REFERENCES

1. M. Hajja, A very short and simple proof of the most elementary theorem in Euclidean geometry, *Forum Geom.* **6** (2006) 167–169.
2. M. Hajja, Stronger forms of the Steiner-Lehmus theorem, *Forum Geom.* **8** (2008) 157–161.
3. M. Hajja, On a morsel of Ross Honsberger, *Math. Gaz.* **93** (2009) 309–312.
4. Sir Thomas L. Heath, *Euclid—The Thirteen Books of the Elements, Vol. 2 (Books III–IX)*, 2nd ed., Dover Publications, New York, 1956.
5. R. Honsberger, *Mathematical Morsels*, Mathematical Association of America, Washington, DC, 1978.
6. R. Honsberger, *Mathematical Gems III*, Mathematical Association of America, Washington, DC, 1985.
7. R. Honsberger, *Episodes in Nineteenth and Twentieth Century Euclidean Geometry*, Mathematical Association of America, Washington, DC, 1995.
8. E. Trost, Problem 4964, *Amer. Math. Monthly* **68** (1961) 384; solution by R. Breusch, *Amer. Math. Monthly* **69** (1962) 672–674. doi:10.2307/2311604
9. A. A. Ungar, *A Gyrovector Space Approach to Hyperbolic Geometry*, Synthesis Lectures on Mathematics and Statistics, Morgan & Claypool, San Rafael, CA, 2009.

Summary In two of his books, Ross Honsberger presented several proofs of the fact that the point A on the circular arc $\overset{\frown}{BC}$ for which $AB + AC$ is maximum is the midpoint of the arc. In this note, we give three more proofs and examine how these proofs and those of Honsberger are related to propositions in Euclid's *Elements* and, less strongly, to other problems in geometry such as the broken chord theorem, Breusch's lemma, Urquhart's theorem, and the Steiner-Lehmus theorem.

Cantor's Other Proofs that $\mathbb{R}$ Is Uncountable

JOHN FRANKS
Northwestern University
Evanston, IL 60208-2730
j-franks@math.northwestern.edu

There are many *theorems* that are widely known among serious students of mathematics, but there are far fewer *proofs* that are part of our common culture. One of the best known proofs is Georg Cantor's diagonalization argument showing the uncountability of the real numbers $\mathbb{R}$. Few people know, however, that this elegant argument was not Cantor's first proof of this theorem, or, indeed, even his second! More than a decade and a half before the diagonalization argument appeared Cantor published a different proof of the uncountability of $\mathbb{R}$. The result was given, almost as an aside, in a paper [**1**] whose most prominent result was the countability of the algebraic numbers. Historian of mathematics Joseph Dauben has suggested that Cantor was deliberately downplaying the most important result of the paper in order to circumvent expected opposition from Leopold Kronecker, an important mathematician of the era who was an editor of the journal in which the result appeared [**4**, pp. 67–69]. A fascinating account of the conflict between Cantor and Kronecker can be found in Hal Hellman's book [**6**]. A decade later Cantor published a different proof [**2**] generalizing this result to perfect subsets of $\mathbb{R}^k$. This still preceded the famous diagonalization argument by six years.

Mathematical culture today is very different from what it was in Cantor's era. It is hard for us to understand how revolutionary his ideas were at the time. Many mathematicians of the day rejected the idea that infinite sets could have different cardinalities. Through much of Cantor's career many of his most important ideas were treated with skepticism by some of his contemporaries (see [**6**] for an interesting account of some of the disputes).

Math. Mag. **83** (2010) 283–289. doi:10.4169/002557010X521822.

As mentioned above, Cantor's first proof was in a paper [**1**] whose main result was the countability of the algebraic numbers—those real numbers which are roots of polynomials with integer coefficients. Since the real numbers are uncountable and the algebraic numbers are only countable there must be infinitely many (in fact, uncountably many) real numbers which are not algebraic. Such numbers are called *transcendental*. The fact that transcendental numbers exist had been established by Joseph Liouville, only about thirty years earlier and was itself still the subject of controversy.

