CHAPTER 8

Liouville

Joseph Liouville

Generality lies at the heart of modern analysis, a trend already evident in the limit theorems of Cauchy or the integrals of Riemann. More than their predecessors, these mathematicians defined key concepts inclusively and drew conclusions valid not for one or two cases but for enormous families. It was a most significant development.

Yet the century witnessed another, seemingly opposite, phenomenon: the growing importance of the explicit example and the specific counterexample. These deserve our attention alongside the general theorems of the preceding pages. In this chapter, we examine Joseph Liouville's discovery of the first transcendental number in 1851; in the next, we consider Karl Weierstrass's astonishingly pathological function from 1872. Each of these was a major achievement of its time, and each reminds us that analysis would be incomplete without the clarification provided by individual examples.

To study transcendentals, we need some background on where the problem originated, how it was refined over the decades, and why its resolution was such a grand achievement. We start, as did calculus itself, in the seventeenth century.

The Algebraic and the Transcendental

It appears to have been Leibniz who first used the term "transcendental" in a mathematical classification scheme. Writing about his newly invented differential calculus, Leibniz noted its applicability to fractions, roots, and similar algebraic quantities, but then added, "It is clear that our method also covers transcendental curves—those that cannot be reduced by algebraic computation or have no particular degree—and thus holds in a most general way"[1]. Here Leibniz wanted to separate those entities that were algebraic, and thus reasonably straightforward, from those that were intrinsically more sophisticated.

The distinction was refined by Euler in the eighteenth century. In his *Introductio*, he listed the so-called algebraic operations as "addition, subtraction, multiplication, division, raising to a power, and extraction of roots," as well as "the solution of equations." Any other operations were transcendental, such as those involving "exponentials, logarithms, and others which integral calculus supplies in abundance" [2]. He even went so far as to mention transcendental *quantities* and gave as an example "logarithms of numbers that are not powers of the base," although he provided no airtight definition nor rigorous proof [3].

Our mathematical forebears had the right idea, even if they failed to express it precisely. To them it was evident that certain mathematical objects, be they curves, functions, or numbers, were accessible via the fundamental operations of algebra, whereas others were sufficiently complicated to transcend algebra altogether and thereby earn the name "transcendental."

After contributions from such late eighteenth century mathematicians as Legendre, an unambiguous definition appeared. A real number was said to be *algebraic* if it solved some polynomial equation with integer coefficients. That is, x_0 is an algebraic number if there exists a polynomial $P(x) = ax^n + bx^{n-1} + cx^{n-2} + \cdots + gx + h$, where $a, b, c, \ldots, g$, and h are integers and such that $P(x_0) = 0$. For instance, $\sqrt{2}$ is algebraic because it is a solution of $x^2 - 2 = 0$, a quadratic equation with integer coefficients. Less obviously, the number $\sqrt{2} + \sqrt[3]{5}$ is algebraic for it solves $x^6 - 6x^4 - 10x^3 + 12x^2 - 60x + 17 = 0$.

From a geometric perspective, an algebraic number is the x-intercept of the graph of $y = P(x)$, where P is a polynomial with integer coefficients (see figure 8.1). If we imagine graphing on the same axes all linear, all quadratic, all cubic—generally all polynomials whose coefficients are integers—then the infinite collection of their x-intercepts will be the algebraic numbers.

An obvious question arises: Is there anything else? To allow for this possibility, we say a real number is *transcendental* if it is not algebraic. Any real number must, by sheer logic, fall into one category or the other.

But are there any transcendentals? A piece of terminology, after all, does not guarantee existence. A mammalogist might just as well define a dolphin to be algebraic if it lives in water and to be transcendental if it does not. Here, the concept of a transcendental dolphin is unambiguous, but no such thing exists.

Mathematicians had to face a similar possibility. Could transcendental numbers be a well-defined figment of the imagination? Might all those (algebraic) x-intercepts cover the line completely? If not, where should one look for a number that is not the intercept of *any* polynomial equation with integer coefficients?

As a first step toward an answer, we note that a transcendental number must be irrational. For, if $x_0 = a/b$ is rational, then x_0 obviously satisfies the first-degree equation $bx - a = 0$, whose coefficients b and $-a$ are integers. Indeed, the rationals are precisely those algebraic numbers satisfying linear equations with integer coefficients.

Of course, not every algebraic number is rational, as is clear from the algebraic irrationals $\sqrt{2}$ and $\sqrt{2} + \sqrt[3]{5}$. Algebraic numbers thus represent a generalization of the rationals in that we now drop the requirement that they solve polynomials of the *first* degree (although we retain the restriction that coefficients be integers).

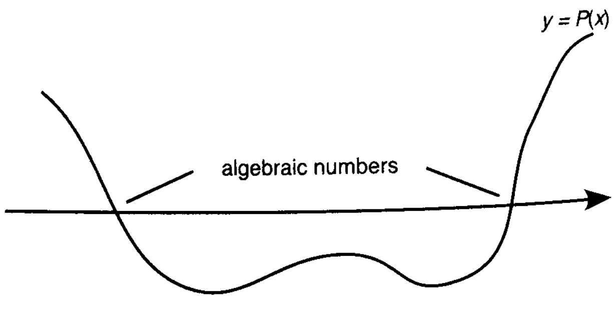

Figure 8.1

Transcendentals, if they exist, must lurk among the irrationals. From the time of the Greeks, roots like $\sqrt{2}$ were known to be irrational, and by the end of the eighteenth century, the irrationality of the constants e and π had been established, respectively, by Euler in 1737 and Johann Lambert (1728–1777) in 1768 [4]. But proving irrationality is a far easier task than proving transcendence.

As we noted, Euler conjectured that the number $\log_2 3$ is transcendental, and Legendre believed that π was as well [5]. However, beliefs of mathematicians, no matter how fervently held, prove nothing. Deep into the nineteenth century, the existence of even a single transcendental number had yet to be demonstrated. It remained possible that these might occupy the same empty niche as those transcendental dolphins.

