1

Introduction

By 1850, most mathematicians thought they understood calculus. Real progress was being made in extending the tools of calculus to complex numbers and spaces of higher dimensions. Equipped with appropriate generalizations of Fourier series, solutions to partial differential equations were being found. Cauchy's insights had been assimilated, and the concepts that had been unclear during his pioneering work of the 1820s, concepts such as uniform convergence and uniform continuity, were coming to be understood. There was reason to feel confident.

One of the small, nagging problems that remained was the question of the convergence of the Fourier series expansion. When does it converge? When it does, can we be certain that it converges to the original function from which the Fourier coefficients were derived? In 1829, Peter Gustav Lejeune Dirichlet had proven that as long as a function is piecewise monotonic on a closed and bounded interval, the Fourier series converges to the original function. Dirichlet believed that functions did not have to be piecewise monotonic in order for the Fourier series to converge to the original function, but neither he nor anyone else had been able to weaken this assumption.

In the early 1850s, Bernard Riemann, a young protegé of Dirichlet and a student of Gauss, would make substantial progress in extending our understanding of trigonometric series. In so doing, the certainties of calculus would come into question. Over the next 60 years, five big questions would emerge and be answered. The answers would be totally unexpected. They would forever change the nature of analysis.

1. **When does a function have a Fourier series expansion that converges to that function?**
2. **What is integration?**
3. **What is the relationship between integration and differentiation?**

4. What is the relationship between continuity and differentiability?
5. When can an infinite series be integrated by integrating each term?

This book is devoted to explaining the answers to these five questions – answers that are very much intertwined. Before we tackle what happened after 1850, we need to understand what was known or believed in that year.

1.1 The Five Big Questions

Fourier Series

Fourier's method for expanding an arbitrary function F defined on $[-\pi, \pi]$ into a trigonometric series is to use integration to calculate coefficients:

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} F(x)\cos(kx)\,dx \quad (k \geq 0), \tag{1.1}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} F(x)\sin(kx)\,dx \quad (k \geq 1). \tag{1.2}$$

The Fourier expansion is then given by

$$F(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}\left[a_k\cos(kx) + b_k\sin(kx)\right]. \tag{1.3}$$

The heuristic argument for the validity of this procedure is that if F really can be expanded in a series of the form given in Equation (1.3), then

$$\begin{aligned}
&\int_{-\pi}^{\pi} F(x)\cos(nx)\,dx \\
&\quad = \int_{-\pi}^{\pi}\left(\frac{a_0}{2} + \sum_{k=1}^{\infty}\left[a_k\cos(kx) + b_k\sin(kx)\right]\right)\cos(nx)\,dx \\
&\quad = \int_{-\pi}^{\pi}\frac{a_0}{2}\cos(nx)\,dx + \sum_{k=1}^{\infty}\int_{-\pi}^{\pi} a_k\cos(kx)\cos(nx)\,dx \\
&\qquad + \sum_{k=1}^{\infty}\int_{-\pi}^{\pi} b_k\sin(kx)\cos(nx)\,dx.
\end{aligned} \tag{1.4}$$

Since n and k are integers, all of the integrals are zero except for the one involving a_n. These integrals are easily evaluated:

$$\int_{-\pi}^{\pi} F(x)\cos(nx)\,dx = \pi a_n. \tag{1.5}$$

Similarly,

$$\int_{-\pi}^{\pi} F(x)\sin(nx)\,dx = \pi b_n. \tag{1.6}$$

This is a convincing heuristic, but it ignores the problem of interchanging integration and summation, and it sidesteps two crucial questions:

1. Are the integrals that produce the Fourier coefficients well-defined?
2. If these integrals can be evaluated, does the resulting Fourier series actually converge to the original function?

Not all functions are integrable. In the 1820s, Dirichlet proposed the following example.

Example 1.1. The **characteristic function of the rationals** is defined as

$$f(x) = \begin{cases} 1, & x \text{ is rational,} \\ 0, & x \text{ is not rational.} \end{cases}$$

This example demonstrates how very strange functions can be if we take seriously the definition of a function as a well-defined rule that assigns a value to each number in the domain. Dirichlet's example represents an important step in the evolution of the concept of function. To the early explorers of calculus, a function was an algebraic rule such as $\sin x$ or $x^2 - 3$, an expression that could be computed to whatever accuracy one might desire.