Cantor's early proofs of uncountability are nearly as simple as his more famous diagonalization proof and deserve to be better known. In this expository note we present all three of these proofs and explore the relationships between them. Understanding multiple proofs of an important result almost always leads to a deeper understanding of the concepts involved.

Cantor's first proof

Recall that a set X is *countably infinite* if there is a bijection (or one-to-one correspondence) between the elements of X and the natural numbers $\mathbb{N} = \{1, 2, 3 \dots\}$. Equivalently, X is *countably infinite* if there is a sequence $\{x_k\}_{k=1}^{\infty}$ of distinct elements in which every element of X occurs precisely once. An infinite set that is not countable is called *uncountable*. So to prove that a set X is uncountable we must show that for every sequence $\{x_k\}_{k=1}^{\infty}$ of distinct elements of X there must be an element of X which is omitted by that sequence. Different sequences will omit different elements, of course, but there is no one sequence which contains every element of X.

Cantor's first proof of the uncountability of $\mathbb{R}$ was published in 1874 and is based on the fact that bounded monotonic sequences of real numbers converge.

THEOREM 1. (CANTOR [**1**]) *If $\{x_k\}_{k=1}^{\infty}$ is a sequence of distinct real numbers there is at least one $z \in \mathbb{R}$ which does not occur in this sequence.*

Proof. Let $\{x_k\}_{k=1}^{\infty} = x_1, x_2, \dots$ be a sequence of distinct real numbers. Define a sequence of closed intervals $I_n = [a_n, b_n]$ as follows. Let a_1 be the smaller of x_1 and x_2 and b_1 be the larger. Define I_1 to be $[a_1, b_1]$. We define I_n recursively. Given the non-trivial interval $I_{n-1} = [a_{n-1}, b_{n-1}]$ let y and y' be the first two elements of the sequence $\{x_k\}_{k=1}^{\infty}$ which lie in the open interval (a_{n-1}, b_{n-1}). (Clearly such y and y' must exist or there are infinitely many choices of elements of the interior of I_{n-1} which are not in the sequence $\{x_k\}_{k=1}^{\infty}$ and our proof is done.) Define a_n to be the smaller of y and y' and b_n to be the larger and let $I_n = [a_n, b_n]$.

From their construction it is clear that these closed intervals are non-trivial and nested. That is, for each index n,

$$a_{n-1} < a_n < b_n < b_{n-1},$$

and hence $I_n \subset I_{n-1}$. So the sequence $\{a_k\}_{k=1}^{\infty}$ is strictly increasing and bounded above (for example any b_n is an upper bound) and the sequence $\{b_k\}_{k=1}^{\infty}$ is strictly decreasing and bounded below.

Cantor then appealed to the fact that bounded monotonic sequences always have limits. He defined:

$$a_\infty = \lim_{n\to\infty} a_n \quad \text{and}$$
$$b_\infty = \lim_{n\to\infty} b_n$$

He observed that since $a_n < b_n$ for all n, we have $a_\infty \le b_\infty$ and the interval $[a_\infty, b_\infty]$ contains at least one point.

If $z \in [a_\infty, b_\infty]$ then

$$a_n < z < b_n \quad \text{for all } n \in \mathbb{N}, \tag{1}$$

and in particular $z \neq a_n$ and $z \neq b_n$.

We will prove by contradiction that z cannot occur in the sequence $\{x_k\}_{k=1}^{\infty}$. To do this we assume z is in the sequence and show this assumption leads to a contradiction. If z does occur in this sequence then there are only finitely many points preceding it in the sequence and hence only finitely many elements of the subsequence $\{a_n\}_{n=1}^{\infty}$ preceding it. Let a_m be the last element of the subsequence $\{a_n\}_{n=1}^{\infty}$ which precedes z in the sequence $\{x_k\}_{k=1}^{\infty}$.

