies horizontally "for an instant"; thus if the y-coordinate of the object has a maximum, then $dy = 0$ there).

Let us use calculus to determine the largest value of $t - t^2$. Let $x = t - t^2$. Then at the maximum of x, we should have $dx = 0$. Now dx is none other than the change of x as t changes to $t + dt$. From this, we infer that

$$dx = \left[(t + dt) - (t + dt)^2\right] - \left[t - t^2\right] = dt - 2t \cdot dt - (dt)^2 = dt - 2t \cdot dt.$$

In the last step above, the $(dt)^2$ term was "ignored." That is, it was simply left out of the computation, based on the argument that the value $(dt)^2$ is "infinitely smaller" than all the rest of the values appearing in the computation. Thus the condition $dx = 0$ gives $dt - 2t \cdot dt = 0$, and thus after dividing by dt, we obtain $1 - 2t = 0$ and $t = 1/2$. Therefore, $t - t^2$ takes on its largest value at $t = 1/2$. The users of calculus were aware of the fact that the argument above lacks the requisite mathematical precision. At the same time, they were convinced that the argument leads to the correct result.

Now let us see a problem involving the construction of tangent lines to a curve. At the tangent point, the direction of the tangent line should be the same as that of the curve. The direction of the curve at a point (x,y) can be computed by connecting the point by a line to a point "infinitesimally close"; this will tell us the slope of the tangent. After an infinitesimally small change in time, the x-coordinate changes to $x + dx$, while the y-coordinate changes to $y + dy$. The point $(x+dx, y+dy)$ is thus a point of the curve that is "infinitesimally close" to (x,y). The slope of the line connecting the points (x,y) and $(x+dx, y+dy)$ is

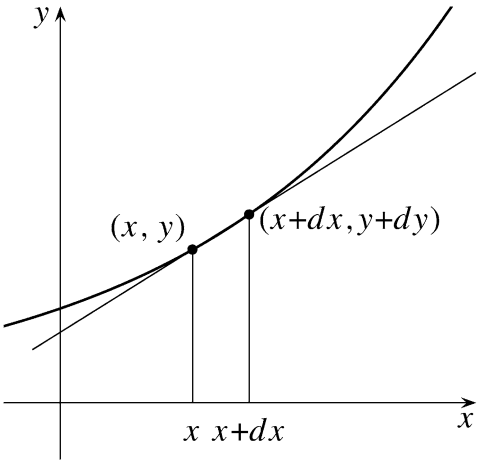

Fig. 1.4

$$\frac{(y+dy)-y}{(x+dx)-x} = \frac{dy}{dx}.$$

This is the quotient of two differentials, thus a *differential quotient*. We get that the slope of the tangent line at the point (x,y) is exactly the differential quotient dy/dx. Computing this quantity is very simple.

Consider, for example, the parabola with equation $y = x^2$. Since the point $(x+dx, y+dy)$ also lies on the parabola, the equation tells us that

$$dy = (y+dy) - y = (x+dx)^2 - x^2 = 2xdx + (dx)^2 = 2xdx,$$

where we "ignore" the term $(dx)^2$ once again. It follows that $dy/dx = 2x$, so at the point (x,y), the slope of the tangent of the parabola given by the equation $y = x^2$ is $2x$. Now consider the point (a,a^2) on the parabola. The slope of the tangent is $2a$ here, so the equation of the tangent is

$$y = 2a \cdot (x-a) + a^2.$$

This intersects the x-axis at $a/2$. Thus—according to mathematicians of the seventeenth century—we can construct the line tangent to the parabola at the point (a,a^2) by connecting the point $(a/2,0)$ to the point (a,a^2).

Finally, let us return to the problem of computing area that we already addressed. Take the parabola given by the equation $y=x^2$ again, and compute the area of the region R that is bordered by the segment $[0,x]$ on the x-axis, the arc of the parabola between the origin and the point (x,x^2), and the segment connecting the points $(x,0)$ and (x,x^2). Let A denote the area in question. Then A itself is a variable quantity. After an infinitesimally small change in time, the value of x changes to $x+dx$, so the region R grows by an infinitesimally narrow "rectangle" with width dx and height x. Thus the change of the area A is $dA = y\cdot dx = x^2\cdot dx$.