An example was provided at long last by the French mathematician Joseph Liouville (1809–1882). Modern students may remember his name from Sturm–Liouville theory in differential equations or from Liouville's theorem ("an entire, bounded function is constant") in complex analysis. He contributed significantly to such applied areas as electricity and thermodynamics and, in an entirely different arena, was elected to the Assembly of France during the tumultuous year of 1848. On top of all of this, for thirty-nine years he edited one of the most influential journals in the history of mathematics, originally titled *Journal de mathématiques pures et appliquées* but often referred to simply as the *Journal de Liouville*. In this way, he was responsible for transmitting mathematical ideas to colleagues around Europe and the world [6].

Within real analysis, Liouville is remembered for two significant discoveries. First was his proof that certain elementary functions cannot have elementary antiderivatives. Anyone who has taken calculus will remember applying clever schemes to find indefinite integrals. Although these matters are no longer addressed with quite as much zeal as in the past, calculus courses still cover techniques like integration by parts and integration by partial fractions that allow us to compute such antiderivatives as $\int x^2 e^{-x} dx = -x^2 e^{-x} - 2xe^{-x} - 2e^{-x} + C$ or the considerably less self-evident

$$\int \sqrt{\tan x}\, dx = \frac{1}{\sqrt{8}} \ln \left| \frac{\tan x - \sqrt{2 \tan x} + 1}{\tan x + \sqrt{2 \tan x} + 1} \right| + \frac{1}{\sqrt{2}} \arctan \left(\frac{\sqrt{2 \tan x}}{1 - \tan x} \right) + C.$$

Note that both the integrands and their antiderivatives are composed of functions from the standard Eulerian repertoire: algebraic, trigonometric, logarithmic, and their inverses. These are "elementary" integrals with "elementary" antiderivatives.

Alas, even the most diligent integrator will be stymied in his or her quest for $\int \sqrt{\sin x}\, dx$ as a finite combination of simple functions. It was Liouville who proved in an 1835 paper why a closed-form answer for certain integrals is impossible. For instance, he wrote that, "One easily convinces oneself by our method that the integral $\int \frac{e^x}{x} dx$, which has greatly occupied geometers, is impossible in finite form" [7]. The hope that easy functions must have easy antiderivatives was destroyed forever.

In this chapter our object is Liouville's other famous contribution: a proof that transcendental numbers exist. His original argument came in 1844, although he refined and simplified the result in a classic 1851 paper (published in his own journal, of course) from which we take the proof that follows [8]. Before providing his example of a hitherto unseen transcendental, Liouville first had to prove an important inequality about irrational algebraic numbers and their rational neighbors.

Liouville's Inequality

As noted, a real number is algebraic if it is the solution to some polynomial equation with integer coefficients. Any number that solves one such equation, however, solves infinitely many. For instance, $\sqrt{2}$ is the solution of the quadratic equation $x^2 - 2 = 0$, as well as the cubic equation $x^3 + x^2 - 2x - 2 = (x^2 - 2)(x + 1) = 0$, the quartic equation $x^4 + 4x^3 + x^2 - 8x - 6 = (x^2 - 2)(x + 1)(x + 3) = 0$, and so on. Our first stipulation, then, is that we use a polynomial of minimal degree. So, for the algebraic number $\sqrt{2}$, we would employ the quadratic above and not its higher degree cousins.

Suppose that x_0 is an *irrational* algebraic number. Following Liouville's notation, we denote its minimal-degree polynomial by

$$P(x) = ax^n + bx^{n-1} + cx^{n-2} + \cdots + gx + h, \tag{1}$$

where $a, b, c, \ldots, g$, and h are integers and $n \geq 2$ (as noted above, if $n = 1$, the algebraic number is rational). Because $P(x_0) = 0$, the factor theorem allows us to write

$$P(x) = (x - x_0)\, Q(x), \tag{2}$$

where Q is a polynomial of degree $n - 1$. Liouville wished to establish a bound upon the size of $|Q(x)|$, at least for values of x in the vicinity of x_0. We give his proof and then follow it with a simpler alternative.

Liouville's Inequality: If x_0 is an irrational algebraic number with minimum-degree polynomial $P(x) = ax^n + bx^{n-1} + cx^{n-2} + \cdots + gx + h$ having integer coefficients and degree $n \geq 2$, then there exists a positive real number A so that, if p/q is a rational number in $[x_0 - 1, x_0 + 1]$, then

$$\left|\frac{p}{q} - x_0\right| \geq \frac{1}{Aq^n}.$$

Proof: The argument has its share of fine points, but we begin with the real polynomial Q introduced in (2). This is continuous and thus bounded on any closed, finite interval, so there exists an $A > 0$ with

$$|Q(x)| \leq A \quad \text{for all } x \text{ in } [x_0 - 1, x_0 + 1]. \tag{3}$$

Now consider any rational number p/q within one unit of x_0, where we insist that the rational be in lowest terms and that its denominator be positive (i.e., that $q \geq 1$). We see by (3) that $|Q(p/q)| \leq A$. We claim as well that $P(p/q) \neq 0$, for otherwise we could factor $P(x) = \left(x - \frac{p}{q}\right)R(x)$, and it can be shown that R will be an $(n - 1)$st-degree polynomial having integer coefficients. Then $0 = P(x_0) = \left(x_0 - \frac{p}{q}\right)R(x_0)$ and yet $\left(x_0 - \frac{p}{q}\right) \neq 0$ (because the rational p/q differs from the irrational x_0), and we would conclude that $R(x_0) = 0$. This, however, makes x_0 a root of R, a polynomial with integer coefficients having lower degree than P, in violation of the assumed minimality condition. It follows that p/q is not a root of $P(x) = 0$.

Liouville returned to the minimal-degree polynomial in (1) and defined $f(p,q) \equiv q^n P(p/q)$. Note that

$$\begin{aligned} f(p,q) &= q^n P(p/q) \\ &= q^n[a(p/q)^n + b(p/q)^{n-1} + c(p/q)^{n-2} + \cdots + g(p/q) + h] \\ &= ap^n + bp^{n-1}q + cp^{n-2}q^2 + \cdots + gpq^{n-1} + hq^n. \end{aligned} \tag{4}$$

From (4), he made a pair of simple but telling observations.