We defined a_{m+1} and b_{m+1} to be the first two elements of the sequence $\{x_k\}_{k=1}^{\infty}$ which lie in the interior of I_m. Since z is in the interior of I_m, by Equation (1), and is not equal to either a_{m+1} or b_{m+1}, it must be that a_{m+1} and b_{m+1} precede z in the sequence $\{x_k\}_{k=1}^{\infty}$. This contradicts the definition of a_m as the last element of the subsequence $\{a_n\}_{n=1}^{\infty}$ preceding z in this sequence. This contradiction implies that the assumption that z is in the sequence $\{x_k\}_{k=1}^{\infty}$ is false and hence proves the result. ∎

Cantor also remarked that, in fact, the sequence $\{x_k\}_{k=1}^{\infty}$ omits at least one point in any non-empty open interval (a, b), because we could choose a_1 and b_1 to be the first two points of the sequence which lie in this interval. According to historian Joseph Dauben, this published proof benefited from some simplifications due to the German mathematician Richard Dedekind who had seen a more complicated early draft [**4**, pp. 50–52].

Indeed, the heart of this proof is the fact that bounded monotonic sequences have limits. Mathematicians in 1874 would have accepted this as a fact, but it is worth remembering that the rigorous foundations for results such as this were still being established. It was only two years earlier, in 1872, that Dedekind had published his monograph, *Stetigkeit und irrationale Zahlen,* or *Continuity and irrational numbers* [**5**]. It was in this monograph that he introduced what we now call "Dedekind cuts" as a foundation for the construction of the real numbers. This construction provided the basis for what in modern terminology is called the *completeness* of the real numbers and in particular the existence of limits for bounded monotonic sequences.

Perfect sets

In 1884 Cantor published a generalization of Theorem 1 which asserts that any perfect subset of $\mathbb{R}^k$ is uncountable. Recall that a subset X of $\mathbb{R}^k$ is said to be *perfect* if X is closed and every point x of X is a limit of a sequence of points in X which are distinct from x.

THEOREM 2. (CANTOR [**2**]) *Suppose X is a perfect subset of $\mathbb{R}^k$. Then X is uncountable.*

Proof. We will again show that if $\{x_n\}_{n=1}^{\infty}$ is a sequence in X, then there is a $z \in X$ which is not a term in this sequence.

Since X is perfect, for every x in X, a ball of any positive radius centered at x contains infinitely many points of X. From this it is easy to see that if B is a closed ball in $\mathbb{R}^k$ centered at a point of X, and y is any point of X, then there is another closed ball B' which is contained in B, is centered at a point of X, and does not contain the point y.

This property is used to construct recursively a sequence $\{z_n\}_{n=1}^{\infty}$ which has a limit z which is not an element of our original sequence. At the same time we construct a

nested sequence of closed balls $\{B_n\}_{n=1}^{\infty}$ with each B_n centered at z_n. Let B_0 be any closed ball of positive diameter D centered at a point z_0 of X. Given B_{n-1} choose a closed ball B_n such that

- The ball B_n is a subset of B_{n-1};
- The center of the ball B_n, which we denote z_n, is a point of X;
- The ball B_n does not contain the point x_n; and
- The diameter of B_n is at most half the diameter of B_{n-1}.

Notice that, for $1 \leq m \leq n$, the point x_m is not in B_n.

From the fact that each successive diameter is at most half of the previous one, it is easy to see by induction that the diameter of $B_n \leq D/2^n$. Cantor observed that the sequence $\{z_n\}_{n=1}^{\infty}$ is what we now call a Cauchy sequence. This is because if $n, m > N$, then z_n and z_m are in B_N so

$$\|z_n - z_m\| < \frac{D}{2^N}.$$

Since the sequence $\{z_n\}_{n=1}^{\infty}$ is a Cauchy sequence it has a limit in $\mathbb{R}^k$ which we will denote z. Since X is a closed set and $z_n \in X$, the limit point z is also in X.

