the Size of a Set

Eric C.R. Hehner
University of Toronto

Abstract. It is popularly believed that Cantor's diagonal argument proves that there are more reals than integers. In fact, it proves only that there is no onto function from the integers to the reals; by itself it says nothing about the sizes of sets. Set size measurement and comparison, like all mathematics, should be chosen to fit the needs of an application domain. Cantor's countability relation is not a useful way to compare set sizes.

Keywords. Cantor, diagonal argument, infinity, set size

Introduction

We can count the elements of a finite set, and the natural number we get is called its size. Comparing the sizes of two finite sets just means comparing two natural numbers. For infinite sets, sizes are not so easily found and compared. The foundational work on measuring and comparing the sizes of infinite sets is by Cantor [0], and that is where we begin.

Cantor and the Uncountable

Cantor's "diagonal argument" is supposed to show that there are more real numbers than integer numbers. We will be content with reals r in the range $0 \le r \le 1$, which we will call the "small" real numbers, and integers i in the range $0 \le i$, commonly called the "natural" numbers. The argument is presented with the aid of a picture, which motivates the name "diagonal".

```
0   →   . 1 4 1 5 9 2 6 5 3 5 8 ···
1   →   . 1 4 2 8 5 7 1 4 2 8 5 ···
2   →   . 1 0 1 0 0 1 0 0 0 1 0 ···
3   →   . 6 5 3 4 9 7 6 4 8 4 2 ···
4   →   . 7 6 4 3 9 6 0 6 3 4 5 ···
    ⋮
```

On row 0, we have a decimal point followed by an infinite sequence of decimal digits, representing a small real number. On row 1, we have a decimal point followed by another infinite sequence of decimal digits, representing another small real number. And so on: for each natural we have a representation of a small real. We now form an infinite sequence of digits that is not on any row, as follows. First, form the diagonal sequence by taking the first digit from the first row, the second digit from the second row, and so on. In our example, that's .14149... (shown in bold). Now form a sequence of digits that differs in every position from the diagonal sequence. We could add 1 modulo

10 to each digit. In our example, that's .25250... . This new sequence differs from the sequence on every row in at least one position. We are supposed to see from this that any attempt to make a list of all small reals (or mapping from naturals onto small reals) fails because, from any list of small reals, we can construct a small real that differs from every small real on the list. We are now supposed to conclude that there are more small real numbers than natural numbers.

Like most mathematical proofs, the proof just presented is informal. In the picture we see three dots in several places, which means "guess what goes here". And the proof is written mostly in a natural language (English), rather than in mathematics. Informal proofs are error-prone; the errors are hidden in the imprecision and ambiguities of the words. I'll begin with a minor flaw, and then move on to a more major concern.

Some small real numbers have two representations. For example, .1234000... = .1233999... , where the left side of the equation ends in an infinite string of 0s and the right side ends in an infinite string of 9s. So it is possible that the sequence of digits we create, which differs from all sequences in the list, nonetheless represents a number that is already represented in the list. The problem can be repaired by banishing sequences ending in an infinite string of 9s from the list. To prove that this repairs the problem requires proving that there are no other ways for a small real to have two representations, and that if a sequence ending in an infinite string of 0s is on the list, then the altered diagonal cannot be its equal ending in an infinite string of 9s. Banishing sequences ending in an infinite string of 0s from the list does not work; it is indeed possible that a sequence ending in an infinite string of 9s is on the list, and the diagonal is its equal ending in an infinite string of 0s. (If we make Cantor's diagonal argument using binary digits, the problem cannot be repaired by banishing sequences.) Another way to repair the problem is to add 2 modulo 10 to each digit of the diagonal sequence. Now the constructed sequence differs from the sequence on every row in at least one position by 2 modulo 10 , so it cannot represent the same number as the row represents. (If we make Cantor's diagonal argument using binary digits, we have to form a diagonal from pairs of bits on each row.)

Program Analogy

Here is an argument that is closely analogous to Cantor's. First, choose a programming language; all programs in this argument will be in this programming language. As before, we have rows labeled with natural numbers. On each row is a program to print an infinite sequence of digits. We can write each program completely without the need for three horizontal dots, but we cannot write infinitely many programs. To get rid of the vertical dots, we write a program to print an infinite sequence of programs. We can certainly write a program to print out all the programs in our language; the programs of a language can be generated from a grammar for the language. But that's not what we want; we want a list of all and only those programs that print an infinite sequence of digits; we want those digit-printing programs whose execution neither halts nor goes into a non-printing infinite loop. Now comes the diagonal construction. If we have a program P to print all and only the infinite-sequence programs S_i , then we can write a new program D to simulate P , producing the S_i

(without printing them), and in turn simulate each S_i up to i digits (without printing them), and then print a different digit. We thus create an infinite-sequence program that differs from all the S_i. Now what should we conclude?

