Cantor's original proof of the uncountability of $\mathbb{R}$

Richard Chapling

v1 18 February 2019

Contrary to popular belief, Cantor's original proof that the set of real numbers is uncountable was not the diagonal argument. In this handout, we give (a modern interpretation of) Cantor's first proof,[1] then consider a way to generalise it to a wider class of objects, which we can use to prove another fact about $\mathbb{R}$ itself.

1 Nested intervals

The simplest way to present Cantor's proof is to use the following basic property of the real numbers, which is a simple consequence of Dedekind-completeness:

Theorem 1 (Nested Interval Property). *Let $(a_i)_{i=0}^{\infty}$ and $(b_i)_{i=0}^{\infty}$ satisfy the inequalities*

$$a_i \leqslant a_{i+1} \leqslant b_{i+1} \leqslant b_i$$

for every $i \in \mathbb{N}$. Then

$$\bigcap_{i=0}^{\infty} [a_i, b_i] \neq \emptyset.$$

Proof. $A \coloneqq \{a_i\}_{i=0}^{\infty}$ has b_k as an upper bound. By Dedekind-completeness, it therefore has a least upper bound, c. Since (a_i) is increasing, we must have $c \geqslant a_i$ for every i, or c will not be an upper bound. On the other hand, for each i, b_i is an upper bound for A, so since c is least, it is no bigger than b_i. So we have shown that $a_i \leqslant c \leqslant b_i$ for every i. In other words, $c \in [a_i, b_i]$ for every i, which is the same as saying that $c \in \bigcap_{i=0}^{\infty}[a_i, b_i]$, and therefore this set is nonempty, as required. □

Note that this fails for the rationals: we can choose a sequence (a_i) approaching $\sqrt{2}$ from below, and one, (b_i), approaching $\sqrt{2}$ from above, and then the intersection of the intervals $[a_i, b_i]$ contains only $\sqrt{2}$. In fact, for the real numbers, this property is completely equivalent to Dedekind-completeness.[2]

2 Cantor's first uncountability proof

Theorem 2 (Cantor, 1874[3]). *$\mathbb{R}$ is uncountable.*

Proof. Suppose $\mathbb{R}$ is countable. Then there is a surjection $\mathbb{N} \to \mathbb{R}$ given by a sequence $(\alpha_n)_{n=0}^{\infty}$. By discarding terms and relabelling, we may assume that this sequence is actually a bijection. We may also suppose, by relabelling if necessary, that $\alpha_0 < \alpha_1$.

Define $a_0 = \alpha_0$ and $b_0 = \alpha_1$. Then $a_0 < b_0$, and strictly between any two distinct reals there is a third, so let $a_1 = \alpha_{k(1)}$, where $k(1)$ is the first integer for which $a_0 < \alpha_{k(1)} < b_0$. Similarly, there is a real number strictly between a_1 and b_0, so there is an element $\alpha_{m(1)}$ of the sequence with $a_1 < \alpha_{m(1)} < b_0$ and $m(1)$ the first integer so that this holds. We continue this construction inductively: given $a_i < b_i$, we find the least integer $k(i+1)$ so that $a_i < \alpha_{k(i+1)} < b_i$ and set $a_{i+1} = \alpha_{k(i+1)}$. We then find the least integer $m(i+1)$ so that $a_{i+1} < \alpha_{m(i+1)} < b_i$ and set $b_{i+1} = \alpha_{m(i+1)}$. We hence obtain an increasing sequence $(a_i)_{i=0}^{\infty}$ and a decreasing sequence $(b_i)_{i=0}^{\infty}$ so that for each $i \geqslant 0$,

$$a_i < a_{i+1} < b_{i+1} < b_i.$$

[1] It is worth pointing out immediately that this is substantially due to Dedekind, who corresponded with Cantor, and sent him a proof which he modified and published without attribution. *Ex ungue Leonem*: it has Dedekind's clawprint all over it, using his least–upper bound property. However, since Cantor published it first, the unjust name has stuck.

[2] There is a lot to be said for actually using this as the basic completeness property, rather than least upper bounds: it fits naturally with the sort of computational mindset that is common now, and can frequently be used to give constructive proofs of results such as Rolle's theorem.