When Augustin-Louis Cauchy showed that any piecewise continuous function is integrable, he cemented the realization that functions could also be purely geometric, representable only as curves. Even in a situation in which a function has no explicit algebraic formulation, it is possible to make sense of its integral, provided the function is continuous.

Dirichlet stretched the concept of function to that of a rule that can be individually defined for each value of the domain. Once this conception of function is accepted, the gates are opened to very strange functions. At the very least, integrability can no longer be assumed.

The next problem is to show that our trigonometric series converges. In his 1829 paper, Dirichlet accomplished this, but he needed the hypothesis that the original function F is piecewise monotonic, that is the domain can be partitioned into a finite number of subintervals so that F is either monotonically increasing or monotonically decreasing on each subinterval.

The final question is whether the function to which it converges is the function F with which we started. Under the same assumptions, Dirichlet was able to show that this is the case, provided that at any points of discontinuity of F, the value

taken by the function is the average of the limit from the left and the limit from the right.

Dirichlet's result implies that the functions one is likely to encounter in physical situations present no problems for conversion into Fourier series. Riemann recognized that it was important to be able to extend this technique to more complicated functions now arising in questions in number theory and geometry. The first step was to get a better handle on what we mean by integration.

Integration

It is ironic that integration took so long to get right because it is so much older than any other piece of calculus. Its roots lie in methods of calculating areas, volumes, and moments that were undertaken by such scientists as Archimedes (287–212 BC), Liu Hui (late third century AD), ibn al-Haytham (965–1039), and Johannes Kepler (1571–1630). The basic idea was always the same. To evaluate an area, one divided it into rectangles or triangles or other shapes of known area that together approximated the desired region. As more and smaller figures were used, the region would be matched more precisely. Some sort of limiting argument would then be invoked, some means of finding the actual area based on an analysis of the areas of the approximating regions.

Into the eighteenth century, integration was identified with the problem of "quadrature," literally the process of finding a square equal in area to a given area and thus, in practice, the problem of computing areas. In section 1 of Book I of his *Mathematical Principles of Natural Philosophy*, Newton explains how to calculate areas under curves. He gives a procedure that looks very much like the definition of the Riemann integral, and he justifies it by an argument that would be appropriate for any modern textbook.

Specifically, Newton begins by approximating the area under a decreasing curve by subdividing the domain into equal subintervals (see Figure 1.1). Above each subinterval, he constructs two rectangles: one whose height is the maximum value of the function on that interval (the circumscribed rectangle) and the other whose height is the minimum value of the function (the inscribed rectangle). The true area lies between the sum of the areas of the circumscribed rectangles and the sum of the areas of the inscribed rectangles.

The difference between these areas is the sum of the areas of the rectangles $aKbl$, $bLcm$, $cMdn$, $dDEo$. If we slide all of these rectangles to line up under $aKbl$, we see that the sum of their areas is just the change in height of the function multiplied by the length of any one subinterval. As we take narrower subintervals, the difference in the areas approaches zero. As Newton asserts: "The ultimate ratios which the inscribed figure, the circumscribed figure, and the curvilinear figure have

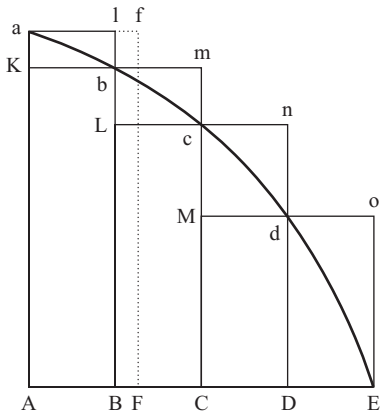

Figure 1.1. Newton's illustration from *Mathematical Principles of Natural Philosophy*. (Newton, 1999, p. 433)

to one another are ratios of equality," which is his way of saying that the ratio of any two of these areas approaches 1. Therefore, the areas are all approaching the same value as the length of the subinterval approaches 0.