For any $n > 0$, all of the points z_m with $m \geq n$ are in the closed ball B_n so their limit z must also be in B_n. But recall that by construction the point x_n is *not* in B_n. Hence for every n it must be that $z \neq x_n$. ∎

Using what we now know about compactness the proof above can be significantly simplified. Having constructed a nested family of balls B_n each of which contains some point of X and with $x_n \notin B_n$, we don't need to worry about centers or diameters or Cauchy sequences. Instead we let $Z_n = X \cap B_n$. Then Z_n is closed and bounded and hence compact. It is also non-empty since each B_n contains at least one point of X. And, of course, $Z_n \subset Z_{n-1}$. These properties imply that the nested intersection $\bigcap_{n=1}^{\infty} Z_n$ is non-empty. If z is a point of this intersection then for each $n \in \mathbb{N}$, $z \in B_n$ and hence $z \neq x_n$. So the point z is not in the sequence $\{x_n\}_{n=1}^{\infty}$.

Of course this line of proof was not available to Cantor. He could not have known that a nested intersection of non-empty compact sets is non-empty—indeed the concept of compactness was unknown at the time he wrote this paper. It was not until 1894 that Émile Borel proved that an open cover of a closed interval has a finite subcover. See [**7**] for a history of the concept of compactness. What we consider the standard properties of compactness were not developed until the 20th century.

Cantor published this result in §16 of [**2**]. It is interesting that it appeared a decade after his first proof (Theorem 1) and still well prior to the so-called diagonalization proof which we discuss in the next section. It certainly bears a resemblance to his first proof but, as we will see, it also strongly foreshadows the more famous diagonalization argument.

The diagonalization proof

More than a decade and a half after his first proof Cantor published the much more famous proof of the uncountability of $\mathbb{R}$ which has become associated with his name. This was the introduction of what is now called the *Cantor diagonalization argument*.

THEOREM 3. (CANTOR [**3**]) *The unit interval* $[0, 1]$ *is not countable.*

Proof. Let X denote the subset of $[0, 1]$ consisting of those numbers which have decimal representations containing only the digits 4 and 9. We choose 4 and 9 for

concreteness; other choices would work as well. We know in general that two different decimal expansions can represent the same real number. For example,

$$0.4999\cdots = 0.5000\ldots,$$

where the first decimal ends in an infinitely repeating sequence of 9's and the second in an infinitely repeating sequence of 0's. But if we allow ourselves only to use the digits 4 and 9 there is only one way to write this number.

Indeed, the representation for any number in the set X using only the digits 4 and 9 is unique. To see this suppose u and v are elements of X, so they have decimal representations using only 4 and 9; or more formally, suppose

$$u = \sum_{i=1}^{\infty} \frac{u_i}{10^i}, \quad \text{and} \quad v = \sum_{i=1}^{\infty} \frac{v_i}{10^i},$$

where each u_i and v_i is either 4 or 9. Suppose these decimal representations differ first in the nth place, so $u_i = v_i$ for $1 \le i < n$ and $u_n \ne v_n$. Let w denote the number with decimal representation equal to the decimal representation of u and v in places 1 to $n-1$ (where they agree) and with 0 in all other places so

$$w = \sum_{i=1}^{n-1} \frac{u_i}{10^i} = \sum_{i=1}^{n-1} \frac{v_i}{10^i}$$

Since u and v disagree in the nth place the larger of them has a 9 in this place and must be greater than $w + 9 \times 10^{-n}$. Similarly, the smaller of u and v has a 4 in the nth place and must be at most $w + 5 \times 10^{-n}$. Hence $|u - v| > 4 \times 10^{-n} > 0$ so $u \ne v$. This shows that two different decimal representations, which use only the digits 4 and 9, must actually represent different numbers.

Now given any sequence $\{x_k\}_{k=1}^{\infty}$ in X we define an element z by specifying its decimal expansion using a process called *diagonalization*. Specifically let

$$z = \sum_{k=1}^{\infty} \frac{z_k}{10^i},$$

where

$$z_k = \begin{cases} 4, & \text{if the } k\text{th decimal digit of } x_k \text{ is 9;} \\ 9, & \text{if the } k\text{th decimal digit of } x_k \text{ is 4.} \end{cases}$$

We conclude that z is in X, since its decimal expansion contains only the digits 4 and 9. But it is not an element of the sequence $\{x_k\}_{k=1}^{\infty}$ since z differs from x_k in the kth decimal place. It follows that it is not possible to enumerate the elements of the set X. In other words, there is no sequence $\{x_k\}_{k=1}^{\infty}$ of elements of X which contains all the elements of X. This proves X is uncountable.