First, $f(p, q)$ is an integer, for its components $a, b, c, \ldots, g, h$, along with p and q, are all integers. Second, $f(p, q)$ cannot be zero, for, if $0 = f(p, q) = q^n P(p/q)$, then either $q = 0$ or $P(p/q) = 0$. The former is impossible because q is a denominator, and the latter is impossible by our discussion above. Thus, Liouville knew that $f(p, q)$ was a *nonzero* integer, from which he deduced that

$$|q^n P(p/q)| = |f(p, q)| \geq 1. \tag{5}$$

The rest of the proof followed quickly. From (3) and (5) and the fact that $P(x) = (x - x_0)\,Q(x)$, he concluded that

$$1 \leq |q^n P(p/q)| = q^n |p/q - x_0| |Q(p/q)| \leq q^n |p/q - x_0| A.$$

Hence $|p/q - x_0| \geq 1/Aq^n$, and the demonstration was complete. Q.E.D.

The role played by inequalities in Liouville's proof is striking. Modern analysis is sometimes called the "science of inequalities," a characterization that is appropriate here and would become ever more so as the century progressed.

We promised an alternate proof of Liouville's result. This time, our argument features Cauchy's mean value theorem in a starring role [9].

Liouville's Inequality Revisited: If x_0 is an irrational algebraic number with minimum-degree polynomial $P(x) = ax^n + bx^{n-1} + cx^{n-2} + \cdots + gx + h$ having integer coefficients and degree $n \geq 2$, then there exists an $A > 0$ such that, if p/q is a rational number in $[x_0 - 1, x_0 + 1]$, then,

$$\left|\frac{p}{q} - x_0\right| \geq \frac{1}{Aq^n}.$$

Proof: Differentiating P, we find $P'(x) = nax^{n-1} + (n-1)bx^{n-2} + (n-2)cx^{n-3} + \cdots + g$. This $(n-1)$st-degree polynomial is bounded on $[x_0 - 1, x_0 + 1]$, so there is an $A > 0$ for which $|P'(x)| \leq A$ for all $x \in [x_0 - 1, x_0 + 1]$. Letting p/q be a rational number within one unit of x_0 and applying the mean value theorem to P, we know there exists a point c between x_0 and p/q for which

$$\frac{P(p/q) - P(x_0)}{p/q - x_0} = P'(c). \tag{6}$$

Given that $P(x_0) = 0$ and c belongs to $[x_0 - 1, x_0 + 1]$, we see from (6) that

$$|P(p/q)| = |p/q - x_0| \cdot |P'(c)| \leq A|p/q - x_0|.$$

Consequently, $|q^n P(p/q)| \leq Aq^n |p/q - x_0|$. But, as noted above, $q^n P(p/q)$ is a nonzero integer, and so $1 \leq Aq^n |p/q - x_0|$. The result follows. Q.E.D.

At this point, an example might be of interest. We consider the algebraic irrational $x_0 = \sqrt{2}$. Here the minimal-degree polynomial is $P(x) = x^2 - 2$, the derivative of which is $P'(x) = 2x$. It is clear that, on the interval $[\sqrt{2} - 1, \sqrt{2} + 1]$, P' is bounded by $A = 2\sqrt{2} + 2$. Liouville's inequality shows that, if p/q is any rational in this closed interval, then $\left|\frac{p}{q} - \sqrt{2}\right| \geq \frac{1}{(2\sqrt{2} + 2)q^2}$.

The numerically inclined may wish to verify this for, say, $q = 5$. In this case, the inequality becomes $\left|\frac{p}{5} - \sqrt{2}\right| \geq \frac{1}{(50\sqrt{2} + 50)} \approx 0.00828$. We then check all the "fifths" within one unit of $\sqrt{2}$. Fortunately, there are only ten such fractions, and all abide by Liouville's inequality:

$p/5$	$\lvert p/5 - \sqrt{2}\rvert$
3/5 = 0.60	0.8142
4/5 = 0.80	0.6142
5/5 = 1.00	0.4142
6/5 = 1.20	0.2142
7/5 = 1.40	0.0142
8/5 = 1.60	0.1858
9/5 = 1.80	0.3858
10/5 = 2.00	0.5858
11/5 = 2.20	0.7858
12/5 = 2.40	0.9858

The example suggests something more: we can in general remove the restriction that p/q lies close to x_0. That is, we specify A^* to be the greater of 1 and A, where A is determined as above. If p/q is a rational within one unit of x_0, then

$$\left|\frac{p}{q} - x_0\right| \geq \frac{1}{Aq^n} \geq \frac{1}{A^* q^n} \text{ because } A^* \geq A.$$

On the other hand, if p/q is a rational *more* than one unit away from x_0, then

$$\left|\frac{p}{q} - x_0\right| \geq 1 \geq \frac{1}{A^*} \geq \frac{1}{A^* q^n}$$ because $A^* \geq 1$ and $q \geq 1$ as well.

The upshot of this last observation is that there exists an $A^* > 0$ for which $\left|\frac{p}{q} - x_0\right| \geq \frac{1}{A^* q^n}$ regardless of the proximity of p/q to x_0.

Informally, Liouville's inequality shows that rational numbers are poor approximators of irrational algebraics, for there must be a gap of at least $\frac{1}{A^* q^n}$ between x_0 and any rational p/q. It is not easy to imagine how Liouville noticed this. That he did so, and offered a clever proof, is a tribute to his mathematical ability. Yet all may have been forgotten had he not taken the next step: he used his result to find the world's first transcendental.

LIOUVILLE'S TRANSCENDENTAL NUMBER

We first offer a word about the logical strategy. Liouville sought an irrational number that was *inconsistent* with the conclusion of the inequality above. This irrational would thus violate the inequality's assumptions, which means it would not be algebraic. If Liouville could pull this off, he would have corralled a specific transcendental. Remarkably enough, he did just that [10].

Theorem: The real number $x_0 \equiv \sum_{k=1}^{\infty} \frac{1}{10^{k!}} = \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^6} + \frac{1}{10^{24}} + \frac{1}{10^{120}} + \cdots$ is transcendental.

Proof: There are three issues to address, and we treat them one at a time.