In Lemma 3 of his book, Newton considers the case where the subintervals are not of equal length (using the dotted line fF in Figure 1.1 in place of lB). He observes that the sum of the differences of the areas is still less than the change in height multiplied by the length of the longest subinterval. We therefore get the same limit for the ratio so long as the length of the longest subinterval is approaching zero.

This method of finding areas is paradigmatic for an entire class of problems in which one is multiplying two quantities such as

- area = height × width,
- volume = cross-sectional area × width,
- moment = mass × distance,
- work = force × distance,
- distance = speed × time, or
- velocity = acceleration × time,

where the value of the first quantity can vary as the second quantity increases. For example, knowing that "distance = speed × time," we can find the distance traveled by a particle whose speed is a function of time, say $v(t) = 8t + 5, 0 \le t \le 4$. If we split the time into four intervals and use the velocity at the start of each interval, we get an approximation to the total distance:

$$\text{distance} \approx 5 \cdot 1 + 13 \cdot 1 + 21 \cdot 1 + 29 \cdot 1 = 68.$$

If we use eight intervals of length $1/2$ and again take the speed at the start of each interval, we get

$$\text{distance} \approx 5 \cdot \frac{1}{2} + 9 \cdot \frac{1}{2} + 13 \cdot \frac{1}{2} + \cdots + 33 \cdot \frac{1}{2} = 76.$$

If we use 1,024 intervals of length $1/256$ and take the speed at the start of each interval, we get

$$\begin{aligned}\text{distance} &\approx 5 \cdot \frac{1}{256} + \frac{161}{32} \cdot \frac{1}{256} + \frac{81}{16} \cdot \frac{1}{256} + \cdots + \frac{1{,}183}{32} \cdot \frac{1}{256} \\ &= \frac{1{,}343}{16} = 83.9375.\end{aligned}$$

As we take more intervals of shorter length, we approach the true distance, which is 84. How do you actually *get* 84? We can think of this as taking infinitely many intervals of infinitely short length.

Leibniz's notation is a brilliant encapsulation of this process:

$$\int f(x)\,dx.$$

The product is $f(x)\,dx$, the value of the first quantity times the infinitesimal increment. The elongated S, $\int$, represents the summation.

This is all precalculus. The insight at the heart of calculus is that if $f(x)$ represents the slope of the tangent to the graph of a function F at x, then this provides an easy method for computing limits of sums of products: If x ranges over the interval $[a, b]$, then the value of this integral is $F(b) - F(a)$. Thus, to find the area under the curve $v = 8t + 5$ from $t = 0$ to $t = 4$, we can observe that $f(t) = 8t + 5$ is the derivative of $F(t) = 4t^2 + 5t$. The desired area is equal to

$$(4 \cdot 4^2 + 5 \cdot 4) - (4 \cdot 0^2 + 5 \cdot 0) = 84.$$

The calculating power of calculus comes from this dual nature of the integral. It can be viewed as a limit of sums of products or as the inverse process of differentiation.

It is hard to find a precise definition of the integral from the eighteenth century. The scientists of this century understood and exploited the dual nature of the integral, but most were reluctant to define it as the sum of products of $f(x)$ times the infinitesimal dx, for that inevitably led to the problem of what exactly is meant by an "infinitesimal." It is a useful concept, but one that is hard to pin down. George Berkeley aptly described infinitesimals as "ghosts of departed quantities." He would object, "Now to conceive a quantity infinitely small, that is, infinitely less than any sensible or imaginable quantity or than any the least finite magnitude is, I confess, above my capacity."[1]

[1] George Berkeley, *The Analyst*, as quoted in Struik (1986, pp. 335, 338).

The result was that when a definition of $\int f(x)\,dx$ was needed, the integral was simply defined as the operator that returns you to the function (or, in modern use, the class of functions) whose derivative is f. One of the early calculus textbooks written for an undergraduate audience was S. F. Lacroix's *Traité élémentaire de Calcul Différentiel et de Calcul Intégral* of 1802 (*Elementary Treatise of Differential Calculus and Integral Calculus*). Translated into many languages, it would serve as the standard text of the first half of the nineteenth century. It provides no explicit definition of the integral, but does state that

> Integral calculus is the inverse of differential calculus. Its goal is to restore the functions from their differential coefficients.

