MATH454 LECTURE NOTES: CANTOR'S ORIGINAL 1874 ARGUMENT

KAMERYN J. WILLIAMS

To begin this course, I thought I'd do something a little different and go historical. Most, if not all of you, have already seen an a proof of Cantor's theorem that the set of real numbers is uncountable. But the argument you've (probably) seen, the diagonalization argument, is from 1891, seventeen years after Cantor first proved his theorem, in what is the first set theory paper. I want to give you Cantor's original argument, in the context of giving a new proof of Liouville's theorem on the existence of transcendental numbers. I will follow Cantor's 1874 paper "On a property of the collection of all real algebraic numbers", though I will use modern notation.

1. On a property of the set of real algebraic numbers

Definition 1. A real number x is said to be *algebraic* if there is a natural number n and integers $a_0, \ldots, a_n$ so that

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

In other words, an algebraic real is one which is a root of a polynomial with integer coefficients.

A real number is *transcendental* if it is not algebraic.

Observe that if x is rational then x is algebraic: If $x = p/q$, where p and q are integers, then $qx - p = 0$. As a consequence, the algebraic numbers are *dense in the reals*: if $a < b$ are real numbers then there is an algebraic number between a and b. (This is simply because the rationals are dense in the reals.) Indeed, it immediately follows that there are infinitely many algebraic numbers between a and b, because we can keep splitting the interval into smaller and smaller pieces and keep finding new algebraic numbers. Also observe that there are irrational algebraic numbers, such as $\sqrt{2}$.

A natural question to ask is whether every real number is algebraic. The answer turns out to be no.

Theorem 2 (Liouville 1844)**.** *Let $a < b$ be real numbers. Then there are infinitely many transcendental numbers in the interval (a, b).*

Liouville's construction—which we will not see—is 'artificial'; the transcendental numbers he constructs hadn't been considered before his proof. Later, well-known numbers like e and π were shown to be transcendental. Showing a given number is transcendental is in general really, really, hard. For example, it is still an open question whether $\pi + e$ is transcendental.

Let's now move to seeing Cantor's proof of Liouville's theorem. His proof goes through two steps, which I will label as lemmata.

Lemma 3 (Cantor)**.** *There is a sequence $x_0, x_1, \ldots, x_n, \ldots$, for $n \in \mathbb{N}$, of all algebraic reals.*

Lemma 4 (Cantor)**.** *Let $a < b$ be real numbers. Then there is no sequence of all real numbers in the interval (a, b).*

Date: August 15, 2019.

Before we prove these lemmata, let's see how they imply Liouville's theorem.

Proof of Liouville's theorem from the lemmata. Let $a < b$ be real numbers and let $x_0, x_1, \ldots, x_n, \ldots$ be the sequence of all algebraic reals. From this sequence we can get a sequence of all algebraic reals in (a, b): Simply let y_0 be the first x_n which is in (a, b), then y_1 be the next, and so on. This process never stops because there are infinitely many algebraic reals in the interval. By the second lemma, this sequence cannot make up the entire interval (a, b). So there must be a real in the interval not in the sequence; that is, there must be a transcendental real in the interval. □

Now let's move on to proving the lemmata.

Proof of Lemma 3. Consider a polynomial

$$a_n x^n + a_{n-1} x^{n+1} + \cdots + a_1 x + a_0 = 0$$

with integer coefficients. As an *ad hoc* definition, say that the *height* of the polynomial is the sum of the absolute values of its coefficients plus its degree. That is, the height is

$$n + |a_n| + |a_{n-1}| + \cdots + |a_1| + |a_0| \,.$$

The key claim is that given a height H, there are only finitely many polynomials of height H. This is because the degree n must be less than H, and given a degree n and height H the absolute values of the coefficients must add to $H - n$, of which there are only finitely many ways to do so. Because a polynomial of degree n has at most n roots, this means that given a height H there are finitely many reals which are roots of a polynomial of height H.