First, we claim that the series defining x_0 is convergent, and this follows easily from the comparison test. That is, $k! \geq k$ guarantees that $\frac{1}{10^{k!}} \leq \frac{1}{10^k}$, and so $\sum_{k=1}^{\infty} \frac{1}{10^{k!}}$ converges because $\sum_{k=1}^{\infty} \frac{1}{10^k} = \frac{1/10}{1 - 1/10} = \frac{1}{9}$.

In short, x_0 is a real number.

Second, we assert that x_0 is irrational. This is clear from its decimal expansion, 0.1100010000000 . . . , where nonzero entries occupy the first place, the second, the sixth, the twenty-fourth, the one-hundred twentieth, and so on, with ever-longer strings of 0s separating the

increasingly lonely 1s. Obviously no finite block of this decimal expansion repeats, so x_0 is irrational.

The final step is the hardest: to show that Liouville's number is transcendental. To do this, we assume instead that x_0 is an algebraic irrational with minimal polynomial of degree $n \geq 2$. By Liouville's inequality, there must exist an $A^* > 0$ such that, for any rational p/q, we have $\left|\frac{p}{q} - x_0\right| \geq \frac{1}{A^* q^n}$ and, as a consequence,

$$0 < \frac{1}{A^*} \leq q^n \left|\frac{p}{q} - x_0\right|. \tag{7}$$

We now choose an arbitrary whole number $m > n$ and look at the partial sum $\sum_{k=1}^{m} \frac{1}{10^{k!}} = \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^6} + \cdots + \frac{1}{10^{m!}}$. If we combine these fractions, their common denominator would be $10^{m!}$, so we could write the sum as $\sum_{k=1}^{m} \frac{1}{10^{k!}} = \frac{p_m}{10^{m!}}$, where p_m is a whole number. Thus, of course, $\frac{p_m}{10^{m!}}$ is a rational.

Comparing this to x_0, we see that

$$\left|\frac{p_m}{10^{m!}} - x_0\right| = \sum_{k=m+1}^{\infty} \frac{1}{10^{k!}} = \frac{1}{10^{(m+1)!}} + \frac{1}{10^{(m+2)!}} + \frac{1}{10^{(m+3)!}} + \cdots.$$

An induction argument establishes that $(m+r)! \geq (m+1)! + (r-1)$ for any whole number $r \geq 1$, and so $\frac{1}{10^{(m+r)!}} \leq \frac{1}{10^{(m+1)!+r-1}} = \frac{1}{10^{(m+1)!}}\left[\frac{1}{10^{r-1}}\right]$. As a consequence,

$$\begin{aligned}
\left|\frac{p_m}{10^{m!}} - x_0\right| &= \frac{1}{10^{(m+1)!}} + \frac{1}{10^{(m+2)!}} + \frac{1}{10^{(m+3)!}} + \cdots \\
&\leq \frac{1}{10^{(m+1)!}} + \frac{1}{10^{(m+1)!} \times 10} + \frac{1}{10^{(m+1)!} \times (10^2)} \\
&\quad + \frac{1}{10^{(m+1)!} \times (10^3)} + \cdots \\
&= \frac{1}{10^{(m+1)!}}\left[1 + \frac{1}{10} + \frac{1}{100} + \frac{1}{1000} + \cdots\right] \\
&= \frac{1}{10^{(m+1)!}}\left[\frac{10}{9}\right] < \frac{2}{10^{(m+1)!}}.
\end{aligned} \tag{8}$$

A contradiction is now at hand because

$$
\begin{aligned}
0 < \frac{1}{A^*} &\le (10^{m!})^n \left| \frac{p_m}{10^{m!}} - x_0 \right| && \text{by (7)} \\
&< (10^{m!})^n \cdot \frac{2}{10^{(m+1)!}} && \text{by (8)} \\
&= \frac{2}{10^{(m+1)!-n(m!)}} = \frac{2}{10^{m!(m+1-n)}} < \frac{2}{10^{m!}},
\end{aligned}
$$

where the last step follows because $m > n$ implies that $m + 1 - n > 1$.

This long string of inequalities shows that, for the value of A^* introduced above, we have $\frac{1}{A^*} < \frac{2}{10^{m!}}$ for all $m > n$, or simply that $2A^* > 10^{m!}$ for all $m > n$. Such an inequality is absurd, for $2A^*$ is a fixed number, whereas $10^{m!}$ explodes to infinity as m gets large. Liouville had (at last) reached a contradiction.

By this time, the reader may need a gentle reminder of what was contradicted. It was the assumption that the irrational x_0 is algebraic. There remains but one alternative: x_0 must be transcendental. And the existence of such a number is what Joseph Liouville had set out to prove. Q.E.D.

In his 1851 paper, Liouville observed that, although many had speculated on the existence of transcendentals, "I do not believe a proof has ever been given" to this end [11]. Now, one had.

Strangely enough, Liouville regarded this achievement as something less than a total success, for his original hope had been to show that the number e was transcendental [12]. It is one thing to *create* a number, as Liouville did, and then prove its transcendence. It is quite another to do this for a number like e that was "already there." With his typical flair, Eric Temple Bell observed that it is

> a much more difficult problem to prove that a *particular* suspect, like e or π, is or is not transcendental than it is to invent a whole infinite class of transcendentals: . . . the suspected number is entire master of the situation, and it is the mathematician in this case, not the suspect, who takes orders. [13]

We might say that Liouville demonstrated the transcendence of a number no one had previously cared about but was unable to do the same for the ubiquitous constant e, about which mathematicians cared passionately.

Still, it would be absurd to label him a failure when he found something his predecessors had been seeking in vain for a hundred years.

That original objective would soon be realized by one of his followers. In 1873, Charles Hermite (1822–1901) showed that e was indeed a transcendental number. Nine years later Ferdinand Lindemann (1852–1939) proved the same about π. As is well known, the latter established the impossibility of squaring the circle with compass and straightedge, a problem with origins in classical Greece that had gone unresolved not just for decades or centuries but for *millennia* [14]. The results of Hermite and Lindemann were impressive pieces of reasoning that built upon Liouville's pioneering research.