After this clarification of what is meant by integration, Lacroix then proceeds to deal with the definite integral which "is found by successively calculating the value of the integral when $x = a$, then when $x = b$, and subtracting the first result from the second."

This would continue to be the standard definition of integration in calculus texts until the 1950s and 1960s. There is no loss in the power of calculus. The many textbook writers who took this approach then went on to explain how the definite integral can be used to evaluate limits of sums of products. Pedagogically, this approach has merit. It starts with the more intuitively accessible definition. Mathematically, this definition of integration is totally inadequate.

Cauchy and Riemann Integrals

Fourier and Cauchy were among the first to fully realize the inadequacy of defining integration as the inverse process of differentiation. It is too restrictive. Fourier wanted to apply his methods to arbitrary functions. Not all functions have antiderivatives that can be expressed in terms of standard functions. Fourier tried defining the definite integral of a nonnegative function as the area between the graph of the function and the x-axis, but that begs the question of what we mean by area. Cauchy embraced Leibniz's understanding as a limit of products, and he found a way to avoid infinitesimals.

To define $\int_a^b f(x)\,dx$, Cauchy worked with finite approximating sums. Given a partition of $[a, b]$: $(a = x_0 < x_1 < \cdots < x_n = b)$, we consider

$$\sum_{k=1}^{n} f(x_{k-1})(x_k - x_{k-1}).$$

If we can force all of these approximating sums to be as close to each as other as we wish simply by limiting the size of the difference between consecutive values in the partition, then these summations have a limiting value that is designated as

the value of the definite integral, and the function f is said to be integrable over $[a, b]$.

Equipped with this definition, Cauchy succeeded in proving that *any* continuous or piecewise continuous function is integrable. The class of functions to which Fourier's analysis could be applied was suddenly greatly expanded.

When Riemann turned to the study of trigonometric series, he wanted to know the limits of Cauchy's approach to integration. Was there an easy test that could be used to determine whether or not a function could be integrated? Cauchy had chosen to evaluate the function at the left-hand endpoint of the interval simply for convenience. As Riemann thought about how far this definition could be pushed, he realized that his analysis would be simpler if the definition were stated in a slightly more complicated but essentially equivalent manner. Given a partition of $[a, b]$: $(a = x_0 < x_1 < \cdots < x_n = b)$, we assign a **tag** to each interval, a number x_j^* contained in that interval, and consider all sums of the form

$$\sum_{k=1}^{n} f(x_k^*)(x_k - x_{k-1}).$$

A partition together with such a collection of tags, $x_j^* \in [x_{j-1}, x_j]$, is called a **tagged partition**. If we can force all of these approximating sums to be as close to each other as we wish simply by limiting the size of the difference between consecutive values in the partition, then these summations have a limiting value. We call this limiting value the definite integral, and the function f is said to be integrable over $[a, b]$. In the next chapter, we shall see why this seemingly more complicated definition of the integral simplifies the process of determining when a function is integrable.

Riemann succeeded in clarifying what is meant by integration. In the process, he was able to clearly identify and delimit the set of functions that are integrable and to make it possible for others to realize that this limit definition introduces serious difficulties, difficulties that eventually would lead to the rejection of Riemann's definition in favor of a radically different approach to integration proposed by Henri Lebesgue. In particular, Riemann's definition greatly complicates the relationship between integration and differentiation.

The Fundamental Theorem of Calculus

The fundamental theorem of calculus is, in essence, simply a statement of the equivalence of the two means of understanding integration, as the inverse process of differentiation and as a limit of sums of products. The precise theorems to which this designation refers today arise from the assumption that integration is

defined as a limiting process. They then clarify the precise relationship between integration and differentiation. The actual statements that we shall use are given by the following theorems.