Given H, let $A(H)$ be the set of all algebraic numbers which are roots of polynomials of height exactly H. We have just seen that $A(H)$ is always finite. We now define a sequence of all algebraic numbers as follows: start by listing all elements of $A(2)$, in ascending order. (Observe that $A(0)$ and $A(1)$ are both empty, so there is nothing from them to list.) This gives the first $|A(2)| = 1$ members of the sequence. Next list all elements of $A(3)$, again in ascending order. This gives the next $|A(3)|$ members of the sequence. Continue onward in this fashion. Before stage n, we have only listed $|A(1)| + \cdots + |A(n-1)|$ many numbers, i.e. finitely many. So what we get really is a sequence, with each member appearing at some natural number index.

It is clear from the construction that every number in the sequence is algebraic. On the other hand, the sequence contains all algebraic numbers, since each algebraic number is the root of a polynomial of some height H. □

The sequence we constructed will have lots of repetitions. Each algebraic number will appear infinitely often! If this bothers you, then when constructing the sequence you can only add a new element if it hasn't yet appeared. This gives a sequence of all algebraic reals, one without any repetitions.

Proof of Lemma 4. Suppose we have a nondegenerate interval (a, b) and a sequence

$$x_0, \ldots, x_1, \ldots, x_n, \ldots.$$

We want to find a number in the interval (a, b) which does not appear in the sequence. To that end, set $a_0 = a$ and $b_0 = b$. Given $a_i < b_i$, we want to define a_{i+1} and b_{i+1} so that $a_i < a_{i+1} < b_{i+1} + b_i$. Namely, we do this by letting a_{i+1} be the first number on the sequence which is in (a_i, b_i) and letting b_{i+1} be the first number on the sequence which is in (a_{i+1}, b_i).

Now, it may be that this process terminates at some finite stage i. (For example, if $a = 0$, $b = 1$, and each x_n is greater than 1, then it terminates at stage 0.) But if this happens, then we have that at most one element of the sequence is in the interval (a_i, b_i). At this point, the conclusion of the lemma is clear.

So suppose we are not in this boring case, and that this process goes on infinitely. That is, we have constructed a descending sequence (a_0, b_0), (a_1, b_1), $\ldots$, (a_i, b_i), $\ldots$, for $n \in \mathbb{N}$, of intervals which get smaller and smaller in length. Consider the sequences $a_0 < a_1 < \cdots a_i < \cdots$ and $b_0 > b_1 > \cdots > b_i > \cdots$. Let A be the supremum (= least upper bound) of the set $\{a_i : i \in \mathbb{N}\}$, which exists by the completeness property of $\mathbb{R}$. And let B be the infimum (= greatest lower bound) of the set $\{b_i : i \in \mathbb{N}\}$, again existing by the completeness property of $\mathbb{R}$.

I claim that no number in the interval $[A, B]$ appears in the sequence $x_0, x_1, \ldots, x_n, \ldots$. To see why, let's look a little more closely at our construction of the sequence of intervals (a_i, b_i). Say that x_n is *rejected at stage* i if x_n is not in the interval (a_i, b_i). The idea is, not being in the interval (a_i, b_i) rules out being in the interval $[A, B]$, since it is a strictly smaller interval. And note that, for the same reason, if x_n is rejected at stage i then x_n is rejected at every further stage. So let's see that each x_n is rejected at some stage. There are two possibilities. The first is that x_n is a_i or b_i for some i. In this case, x_n won't be in (a_{i+1}, b_{i+1}), so x_n is rejected at stage $i+1$. The second case is that this doesn't happen. Note that we must have eventually looked at x_n during stage i some large enough i, since each stage consumes two numbers from the sequence. But by the definition of the intervals (a_i, b_i) this means that when we looked at x_n during stage i to try to decide whether it was within (a_i, b_i), we had that it was out of bounds. So then x_n was rejected at stage i.

Since each x_n fails to be in (a_i, b_i) for some large enough i, we get that each x_n fails to be in $[A, B]$, as desired. □

(Kameryn J. Williams) University of Hawai'i at Mānoa, Department of Mathematics, 2565 McCarthy Mall, Keller 401A, Honolulu, HI 96822, USA

E-mail address: kamerynw@hawaii.edu

URL: http://kamerynjw.net