To this day, determining whether a given number is transcendental ranks among the most difficult challenges in mathematics. Much work has been done on this front and many important theorems have been proved, but there remain vast holes in our understanding. Among the great achievements, we should mention the 1934 proof of A. O. Gelfond (1906–1968), which demonstrated the transcendence of an entire family of numbers at once. He proved that if a is an algebraic number other than 0 or 1 and if b is an *irrational* algebraic, then a^b must be transcendental. This deep result guarantees, for instance, that $2^{\sqrt{2}}$ or $(\sqrt{2} + \sqrt[3]{5})^{\sqrt{7}}$ are transcendental. Among other candidates now known to be transcendental are e^{π}, $\ln(2)$, and $\sin(1)$.

However, as of this writing, the nature of such "simple" numbers as π^e, e^e, and π^{π} is yet to be established. Worse, although mathematicians believe in their bones that both $\pi + e$ and $\pi \times e$ are transcendental, no one has actually proved this [15]. We repeat: demonstrating transcendence is very, very hard.

Returning to the subject at hand, we see how far mathematicians had come by the mid-nineteenth century. Liouville's technical abilities in manipulating inequalities as well as his broader vision of how to attack so difficult a problem are impressive indeed. Analysis was coming of age.

Yet this proof will serve as a dramatic counterpoint to our main theorem from chapter 11. There, we shall see how Georg Cantor found a remarkable shortcut to reach Liouville's conclusion with a fraction of the work. In doing so, he changed the direction of mathematical analysis. The Liouville–Cantor interplay will serve as a powerful reminder of the continuing vitality of mathematics.

For now, Cantor must wait a bit. Our next object is the ultimate in nineteenth century rigor: the mathematics of Karl Weierstrass and the greatest analytic counterexample of all.

CHAPTER 11

Cantor

Georg Cantor

"The essence of mathematics lies in its freedom" [1]. So wrote Georg Cantor (1845–1918) in 1883. Few mathematicians so thoroughly embraced this principle and few so radically changed the nature of the subject. Joseph Dauben, in his study of Cantor's works, described him as "one of the most imaginative and controversial figures in the history of mathematics" [2]. The present chapter should demonstrate why this assessment is valid.

Cantor came from a line of musicians, and it is possible to see in him tendencies more often associated with the romantic artist than with the pragmatic technician. His research eventually carried him beyond mathematics to the borders of metaphysics and theology. He raised many an eyebrow with claims that Francis Bacon had written the Shakespearean canon and that his own theory of the infinite proved the existence of God. As an uncompromising advocate of such beliefs, Cantor had a way of alienating friend and foe alike.

Meanwhile, his life was troubled. He suffered bouts of severe depression, almost certainly a bipolar disorder whose recurrences robbed him of the "mental freshness" he so coveted [3]. Time and again Cantor was sent to what were called neuropathic hospitals to endure whatever treatment they could offer. In 1918 he died in a psychiatric institution after a life with more than its share of unhappiness.

None of this detracts from Cantor's mathematical triumph. For all of his misfortune, Georg Cantor revolutionized the subject whose freedom he so loved.

The Completeness Property

As a young man, Cantor had studied with Weierstrass at the University of Berlin. There he wrote an 1867 dissertation on number theory, a field very different from that for which he would become known. His research led him to Fourier series and eventually to the foundations of analysis.

As we have seen, developments in the nineteenth century placed calculus squarely upon the foundation of limits. It had become clear that limits, in turn, rested upon properties of the real number system, foremost among which is what we now call *completeness*. Today's students may encounter completeness in different but logically equivalent forms, such as:

C1. Any nondecreasing sequence that is bounded above converges to some real number.
C2. Any Cauchy sequence has a limit.
C3. Any nonempty set of real numbers with an upper bound has a least upper bound.

Readers in need of a quick refresher are reminded that $\{x_k\}$ is a *Cauchy sequence* if, for every $\varepsilon > 0$, there exists a whole number N such that, if m and n are whole numbers greater than or equal to N, then $|x_m - x_n| < \varepsilon$. In words, a Cauchy sequence is one whose terms get and stay close to one another. This idea put in a brief appearance in chapter 6.

Likewise, M is said to be an *upper bound* of a nonempty set A if $a \le M$ for all elements a in A, and λ is a *least* upper bound, or *supremum*, of A if (1) λ is an upper bound of A and (2) if M is any upper bound of A, then $\lambda \le M$. These concepts appear in any modern analysis text.

There is one other version of completeness, cast in terms of nested intervals, that will play an important role in the next few chapters. Again, we need a few definitions to clarify what is going on.

A closed interval $[a, b]$ is nested within $[A, B]$ if the former is a subset of the latter. This amounts to nothing more than the condition that $A \le a \le b \le B$. Suppose further that we have a sequence of closed, bounded intervals, each nested within its predecessor, as in $[a_1, b_1] \supseteq [a_2, b_2] \supseteq [a_3, b_3] \supseteq \cdots \supseteq [a_k, b_k] \supseteq \cdots$. Such a sequence is said to be *descending*. With this we can introduce another version of completeness:

C4. Any descending sequence of closed, bounded intervals has a point that belongs to each of the intervals.

It is worth recalling why the intervals in question must be both closed and bounded. The descending sequence of closed (but not bounded) intervals

$$[1, \infty) \supseteq [2, \infty) \supseteq [3, \infty) \supseteq \cdots \supseteq [k, \infty) \supseteq \cdots$$

has no point common to all of them, and the descending sequence of bounded (but not closed) intervals

$$(0, 1) \supseteq (0, 1/2) \supseteq (0, 1/3) \supseteq \cdots \supseteq (0, 1/k) \supseteq \cdots$$

likewise has an empty intersection (to use set-theoretic terminology). Although our nineteenth century predecessors often neglected such distinctions, we shall arrange for our intervals to be both closed and bounded before applying C4.

Each of these four incarnations of completeness guarantees that some real number *exists*, be it the limit to which a sequence converges, or the least upper bound that a set possesses, or a point common to each of a collection of nested intervals. As mathematicians probed the logical foundations of calculus, they realized that such existence was often sufficient for their theoretical purposes. Rather than identify a real number explicitly, it may be enough to know that a number is out there somewhere. Completeness provides that assurance.