There is a subtle point here. We have *not* found one z which is omitted from every sequence $\{x_k\}_{k=1}^{\infty}$. Instead we have shown that for each sequence $\{x_k\}_{k=1}^{\infty}$ there is an omitted z—different sequences will omit different elements of X.

Since $[0, 1]$ contains the uncountable set X, it must also be uncountable. (We are using the fact that a subset of a countable set is also countable.) ■

There are two parts to this proof. In the first part we showed that there is a subset X whose elements can be uniquely specified by a decimal expansion containing only the digits 4 and 9, i.e., an infinite sequence of 4's and 9's. In fact, Cantor did not include this part of the proof in his original paper. It is not difficult to show and he probably

considered it obvious. He also did not use 4 and 9 but instead used the letters m and w to represent arbitrary distinct digits. Essentially the same argument given above will show that two decimal representations of a single number must be identical if they both use only the same two digits.

The second part of the proof uses what has come to be called a *diagonalization argument* to show that the collection of all such infinite sequences is not countable. The term diagonalization is used because one way to view the construction of z given in the proof is to use the sequence $\{x_n\}$ in X to make an infinite matrix M. The first row of the matrix M consists of the decimal digits of x_1, the second row the decimal digits of x_2, and the nth row the decimal digits of x_n. So M_{ij} is the jth decimal digit of x_i. Then the element z which does not occur in the sequence is obtained from the *diagonal* of M. More precisely z_n, the nth decimal digit of z, is 4 if $M_{nn} = 9$ and 9 if $M_{nn} = 4$. Then z does not correspond to any row of the matrix M because the nth decimal digit of z is different from the diagonal entry M_{nn}. So z does not correspond to any row of the matrix M and hence z is not in the sequence $\{x_n\}$.

As mentioned above the proof for perfect subsets of $\mathbb{R}^k$ (Theorem 2 above) strongly foreshadows the diagonalization argument. To see this, let X be the subset of $[0, 1]$ consisting of those numbers with decimal representations containing only the digits 4 and 9. It is an easy exercise to show that X is perfect, though we will not need this fact. Let $X_0 = X$ and let X_n be the subset of X_{n-1} consisting of all of those points whose nth decimal digit (4 or 9) is different from the nth decimal digit of x_n. Then $\{X_n\}_{n=0}^{\infty}$ is a nested family of compact sets and $\bigcap_{n=1}^{\infty} X_n$ consists of the single point produced by the diagonalization in the proof of Theorem 3.

There is a slightly different and very clever way to make the diagonalization part of Cantor's argument. Recall that $\mathcal{P}(\mathbb{N})$, the power set of the natural numbers $\mathbb{N}$, is the set of all subsets of $\mathbb{N}$. We first observe that there is a bijection from X, the set of infinite sequences of 4's and 9's, to $\mathcal{P}(\mathbb{N})$. This bijective correspondence is given by

$$A \longleftrightarrow \{x_n\}_{n=1}^{\infty}$$

where A is a subset of $\mathbb{N}$ and $x_i = 9$ if $i \in A$ and $x_i = 4$ otherwise. Thus, it suffices to show that the set $\mathcal{P}(\mathbb{N})$ is uncountable. This can be done as a special case of a more general argument.