Theorem 1.1 (FTC, evaluation). *If f is the derivative of F at every point on $[a, b]$, then under suitable hypotheses we have that*

$$\int_a^b f(t)\,dt = F(b) - F(a). \tag{1.7}$$

Theorem 1.2 (FTC, antiderivative). *If f is integrable on the interval $[a, b]$, then under suitable hypotheses we have that*

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x). \tag{1.8}$$

The first of these theorems tells us how we can use any antiderivative to obtain a simple evaluation of a definite integral. The second shows that the definite integral can be used to create an antiderivative, the definite integral of f from a to x is a function of x whose derivative is f. Both of these statements would be meaningless if we had defined the integral as the antiderivative. Their meaning and importance comes from the assumption that $\int_a^b f(t)\,dt$ is defined as a limit of summations.

In both cases, I have not specified the hypotheses under which these theorems hold. There are two reasons for this. One is that much of the interesting story that is to be told about the creation of analysis in the late nineteenth century revolves around finding necessary and sufficient conditions under which the conclusions hold. When working with Riemann's definition of the integal, the answer is complicated. The second reason is that the hypotheses that are needed depend on the way we choose to define the integral. For Lebesgue's definition, the hypotheses are quite different.

A Brief History of Theorems 1.1 and 1.2[2]

The earliest reference to Theorem 1.1 of which I am aware is Siméon Denis Poisson's 1820 *Suite du Mémoire sur les Intégrales Définies*. There he refers to it as "the fundamental proposition of the theory of definite integrals." Poisson's work is worth some digression because it illustrates the importance of how we define the definite integral and the difficulties encountered when it is defined as the difference of the values of an antiderivative at the endpoints.

[2] With thanks to Larry D'Antonio and Ivor Grattan-Guinness for uncovering many of these references.

Siméon Denis Poisson (1781–1840) studied and then taught at the École Polytechnique. He succeeded to Fourier's professorship in mathematics when Fourier departed for Grenoble to become prefect of the department of Isère. It was Poisson who wrote up the rejection of Fourier's *Theory of the Propoagation of Heat in Solid Bodies* in 1808. When, in 1815, Poisson published his own article on the flow of heat, Fourier pointed out its many flaws and the extent to which Poisson had rediscovered Fourier's own work.

Poisson, as a colleague of Cauchy at the École Polytechnique, almost certainly was aware of Cauchy's definition of the definite integral even though Cauchy had not yet published it. But the relationship between Poisson and Cauchy was far from amicable, and it would have been surprising had Poisson chosen to embrace his colleague's approach. Poisson defines the definite integral as the difference of the values of the antiderivative. It would seem there is nothing to prove. What Poisson does prove is that if F has a Taylor series expansion and $F' = f$, then

$$F(b) - F(a) = \lim_{n\to\infty} \sum_{j=1}^{n} t\ f\left(a + (j-1)t\right), \quad \text{where } t = \frac{b-a}{n}.$$

Poisson begins with the observation that for $1 \le j \le n$ and $t = (b-a)/n$, there is a $k \ge 1$ and a collection of functions R_j such that

$$F(a + jt) = F(a + (j-1)t) + tf(a + (j-1)t) + t^{1+k} R_j(t),$$

and therefore

$$\begin{aligned} F(b) - F(a) &= \sum_{j=1}^{n} \left[F(a + jt) - F(a + (j-1)t)\right] \\ &= \sum_{j=1}^{n} tf(a + (j-1)t) + t^{1+k} \sum_{j=1}^{n} R_j(t). \end{aligned}$$

Poisson now asserts that the functions $R_j(t)$ stay bounded. In fact, we know by the Lagrange remainder theorem that we can take $k = 1$ and these functions are bounded by the supremum of $|f'(x)|/2$ over all x in $[a, b]$. It follows that $t^{1+k} \sum_{j=1}^{n} R_j(t)$ approaches 0 as n approaches infinity.

The confusion over the meaning of the definite integral is revealed in Poisson's attempt to complete the proof by connecting this limit back to the definite integral. He appeals to the Leibniz conception of the integral as a sum of products:

> Using the language of infinitesimals, we shall say what we needed to show, that $F(b) - F(a)$ is the sum of the values of $f(x)\,dx$ as x increases by infinitesimal amounts from $x = a$ to $x = b$, dx being the difference between two consecutive values of this variable.[3]

[3] Poisson (1820, pp. 323–324).