One might ask: if the completeness property is so important, how do we prove it? The answer required mathematicians to understand the real number system itself. From the whole numbers, it is a straightforward task to define the integers (positive, negative, and zero) and from there to define the rationals. But can we create the real numbers from more elementary systems, just as the rationals were defined in terms of the integers?

Affirmative answers to this question came from Cantor and, independently, from his friend Richard Dedekind (1831–1916). Cantor's construction of the reals was based on equivalence classes of Cauchy sequences of rational numbers. Dedekind's approach employed partitions of the rationals into disjoint classes, the so-called "Dedekind cuts." A thorough discussion of these matters would carry us far afield, for constructing the real numbers from the rationals is a bit esoteric for this book and, truth be told, a bit esoteric for most analysis courses. Nonetheless, Cantor and Dedekind did it successfully and then used their ideas to prove the completeness property as a theorem in their newly created realm.

This achievement can be seen as the final step in the separation of calculus from geometry. Dedekind and Cantor had gone back to the arithmetic basics—the whole numbers—from which the reals, then the completeness property, and eventually all of analysis could be developed. Their achievement received the apt but nearly unpronounceable moniker: "the arithmetization of analysis."

THE NONDENUMERABILITY OF INTERVALS

It is not for defining the real numbers that Cantor has been chosen to headline this chapter. Rather it is for his 1874 paper, *"Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen"* (On a Property of the Totality of All Real Algebraic Numbers) [4]. This was a landmark in the history of mathematics, one that demonstrated, in Dauben's words, "[Cantor's] gift for posing incisive questions and for sometimes finding unexpected, even unorthodox answers" [5].

Oddly, the significance of the paper was obscured by its title, for the result about algebraic numbers was but a corollary, albeit a most interesting one, to the paper's truly revolutionary idea. That idea, simply stated, is that a sequence cannot exhaust an open interval of real numbers. As we shall see, Cantor's argument involved the completeness property, thus placing it properly in the domain of real analysis.

Theorem: If $\{x_k\}$ is a sequence of distinct real numbers, then any open, bounded interval (α, β) of real numbers contains a point not included among the $\{x_k\}$.

Proof: Cantor began with an interval (α, β) and considered the sequence in consecutive order: $x_1, x_2, x_3, x_4, \ldots$. If none or just one of these terms lies among the infinitude of real numbers in (α, β), then the proposition is trivially true.

Suppose, instead, that the interval contains at least two sequence points. We then identify the first two terms, by which we mean those with the two smallest subscripts, that fall within (α, β). We denote the smaller of these by A_1 and the larger by B_1. This step is illustrated in figure 11.1. Note that the initial few terms of the sequence fall outside of (α, β) but that x_4 and x_7 fall within it. By our definition, $A_1 = x_7$ (the smaller) and $B_1 = x_4$ (the greater).

We make two simple but important observations:

1. $\alpha < A_1 < B_1 < \beta$, and
2. if a sequence term x_k falls within the open interval (A_1, B_1), then $k \geq 3$.

The second of these recognizes that at least two sequence terms are used up in identifying A_1 and B_1, so any term lying strictly between A_1 and B_1 must have subscript $k = 3$ or greater. In figure 11.1, the next such candidate would be x_8.

Cantor then examined (A_1, B_1) and considered the same pair of cases: either this open interval contains none or just one of the terms of $\{x_k\}$ or it contains at least two of them. In the first case the theorem is true, for there are infinitely many other points in (A_1, B_1), and thus in (α, β), that do not belong to the sequence $\{x_k\}$. In the second case, Cantor repeated the earlier process by choosing the next two terms of the sequence, that is, those with the smallest subscripts, that fall within (A_1, B_1). He labeled the smaller of these A_2, and the larger B_2. If we look at figure 11.2 (which includes more terms of the sequence than did figure 11.1), we see that $A_2 = x_{10}$ and that $B_2 = x_{11}$.

Here again it is clear that

1. $\alpha < A_1 < A_2 < B_2 < B_1 < \beta$, and
2. if x_k falls within the open interval (A_2, B_2), then $k \geq 5$.

As before, the latter observation follows because at least four terms of the sequence $\{x_k\}$ must have been consumed in finding A_1, B_1, A_2, and B_2.

Figure 11.1

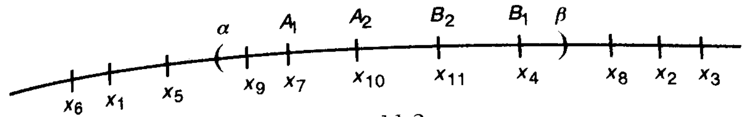

Figure 11.2

Cantor continued in this manner. If at any step there were one or fewer sequence terms remaining within the open subinterval, he could immediately find a point—indeed infinitely many of them—belonging to (α, β) but not to the sequence $\{x_k\}$. The only potential difficulty arose if the process never terminated, thereby generating a pair of infinite sequences $\{A_r\}$ and $\{B_r\}$ such that

1. $\alpha < A_1 < A_2 < A_3 < \cdots < A_r < \cdots < B_r < \cdots < B_3 < B_2 < B_1 < \beta$, and
2. if x_k falls within the open interval (A_r, B_r), then $k \geq 2r + 1$.

We then have a descending sequence of closed and bounded intervals $[A_1, B_1] \supseteq [A_2, B_2] \supseteq [A_3, B_3] \supseteq \cdots$, each nested within its predecessor. By the completeness property (C4), there is at least one point common to all of the $[A_r, B_r]$. That is, there exists a point c belonging to $[A_r, B_r]$ for *all* $r \geq 1$. To finish the proof, we need only establish that c lies in (α, β) but is not a term of the sequence $\{x_k\}$.

The first observation is immediate, for c is in $[A_1, B_1] \subset (\alpha, \beta)$ and so c indeed falls within the original open interval (α, β).