[3] Cantor, G., 'Ueber eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen.', *Journal für die reine und angewandte Mathematik (Crelle's Journal)*, Bd. 77 (1874), pp. 258–262. Available online at `http://www.digizeitschriften.de/dms/resolveppn/?PID=GDZPPN002155583`

We can now apply the Nested Intervals Property, which implies that there is a point $c \in \bigcap_i [a_i, b_i]$. In particular, we have $a_i < c < b_i$ for every i by using the above inequality.

It remains to show that c is not one of the α_i. Suppose that it is; then $c = \alpha_n$ for some n. Since there are $k(i)$ arbitrarily far into the sequence, we can find an ℓ so that $k(\ell) > n$. But by the definition of $k(\ell)$, every α_p with $p < m(\ell)$ is outside the interval $[a_\ell, b_\ell]$, contradicting that $a_\ell < \alpha_n < b_\ell$. Hence c cannot be in the sequence. Since this is true of *any* sequence, there can be no surjection $\mathbb{N} \to \mathbb{R}$, so $\mathbb{R}$ is uncountable. □

It is instructive to notice where this proof fails for $\mathbb{Q}$: namely, the least upper bound of the a_i need not be a rational number. Completeness is absolutely essential for the Nested Interval Property, and hence for this result.

3 What has Cantor actually proven?

Just as the diagonal argument may be adapted to prove a much stronger result than that $[0, 1]$ is uncountable, so this proof may be extended to a wider context than just the real numbers.

Examining the proof carefully, we see that nowhere did we need that $\mathbb{R}$ is a field. Indeed, we have used only two facts about the real numbers:

1. For any $x, y \in \mathbb{R}$ with $x < y$, there is a $z \in \mathbb{R}$ with $x < z < y$.
2. Any $\emptyset \neq A \subset \mathbb{R}$ that has an upper bound has a least upper bound.

The first property is expressed by saying that $\mathbb{R}$ is *densely ordered*. The second is familiar: it is Dedekind-completeness.

This suggests that we can apply Cantor's proof to any object that has these two properties and show it is uncountable. We shall do this in this section. We first make a small digression to pick up the definitions we need to carry out the appropriate generalisation of this idea. Afterwards, we shall consider a specific example.

Definition 3. A *total order* (or *totally ordered set*) is a set S equipped with a binary relation $<$ with the following properties:

Transitivity $\forall x, y, z \in S$, if $x < y$ and $y < z$ then $x < z$.

Trichotomy $\forall x, y \in S$, exactly one the following three statements holds:

$$x < y \qquad y < x \qquad x = y.$$

From this we define supplementary relations $>$, $\leqslant$, and $\geqslant$ exactly as one might expect:

$$x > y \iff y < x \qquad x \leqslant y \iff \neg(y < x) \qquad x \geqslant y \iff \neg(x < y).$$

(The latter two could be written in the positive as "$x < y$ or $x = y$" and "$y < x$ or $x = y$" using trichotomy.)

Next, we extend the definition of upper bound to total orders. This is done exactly as one might expect.

Definition 4. Given a total order $(S, <)$ and a set $A \subseteq S$, an element $u \in S$ is called an *upper bound for A* if for every $a \in A$, $a \leqslant u$. A *least upper bound for A* is an upper bound u so that there is no $v < u$ that is also an upper bound for A.

Both of these definitions agree with those for $\mathbb{R}$ with its usual order. Upper bounds need not exist: $\mathbb{N}$ is a total order with no upper bound, for example; likewise $\mathbb{R}$. By trichotomy, a least upper bound must be unique.

Definition 5. A totally ordered set L is called a *linear continuum* if it has the following three properties:

Multiplicity L contains more than one element.

Densely ordered $\forall x, y \in L$ with $x < y$, $\exists z \in L$ with $x < z < y$.

Dedekind-completeness Any $\emptyset \neq A \subset L$ that has an upper bound has a least upper bound.

The first property avoids the one-element set which has the other two properties trivially: think of it as similar to the axiom that $0 \neq 1$ in a field.

The following are examples of linear continua:

1. $\mathbb{R}$
2. The real intervals $[a,b]$, (a,b), $[a,b)$ and $(a,b]$, where $a,b \in \mathbb{R}$ and $a < b$.[4]
3. The long line. ($\mathbb{R}$ can be represented as countably many copies of the interval $[0,1)$ with the right end of each attached to the left end of its successor in the obvious way. The long line is the same idea, but with uncountably many copies.)
4. If L is a linear continuum and $a,b \in L$ with $a < b$, the following four sets are also linear continua:

$$(a,b)_L \coloneqq \{x \in L : a < x < b\}, \qquad [a,b)_L \coloneqq \{x \in L : a \leqslant x < b\},$$
$$(a,b]_L \coloneqq \{x \in L : a < x \leqslant b\}, \qquad [a,b]_L \coloneqq \{x \in L : a \leqslant x \leqslant b\}.$$

   These are, not surprisingly, called respectively *open*, *half-open*, and *closed intervals of L*. If it is clear that we are talking about L, we shall omit the subscripted L. (We see that these definitions agree precisely with our usual definitions if $L = \mathbb{R}$ with the usual ordering.)