Let us look for a variable z whose differential is exactly $x^2\cdot dx$. We saw before that $d(x^2) = 2x\cdot dx$. A similar computation gives $d(x^3) = 3x^2\cdot dx$. Thus the choice $z=x^3/3$ works, so $dz = x^2\cdot dx$. This means that the differentials of the unknowns A and z are the same: $dA = dz$. This, in turn, means that $d(A-z) = dA-dz = 0$, that is, that $A-z$ does not change, so it is constant. If $x=0$, then A and z are both equal to zero, so the constant $A-z$ is zero, so $A=z=x^3/3$. When $x=1$, we obtain Archimedes' result. Again, the users of calculus were convinced that they had arrived at the correct value.

We can see that calculus is a very efficient method, and many different types of problems can be tackled with its help. Calculus, as a stand-alone method, was developed by a list of great mathematicians (Barrow, Cavalieri, Fermat, Kepler, and many others), and was completed—by seventeenth-century standards—by Isaac Newton (1643–1727) and G. W. Leibniz (1646–1716). Mathematicians immediately seized on this method and produced numerous results. By the end of the century, it was time for a large-scale monograph that would summarize what had been obtained thus far. This was L'Hospital's book *Infinitesimal Calculus* (1696), which remained the most important textbook on calculus for nearly all of the next century.

From the beginning, calculus was met with heavy criticism, which was completely justified. For the logic of the method was unclear, since it worked with imprecise definitions, and the arguments themselves were sometimes obscure. The great mathematicians of antiquity were no doubt turning over in their graves. The "proofs" outlined above seem to be convincing, but they leave many questions unanswered. What does an infinitely small quantity really mean? Is such a quantity zero, or is it not? If it is zero, we cannot divide by it in the differential quotient dy/dx. But if it is not zero, then we cannot ignore it in our computations. Such a contradiction in a mathematical concept cannot be overlooked. Nor was the method of computing the maximum very convincing. Even if we accept that the differential at the maximum is zero (although the argument for this already raises questions and eyebrows), we would need the converse of the statement: if the differential is zero, then there is a maximum. However, this is not always the case. We know that $d(x^3) = 3x^2\cdot dx = 0$ if $x=0$, while x^3 clearly does not have a maximum at 0.

An important part of the criticism of calculus is aimed at the contradictions having to do with infinite series. Adding infinitely many numbers together (or more generally, just the concept of infinity) was found to be problematic much earlier, as

the Greek philosopher Zeno[1] had already shown. He illustrated the problem with his famous paradox of Achilles and the tortoise. The paradox states that no matter how much faster Achilles runs than the tortoise, if the tortoise is given a head start, Achilles will never pass it. For Achilles needs some time to catch up to where the tortoise is located when Achilles starts to run. However, once he gets there, the tortoise has already moved away to a further point. Achilles requires some time to get to that further point, but in the meantime, the tortoise travels even farther, and so on. Thus Achilles will never catch up to the tortoise.

Of course, we all know that Achilles will catch up to the tortoise, and we can easily compute when that happens. Suppose that Achilles runs ten yards every second, while the tortoise travels at one yard per second (to make our computations simpler, we let Achilles race against an exceptionally fast tortoise). If the tortoise gets a one-yard advantage, then after x seconds, Achilles will be $10x$ yards from the starting point, while the tortoise will be $1+x$ yards away from that point. Solving the equation $10x = 1+x$, we get that Achilles catches up to the tortoise after $x = 1/9$ seconds.

Zeno knew all of this too; he just wanted to show that an intuitive understanding of summing infinitely many components to produce movement is impossible and leads to contradictions. Zeno's argument expressed in numbers can be summarized as follows: Achilles needs to travel 1 yard to catch up to the starting point of the tortoise, which he does in $1/10$ of a second. During that time, the tortoise moves $1/10$ of a yard. Achilles has to catch up to this point as well, which takes $1/100$ of a second. During that time, the tortoise moves $1/100$ yards, to which Achilles takes $1/1000$ seconds to catch up, and so on. In the end, Achilles needs to travel infinitely many distances, and this requires $(1/10)+(1/100)+(1/1000)+\cdots$ seconds in total. Thus we get that

$$\frac{1}{10}+\frac{1}{100}+\frac{1}{1000}+\cdots=\frac{1}{9}. \tag{1.2}$$

With this, we have reduced Zeno's paradox to the question whether can we put infinitely many segments (distances) next to each other so that we get a bounded segment (finite distance), or in other words, can the sum of infinitely many numbers be finite?