To argue as Cantor did, we would have to conclude that there are more infinite-sequence programs than natural numbers. If the inability to list all the small reals is cause for concluding that there are more small reals than natural numbers, then the inability to list all the infinite-sequence programs should equally be cause for concluding that there are more infinite-sequence programs than natural numbers. Now we have the following uncomfortable situation:

The infinite-sequence programs are a subset of all programs, so
- there are not more infinite-sequence programs than programs.

We can list all programs, so
- there are not more programs than natural numbers.

We cannot list the infinite-sequence programs, so
- there are more infinite-sequence programs than natural numbers.

The first two bullet-points together contradict the last bullet-point; a Cantor-type conclusion concerning the sizes of sets of programs is absurd. The proper conclusion is simply that there is no program (in some programming language) to generate all and only the infinite-sequence programs (in that same programming language). It is not a conclusion about the sizes of sets. Likewise in the original Cantor argument, the proper conclusion is not about the sizes of sets; it is simply that there is no sequence of all infinite sequences.

Like the original Cantor argument, this program-analogy version is informal, and the informality may hide serious errors. Two paragraphs ago, "the inability to list all the small reals" was compared to "the inability to list all the infinite-sequence programs". Cantor was not concerned with how the list of small reals might be generated; he just supposed that we have such a list. In the program analogy, we suppose that the list of infinite-sequence programs is generated by a program, rather than just supposing we have such a list, no matter how it might be generated. This sparks a debate: is there any other way to generate an infinite list? what does it mean to have an infinite list, with no way to generate it? I prefer not to engage in that debate. I prefer to consider informal proofs to be rough sketches, and to insist that proofs be finished by formalizing them.

Formal Proof

A mathematical formalism consists of rules for writing formulas (expressions), and rules for saying which formulas are theorems. These rules must be so precise and unambiguous that they are machine-checkable. That is what makes proofs objective, not a matter of opinion. Here is a formal proof using the formalism of [1].

Define Cantor's relation $\lesssim$ between sets A and B as

$$A \lesssim B \;=\; \exists f\colon A \to B \cdot\; \forall b\colon B \cdot\; \exists a\colon A \cdot\; f\,a = b$$

The relation $\lesssim$ is a pre-order (reflexive and transitive). How we pronounce a formula is not part of the formalism, but it may be helpful to make a suggestion. We can pronounce $A \lesssim B$ as "A is less countable than or equally as countable as B". The part to the right of the defining equals can be read "there is a function f from A to B such that for any element b of B there is an element a of A such that f applied to a equals b". More briefly, it can be read "there is an onto function from A to B". Next, let nat be the natural numbers. Then $nat \to nat$ is the set of functions from nat to nat, that is, the set of infinite sequences of naturals. We prove

$$\neg(nat \lesssim nat \to nat)$$

which says that there is no onto function from the naturals to the infinite sequences of naturals. Here is the proof. (The notation $\langle m\colon nat \to f\ m\ m + 1\rangle$ expresses a function that maps natural m to $f\ m\ m + 1$. $\bot$ is false and $\top$ is true.)

	$\neg(nat \lesssim nat \to nat)$	definition of $\lesssim$
$=$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \forall g\colon nat \to nat \cdot\; \exists n\colon nat \cdot\; f\,n = g$	specialize g
$\Leftarrow$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \exists n\colon nat \cdot\; f\,n = \langle m\colon nat \to f\ m\ m + 1\rangle$	function equality
$=$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \exists n\colon nat \cdot\; \forall p\colon nat \cdot\; f\,n\,p = \langle m\colon nat \to f\ m\ m + 1\rangle p$	apply function
$=$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \exists n\colon nat \cdot\; \forall p\colon nat \cdot\; f\,n\,p = f\,p\,p + 1$	specialize p to n
$\Leftarrow$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \exists n\colon nat \cdot\; f\,n\,n = f\,n\,n + 1$	cancellation
$=$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \exists n\colon nat \cdot\; 0 = 1$	basic arithmetic
$=$	$\neg\exists f\colon nat \to nat \to nat \cdot\; \exists n\colon nat \cdot\; \bot$	idempotence (unused quantifiers)
$=$	$\neg\bot$	boolean law
$=$	$\top$	

The proof is formal, not a wordy argument. It can be submitted to a proof checker, which complains that the step "specialize g" requires the function $\langle m\colon nat \to f\ m\ m + 1\rangle$ to be of type $nat \to nat$. To satisfy this requirement, we prove