The following are not examples:

1. $\mathbb{N}$ (not densely ordered)
2. $\mathbb{Q}$ (not Dedekind-complete)
3. $\mathbb{R} \setminus \{0\}$ (the set $\{-1/n\}_{n=1}^{\infty}$ does not have a least upper bound)
4. The set of all closed intervals of a linear continuum (not totally ordered).

We shall see, however, that we can make some sets of intervals into linear continua.
Again, there is a Nested Intervals result for any linear continuum:

Theorem 6 (Nested Interval Property). *Let $(a_i)_{i=0}^{\infty}$ and $(b_i)_{i=0}^{\infty}$ satisfy the inequalities*

$$a_i \leqslant a_{i+1} \leqslant b_{i+1} \leqslant b_i$$

for every $i \in \mathbb{N}$. Then

$$\bigcap_{i=0}^{\infty} [a_i, b_i] \neq \emptyset.$$

The proof is identical to the proof for $\mathbb{R}$, apart from replacing real intervals by L-intervals, and so we omit it.
We now translate Cantor's result to linear continua:

Theorem 7. *Let L be a linear continuum. Then L is uncountable.*

We will give the proof here, but do notice that it is actually almost identical to the proof for $\mathbb{R}$ given earlier.

Proof. Suppose that L is countable, then there is a sequence $(\alpha_i)_{i=0}^{\infty}$ so that $\mathbb{N} \ni i \mapsto \alpha_i \in L$ is a bijection. By swapping the two terms if necessary, we can assume that $\alpha_0 < \alpha_1$.

Define $a_0 = \alpha_0$ and $b_0 = \alpha_1$. Suppose now that we have defined $a_i, b_i \in L$ with $a_i < b_i$. Since L is densely ordered and $i \mapsto \alpha_i$ is surjective, there are terms of the sequence strictly between a_i and b_i, and in particular, there is a $k(i+1)$ so that $a_i < \alpha_{k(i+1)} < b_i$ and $k(i+1)$ is the least such integer; define $a_{i+1} = \alpha_{k(i+1)}$. Similarly, there is an $m(i+1)$ so that $a_{i+1} < \alpha_{m(i+1)} < b_i$ and $m(i+1)$ is the least such integer; define $b_i = \alpha_{m(i+1)}$. We have thus defined inductively an increasing sequence $(a_i)_{i=0}^{\infty}$ and a decreasing sequence $(b_i)_{i=0}^{\infty}$ with

$$a_i < a_{i+1} < b_{i+1} < b_i$$

for each $i \geqslant 0$.

We can now apply the Nested Intervals Property, which implies that there is a point $c \in \bigcap_i [a_i, b_i]$. In particular, we have $a_i < c < b_i$ for every i by using the above inequality.

It remains to show that c is not one of the α_i. Suppose that it is; then $c = \alpha_n$ for some n. Since there are $k(i)$ arbitrarily far into the sequence, we can find an ℓ so that $k(\ell) > n$. But by the definition of $k(\ell)$, every α_p with $p < m(\ell)$ is outside the interval $[a_\ell, b_\ell]$, contradicting that $a_\ell < \alpha_n < b_\ell$. Hence c cannot be in the sequence. Since this is true of *any* sequence, there can be no surjection $\mathbb{N} \to L$, so L is uncountable. □

[4] We can also allow $a = -\infty$ or $b = +\infty$, where $-\infty$ is an object taken to be smaller than every real number, $+\infty$ larger.

3.1 Application: disjoint interval covers

A *cover* of a set S is a set Q of subsets of S so that each element of S is an element of at least one element of Q. (A cover where the sets are disjoint is a *partition*.)

As usual, two intervals are *disjoint* if they have no points in common.

Proposition 8. *Let L be a linear continuum, and let S be a cover of L by at least two disjoint closed intervals. Then S has an ordering given by*

$$[a,b] \prec [c,d] \iff (\forall x \in [a,b], y \in [c,d])(x < y),$$

and with this ordering, S is also a linear continuum.