Could c appear as a term of the sequence $\{x_k\}$? If so, then $c = x_N$ for some subscript N. Because c lies in all of the closed intervals, it lies in $[A_{N+1}, B_{N+1}]$, and thus

$$A_N < A_{N+1} \leq c \leq B_{N+1} < B_N.$$

It follows that $c = x_N$ lies in the open interval (A_N, B_N), and so, according to (2) above, $N \geq 2N + 1$. This, of course, is absurd. We conclude that c can be none of the terms in the sequence $\{x_k\}$.

To summarize, Cantor had demonstrated that in (α, β) there is a point not appearing in the original sequence $\{x_k\}$. The existence of such a point was the object of the proof. Q.E.D.

Today, this theorem is usually preceded by a bit of terminology. We define a set to be *denumerable* if it can be put into a one-to-one correspondence with the set of whole numbers. Sequences are trivially denumerable, with the required correspondence appearing as the subscripts. An infinite set that cannot be put into a one-to-one correspondence with the whole numbers is said to be *nondenumerable*. We then characterize the result above as proving that any open interval of real numbers is nondenumerable.

The evolution of Cantor's thinking on this matter is interesting. Through the early 1870s, he had pondered the fundamental properties of the real numbers, trying to isolate exactly what set them apart from the rationals. Obviously, completeness was a key distinction that somehow embodied what was meant by "the continuum" of the reals.

But Cantor began to suspect there was a difference in the *abundance* of numbers in these two sets—what we now call their "cardinality"—and in November of 1873 shared with Dedekind his doubts that the whole numbers could be matched in a one-to-one fashion with the real numbers. Implicitly this meant that, although both collections were infinite, the reals were more so.

Try as he might, Cantor could not prove his hunch. He wrote Dedekind, in some frustration, "as much as I am inclined to the opinion that [the whole numbers] and [the real numbers] permit no such unique correspondence, I cannot find the reason" [6]. A month later, Cantor had a breakthrough. As a Christmas gift to Dedekind, he sent a draft of his proof and, after receiving suggestions from the latter, cleaned it up and published what we saw above. Persistence had paid off.

Readers who know Cantor's "diagonalization" proof of nondenumerability may be surprised to see that his 1874 reasoning was wholly different. The diagonal argument, which Cantor described as a "much simpler demonstration," appeared in an 1891 paper [7]. In contrast to the 1874 proof, which, as we have seen, invoked the completeness property, diagonalization was applicable to situations where completeness was irrelevant, far from the constraints of analysis proper.

Although the later argument is more familiar, the earlier one represents the historic beginning and so has been included here. We stress again that Cantor's original proof did not use terms like denumerability nor raise specific questions about infinite cardinalities. All this would come later. In 1874, he simply showed that a sequence cannot exhaust an open interval.

But why should anyone care? It was a good question, and Cantor had a spectacular answer.

The Existence of Transcendentals, Revisited

We recall that Cantor's paper was titled, "On a Property of the Totality of All Real Algebraic Numbers." To this point, algebraic numbers have yet to be mentioned, nor have we said anything about the "property" of these numbers to which the title refers. The time has come to address those omissions.

As we saw, a real number is algebraic if it is the solution to a polynomial equation with integer coefficients. There are infinitely many of these (for instance, any rational number), and it was no easy matter for Liouville to find a number that lay outside the algebraic realm.

Cantor, upon considering the matter, claimed that it was possible to list the algebraic numbers in a sequence. At first glance, this may seem preposterous. It would require him to generate a sequence with the twin properties that (1) every term was an algebraic number and (2) every algebraic number was somewhere in the sequence. A clever eye would be necessary to do this in an orderly and exhaustive fashion, but Cantor was nothing if not clever. He began by introducing a new idea.

Definition: If $P(x) = ax^n + bx^{n-1} + cx^{n-2} + \cdots + gx + h$ is an nth-degree polynomial with integer coefficients, we define its *height* by $(n-1) + |a| + |b| + |c| + \cdots + |h|$.

For instance, the height of $P(x) = 2x^3 - 4x^2 + 5$ is $(3-1) + 2 + 4 + 5 = 13$ and that of $Q(x) = x^6 - 6x^4 - 10x^3 + 12x^2 - 60x + 17$ is $(6-1) + 1 + 6 + 10 + 12 + 60 + 17 = 111$.

Clearly the height of a polynomial with integer coefficients will itself be a whole number. Further, any algebraic number has a minimal-degree polynomial whose coefficients we can assume to have no common divisor other than 1. These conventions simplify the task at hand.

Cantor in turn collected all algebraic numbers that arise from polynomials of height 1, then those that arise from polynomials of height 2, then of height 3, and so on. This was the key to arranging algebraic numbers into an infinite sequence, here denoted by $\{a_k\}$.

To see the process in action, we observe that the only polynomial with integer coefficients of height 1 is $P(x) = 1 \cdot x^1 = x$. The solution to the associated equation $P(x) = 0$ is the first algebraic number, namely $a_1 = 0$.

There are four polynomials with height 2:

$$P_1(x) = x^2,\ P_2(x) = 2x,\ P_3(x) = x + 1,\ P_4(x) = x - 1.$$

Setting the first and second equal to zero yields the solution $x = 0$, which we do not count again. Setting $P_3(x) = 0$ gives $a_2 = -1$ and $P_4(x) = 0$ gives $a_3 = 1$.

We continue. There are eleven polynomials of height 3:

$$P_1(x) = x^3,\ P_2(x) = 2x^2,\ P_3(x) = x^2 + 1,\ P_4(x) = x^2 - 1,$$
$$P_5(x) = x^2 + x,\ P_6(x) = x^2 - x,\ P_7(x) = 3x,\ P_8(x) = 2x + 1,$$
$$P_9(x) = 2x - 1,\ P_{10}(x) = x + 2,\ P_{11}(x) = x - 2.$$

Upon setting these equal to zero, we get four new algebraic numbers:

$$a_4 = -\frac{1}{2},\ a_5 = \frac{1}{2},\ a_6 = -2, \text{ and } a_7 = 2.$$

As his title indicated, Cantor was restricting his attention to *real* algebraic numbers, so $0 = P_3(x) = x^2 + 1$ added nothing to the collection.