If the terms of an infinite series form a geometric sequence, then its sum can be determined using simple arithmetic—at least formally. Consider the series $1+x+x^2+\cdots$, where x is an arbitrary real number. If $1+x+x^2+\cdots=A$, then

$$A = 1+x\cdot(1+x+x^2+\cdots) = 1+x\cdot A,$$

which, when $x \neq 1$, gives us the equality

$$1+x+x^2+\cdots=\frac{1}{1-x}. \tag{1.3}$$

[1] Zeno (333–262 BCE) Greek philosopher.

If in (1.3) we substitute $x = 1/10$ and subtract 1 from both sides, then we get (1.2). In the special case $x = 1/2$, we get the identity $1 + 1/2 + 1/4 + \cdots = 2$, which is immediate from Figure 1.5 as well.

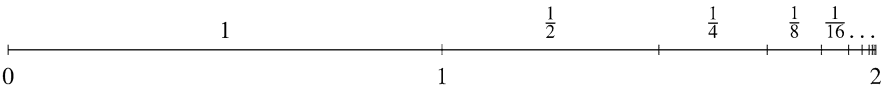

Fig. 1.5

However, the identity (1.3) can produce strange results as well. If we substitute $x = -1$ in (1.3), then we get that

$$1 - 1 + 1 - 1 + \cdots = \tfrac{1}{2}. \tag{1.4}$$

This result is strange from at least two viewpoints. On one hand, we get a fraction as a result of adding integers. On the other hand, if we put parentheses around pairs of numbers, the result is

$$(1-1) + (1-1) + \cdots = 0 + 0 + \cdots = 0.$$

In fact, if we begin the parentheses elsewhere, we get that

$$1 - (1-1) - (1-1) - \ldots = 1 - 0 - 0 - \ldots = 1.$$

Thus three different numbers qualify as the sum of the series $1 - 1 + 1 - 1 + \cdots$: $1/2$, 0, and 1.

We run into a different problem if we seek the sum of the series $1 + 1 + 1 + \cdots$. If its value is y, then

$$1 + y = 1 + (1 + 1 + 1 + \cdots) = 1 + 1 + 1 + \cdots = y.$$

Such a number y cannot exist, however. We could say that the sum must be ∞, but can we exclude $-\infty$? We could argue that we are adding positive terms so cannot get a negative number, but are we so sure? If we substitute $x = 2$ in (1.3), then we get that

$$1 + 2 + 4 + \cdots = -1, \tag{1.5}$$

and so, it would seem, the sum of positive numbers can actually be negative.

These strange, impossible, and even contradictory results were the subject of many arguments up until the beginning of the nineteenth century. To resolve the contradictions, some fantastic ideas were born: to justify the equality $1 + 2 + 4 + \cdots = -1$, there were some who believed that the numbers "start over" and that after infinity, the negative numbers follow once again.

The arguments surrounding calculus lasted until the end of the nineteenth century, and they often shifted to philosophical arguments. Berkeley maintained that

the statements of calculus are no more scientific than religious beliefs, while Hegel argued that the problems with calculus could be solved only through philosophical arguments.

These problems were eventually solved by mathematicians by replacing the intuitive but hazy definitions leading to contradictions with precisely defined mathematical concepts. The variable quantities were substituted by functions, and the differential quotients by derivatives. The sums of infinite series were defined by Augustin Cauchy (1789–1857) as the limit of partial sums.[2] As a result of this clarifying process—in which Cauchy, Karl Weierstrass (1815–1897), and Richard Dedekind (1831–1916) played the most important roles—by the end of the nineteenth century, the theory of **differentiation and integration** (or **analysis**, for short) reached the logical clarity that mathematics requires.

The creation of the precise theory of analysis was one of the greatest intellectual accomplishments of modern Western culture. We should not be surprised when this theory—especially its basics, and first of all its central concept, the limit—is found to be difficult. We want to facilitate the mastery of this concept as much as possible, which is why we begin with limits of sequences. But before everything else, we must familiarize ourselves with the foundations on which this branch of mathematics, analysis, is based.

[2] See the details of this in Chapter 7.