Before proving this, we derive a simpler characterisation for this ordering. First, note that if $[a,b] \prec [c,d]$, *a fortiori* we must have $b < c$, since b and c are elements of their respective intervals. Since then $a \leqslant b < c$, we find $a < c$. We now wish to show that the converse is also true. Suppose $a < c$. The intervals are meant to be disjoint, so

$$[a,b] \cap [c,d] = \{x \in L : a \leqslant x \leqslant b \text{ and } c \leqslant x \leqslant d\} = \{x \in L : c \leqslant x \leqslant \min\{b,d\}\}$$

must be empty, and thus we must have $\min\{b,d\} < c$. Since $c \leqslant d$, this obviously means that $b < c$. But then, if $x \in [a,b]$ and $y \in [c,d]$, we have $x \leqslant b < c \leqslant y$, so by transitivity, $x < y$ and so $[a,b] \prec [c,d]$. Therefore, on S we have

$$[a,b] \prec [c,d] \iff a < c.$$

The version on the right is *much* easier to work with.

Proof. S is specified to contain more than one element, so we need to check that $\prec$ is a total ordering and that it turns S into a linear continuum.

$\prec$ is a total ordering Let $[a_1,b_1]$ and $[a_2,b_2]$ be two intervals in the cover. Then by trichotomy of the original order relation, we have three cases:

- If $a_1 = a_2$, then by disjointness we must have $[a_1,b_1] = [a_2,b_2]$.
- If $a_1 < a_2$, then by definition $[a_1,b_1] \prec [a_2,b_2]$.
- If $a_2 < a_1$, then by definition $[a_2,b_2] \prec [a_1,b_1]$.

Hence we have trichotomy for $\prec$; the only difficult case is equality, where we had to use disjointness. It remains to show transitivity.

Suppose that $[a_1,b_1] \prec [a_2,b_2]$ and $[a_2,b_2] \prec [a_3,b_3]$. Then we have $a_1 < a_2$ and $a_2 < a_3$ by definition of $\prec$, and then since $<$ is a total order, $a_1 < a_3$, and hence $[a_1,b_1] \prec [a_3,b_3]$, which is transitivity.

S is densely ordered Given $[a_1,b_1],[a_2,b_2] \in S$ with $[a_1,b_1] \prec [a_2,b_2]$, by disjointness $b_1 < a_2$. Since L is densely ordered, there is $c \in L$ with $b_1 < c < a_2$. Then c is not in either interval, so since S is a cover, there must be an interval $[a_3,b_3] \ni c$. By disjointness, we must have $b_1 < a_3$ and $b_3 < a_2$. But also $a_i \leqslant b_i$, so by transitivity of $<$ it follows that $a_1 < a_3 < a_2$, so $[a_1,b_2] \prec [a_3,b_3] \prec [a_2,b_2]$ as required.

S is Dedekind-complete It should be reasonably obvious how most of this is going to work. Take a collection $A \coloneqq \{[a_i,b_i]\}_{i \in I} \subset S$, with an upper bound $[c,d]$, i.e. so that for each i, $[a_i,b_i] \preceq [c,d]$. Now, $\{a_i\} \subset L$ is a subset of L with c as an upper bound, so by Dedekind-completeness of L, there is $\alpha \in L$ with $a_i \leqslant \alpha \leqslant c$ for every a_i and upper bound $[c,d]$ of A.

- If $\alpha = a_j$ for some $a_j \in A$, then $[a_j,b_j]$ is a least upper bound: that there is only one choice for b_j again follows from disjointness.
- If $a_i < \alpha$ for every $a_i \in A$, since S is a cover, there must be exactly one interval $\alpha \in [\gamma,\delta] \in S$. Further, we must have $\gamma = \alpha$, for if $\gamma < \alpha$, disjointness would force that $a_i < \gamma$ for every i, so γ would then be a smaller upper bound than the least upper bound α, a contradiction. Then $[\alpha,\delta]$ is an upper bound by the definition of α and disjointness, and also by the definition of α, it must be the least such, as required. □

Corollary 9. *Let S be a cover of $\mathbb{R}$ by disjoint closed intervals. Then S is uncountable.*

This follows from the previous result and the generalisation of Cantor's theorem. The same is true of any interval of $\mathbb{R}$, provided the cover contains at least two intervals.