PROPOSITION 4. *Suppose S is a non-empty set and $f : S \to \mathcal{P}(S)$ is a function from S to its power set. Then f is not surjective.*

Proof. For each $x \in S$ either $x \in f(x)$ or $x \notin f(x)$. Let $Y = \{y \in S \mid y \notin f(y)\}$. Let x be any element of S. From the definition of Y we observe that x is in Y if and only if x is not in the set $f(x)$. Hence the sets Y and $f(x)$ can never be equal since one of them contains x and the other does not. Therefore, there is no x with $f(x) = Y$, so f is not surjective. ∎

This proposition implies that any set S has a cardinality which is less than that of its power set $\mathcal{P}(S)$ and, in particular, $\mathcal{P}(\mathbb{N})$ is uncountable. The proof of Proposition 4 is really just a disguised version of the diagonalization argument and consequently this proposition is also sometimes referred to as Cantor's diagonalization theorem.

Acknowledgment Supported in part by NSF grant DMS0901122.

REFERENCES

1. Georg Cantor, Ueber eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen, *Journal für die reine und angewandte Mathematik* (1874) 258–261. Reprinted and translated in [**8**]. doi:10.1515/crll.1874.77.258

2. Georg Cantor, Ueber unendliche lineare Punktmannichfaltigkeiten, Nr.6, *Math. Annalen* **23** (1884) 453–488. doi:10.1007/BF01446598
3. Georg Cantor, Über eine elementare Frage de Mannigfaltigketslehre, *Jahresber. Deutsch. Math.-Verein.* **1** (1890–1891) 75–78.
4. Joseph Dauben, *Georg Cantor: His Mathematics and Philosophy of the Infinite*, Princeton University Press, Princeton, NJ, 1990.
5. Richard Dedekind, *Essays on the Theory of Numbers. I: Continuity and Irrational Numbers; II: The Nature and Meaning of Numbers*, authorized translation by Wooster Woodruff Beman, Dover, New York, 1963.
6. Hal Hellman, *Great Feuds in Mathematics: Ten of the Liveliest Disputes Ever*, John Wiley, Hoboken, NJ, 2006.
7. Manya Janaky Raman, *Understanding Compactness: A Historical Perspective*, M.A. thesis, University of California Berkeley, 1997.
8. Jacqueline Stedall, *Mathematics Emerging: A Sourcebook 1540–1900,* Oxford University Press, New York, 2008.

Summary This expository note describes some of the history behind Georg Cantor's proof that the real numbers are uncountable. In fact, Cantor gave three different proofs of this important but initially controversial result. The first was published in 1874 and the famous diagonalization argument was not published until nearly two decades later. We explore the different ideas used in each of his three proofs.

Nothing Lucky about 13

B. SURY
Stat-Math Unit, Indian Statistical Institute
8th Mile Mysore Road
Bangalore 560 059 India
sury@isibang.ac.in

Recently, a high school teacher came across the following problem which he passed on to a forum for mathematics teachers:

$$\text{Evaluate } \cos\left(\frac{2\pi}{13}\right)+\cos\left(\frac{6\pi}{13}\right)+\cos\left(\frac{8\pi}{13}\right).$$

One could solve this in a number of elementary ways, and as we will show below, the value turns out to be $\frac{-1+\sqrt{13}}{4}$. The point here is to find what is special about 13 and about 2, 6, 8.

Without further ado, let us break the illusion that 13 might be particularly "lucky" to admit such a simple expression: We show a corresponding result for every prime number congruent to 1 modulo 4 and, indeed, for every prime.

Here we will explain briefly how to prove for any prime number $p \equiv 1$ modulo 4 the identity

$$\sum_{a\in Q}\cos\left(\frac{2a\pi}{p}\right)=\frac{-1+\sqrt{p}}{2}, \tag{1}$$

where the sum is over the set Q of quadratic residues mod p; that is, $a \in Q$ if $1 \le a \le p-1$ and for some integer b, $a \equiv b^2 \bmod p$. When $p \equiv 1 \bmod 4$, then -1 is a square mod p; indeed, for those who know it, we mention that Wilson's congruence $(p-1)! \equiv -1 \bmod p$ simplifies to $((\frac{p-1}{2})!)^2 \equiv -1 \bmod p$ in the case $p \equiv 1 \bmod 4$. Thus the squares mod p (as well as the nonsquares mod p) come in pairs $a, -a$ with

Math. Mag. **83** (2010) 289–293. doi:10.4169/002557010X521840.