And on we go. There are twenty-eight polynomials of height 4, and from these we harvest a dozen additional algebraic numbers, some of which are irrational. For instance, the polynomial $P(x) = x^2 + x - 1$ is of height 4 and contributes $\frac{-1+\sqrt{5}}{2}$ and $\frac{-1-\sqrt{5}}{2}$.

As the heights increase, more and more algebraic numbers appear. Conversely, any specific algebraic number must arise from *some* polynomial with integer coefficients, and this polynomial, in turn, has a height. For instance, the algebraic number $\sqrt{2} + \sqrt[3]{5}$, which we encountered in chapter 8, is a solution to the polynomial equation $x^6 - 6x^4 - 10x^3 + 12x^2 - 60x + 17 = 0$ with height 111.

A few simple observations allowed Cantor to wrap up his argument:

- For a given height, there are only finitely many polynomials with integer coefficients.
- Each such polynomial can generate only finitely many new algebraic numbers (because an *n*th-degree polynomial equation can have no more than *n* solutions).
- Hence, for each height there can be only finitely many new algebraic numbers.

This means that, upon "entering" a given height in our quest for algebraic numbers, we must emerge from that height after finitely many steps. We cannot get "stuck" in a height trying to list an infinitude of new algebraic numbers.

Consequently, the number $\sqrt{2} + \sqrt[3]{5}$ with its polynomial of height 111 has to show up somewhere in our sequence $\{a_k\}$. It will take a while, but the process must, after finitely many steps, bring us to height 111, and then, as we run through the polynomials of this height, we reach $x^6 - 6x^4 - 10x^3 + 12x^2 - 60x + 17$ after finitely many more. This will determine the position of $\sqrt{2} + \sqrt[3]{5}$ in the sequence $\{a_k\}$. The same can be said of any real algebraic number. So, the "property" of the algebraic numbers mentioned in Cantor's title is, in modern parlance, its denumerability.

Now he combined his two results: first, that a sequence cannot exhaust an interval and, second, that the algebraic numbers form a sequence. Individually, these are interesting. Together, they allowed him to conclude that the algebraic numbers cannot account for all points on an open interval. Consequently, within any (α, β), there must lie a transcendental.

Or, to put it directly, transcendental numbers *exist*.

Of course, this was what Liouville had demonstrated a few decades earlier when he showed that $\sum_{k=1}^{\infty} \frac{1}{10^{k!}} = \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^6} + \frac{1}{10^{24}} + \frac{1}{10^{120}} + \cdots$ was transcendental. To prove the existence of transcendental numbers, he went out and found one.

Cantor reached the same end by very different means. Early in his 1874 paper, he had promised "a new proof of the theorem first demonstrated by Liouville," and he certainly delivered [8]. But his argument, as we have seen, contained no example of a specific transcendental. It was strikingly nonexplicit.

To contrast the two approaches, we offer the analogy of finding a needle in a haystack. We envision Liouville, industrious to a fault, putting on his old clothes, hiking out to the field, and rooting around in the hay under a broiling sun. Hours later, drenched with perspiration, he pricks his finger on the elusive quarry; a needle! Cantor, by contrast, stays indoors using pure reason to show that the mass of the haystack exceeds the mass of the hay in it. He deduces that there must be something else, that is, a needle, to account for the excess. Unlike Liouville, he remains cool and spotless.

Some mathematicians were troubled by a nonconstructive proof that relied upon the properties of infinite sets. Compared to Liouville's lengthy argument, Cantor's seemed too easy, almost like sleight-of-hand. The young Bertrand Russell (1872–1970) may not have been alone in his initial reaction to Cantor's ideas:

> I spent the time reading Georg Cantor, and copying out the gist of him into a notebook. At that time I falsely supposed all his arguments to be fallacious, but I nevertheless went through them all in the minutest detail. This stood me in good stead when later on I discovered that all the fallacies were mine [9].

Like Russell, mathematicians came to appreciate Cantor for the innovator he was. His 1874 paper ushered in a new era for analysis, where the ideas of set theory would be employed alongside the $\varepsilon - \delta$ arguments of the Weierstrassians.

Cantor's work had consequences, many of which were truly astonishing. For instance, it is easy to show that if the algebraic numbers and the transcendental numbers are *each* denumerable, then so is their union, the set of all real numbers. Because this is not so, Cantor knew that the transcendentals form a nondenumerable set and thus far outnumber their algebraic cousins. Eric Temple Bell put it this way: "The algebraic numbers are spotted over the plane like stars against a black sky; the dense blackness is the firmament of the transcendentals" [10]. This is a delightfully unexpected realization, for the plentiful numbers seem scarce, and the scarce ones seem plentiful. In a sense, Cantor showed that the transcendentals are the hay and not the needles.

A related but more far-reaching consequence was the distinction between "small" and "large" infinite sets. Cantor proved that a denumerable set, although infinite, was *insignificantly* infinite when compared to a nondenumerable counterpart. As his ideas took hold, mathematicians came to regard denumerable sets as so much jetsam, easily expendable when addressing questions of importance.

As we shall see, dichotomies between large and small sets would arise in other analytic settings. At the turn of the nineteenth century, René Baire found a "large/small" contrast in what he called a set's "category," and Henri Lebesgue found another in what he called its "measure." Although cardinality, category, and measure are distinct concepts, each provided a means of comparing sets that would prove valuable in mathematical analysis.

Cantor addressed other questions about infinite sets. One was, "Are there nondenumerable sets having greater cardinality than intervals?" This he answered in the affirmative. Another was, "Are there infinite sets of an intermediate cardinality between a denumerable sequence and a nondenumerable interval?" This he never succeeded in resolving. With Cantor's founding vision and continuing research, set theory took on a life of its own, quite apart from the concerns of analysis proper. But it all grew out of his 1874 paper.

Unlike many revolutionaries down through history, Georg Cantor lived to see his ideas embraced by the wider community. An early enthusiast was Russell, who described Cantor as "one of the greatest intellects of the nineteenth century" [11]. This is no small praise from a mathematician, philosopher, and eventual Nobel laureate.

Another of Cantor's admirers was the Italian prodigy Vito Volterra. His work, which beautifully combined Weierstrassian analysis and Cantorian set theory, is the subject of our next chapter.