	$\top$	context
$=$	$f\colon nat \to nat \to nat$	function inclusion
$=$	$\forall m\colon nat \cdot\; f\,m\colon nat \to nat$	function inclusion
$=$	$\forall m, p\colon nat \cdot\; f\,m\,p\colon nat$	specialize p to m
$\Rightarrow$	$\forall m\colon nat \cdot\; f\,m\,m\colon nat$	nat construction (Peano)
$\Rightarrow$	$\forall m\colon nat \cdot\; f\,m\,m + 1\colon nat$	function inclusion
$=$	$\langle m\colon nat \to f\ m\ m + 1\rangle : nat \to nat$	

The proof checker needed to be guided to a Peano axiom, instantiated to say that if $f\,m\,m$ is in nat then $f\,m\,m + 1$ is in nat. The proof is now complete and correct, with no hidden assumptions or errors. It says there is no onto function from the naturals to the infinite sequences of naturals. In other

words, the naturals are more countable than the infinite sequences of naturals. But it says nothing about the sizes of sets. For that, we need one more ingredient.

Set Size Comparison

Here are some ways to compare the sizes of finite sets.

0 If A is a subset of B, then A is not larger than B.
1 If A is a proper subset of B, then A is smaller than B.
2 If there is an onto function from A to B, then A is not smaller than B.
3 If there is no onto function from A to B, then A is smaller than B.
4 If there is a one-to-one correspondence between A and B, then they have the same size.
5 If there is a two-to-one correspondence from A to B, then there are twice as many elements of A as B.
6 If there is an ordering on the elements of A and B such that between every pair of elements of A there is an element of B, then B is at least as large as A (within one element).

And there may be other ways. We may use these ways of comparing the sizes of finite sets to suggest ways of comparing the sizes of infinite sets. Cantor chose 2, 3, and 4 for comparing infinite sets. Formally, he adopted the axioms

2 $A \precsim B \implies \neg(|A| < |B|)$
3 $\neg(A \precsim B) \implies |A| < |B|$
4 $(A \precsim B) \wedge (B \precsim A) \implies |A| = |B|$

Altogether, they can be written as

$(A \precsim B) = (|A| \geq |B|)$

Cantor chose the countability ordering $\precsim$ to be the size ordering: less countable is larger. Cantor rejects 1; he says there are equally many even naturals as naturals, even though the even naturals are a proper subset of the naturals. Cantor rejects 5; he says there are equally many naturals as even naturals, not twice as many, even though there is a two-to-one correspondence from the naturals to the even naturals. (There is a two-to-one correspondence from the naturals to the naturals, and Cantor did not want to say there are twice as many naturals as naturals!) Cantor rejects 6; he says there are more reals than rationals, even though there is a rational between every two reals. Depending on the formalization of "having an infinite list", Cantor may have to reject 0; by the transitivity of ordering, he may have to say there are more infinite-sequence programs than programs, even though the infinite-sequence programs are a subset of all programs. Cantor chose to use 2, 3, and 4, rejecting the other properties, for comparing infinite set sizes. But he could have made a different choice.

Mathematical Design

A Platonist mathematician believes that mathematical objects, like sets and numbers, exist in some sense, and they discover facts, or truths, about them. They believe that Cantor discovered the fact that set sizes are compared by the countability pre-order. Thanks to the Platonist legacy, it is now standard to say that the following three statements are facts, or that they are true:

- There is the same number of even naturals as naturals.
- There are more reals than rationals.
- There are infinitely many infinities.

In contrast to a Platonist, a Formalist mathematician believes that mathematics is a language, or notation, designed by people to describe the world in a quantitative and calculational way. Numbers and sets and so on are the words or expressions of the language. They are not existing abstract objects, but they can be used to express or describe existing real objects in the domains of application of mathematics. The application area motivates and justifies the design of a mathematical formalism. We design it so that its theorems represent the truths in the application area.

The formalism in [1] was designed for reasoning about computer programs. An infinite number ∞ is needed to account for program executions that take infinite time (nonterminating executions). The infinite sets of naturals, integers, and rationals are useful because their theories are much simpler and easier to reason with than the set of integers modulo 2^{32} and the set of 64-bit floating-point numbers; these infinite sets leave out the complications of finite boundaries that are irrelevant for many computations (even the IEEE floating-point standard includes ∞). Sometimes, as in the diagonal argument about programs that we saw earlier, we want to reason about programs that take infinite time and use infinitely many integers. In the theory in [1], there's just one infinite number (and its negation); the application has no need for any other infinities. This infinity absorbs finite additions ($\infty+1=\infty$) and subtractions ($\infty-1=\infty$); it also absorbs positive finite multiplications ($\infty\times 2=\infty$) and divisions ($\infty/2=\infty$). As a result, there are the same number of even naturals as naturals, AND there are half as many even naturals as naturals.

For some applications, Hölder's process gives a useful comparison of the sizes of sets of naturals. Given a set A of naturals, define

$S\ 0\ n\ =$ **if** $n\in A$ **then** 1 **else** 0

$S\ (m+1)\ n\ =\ (\Sigma i{:}\ 0\le i\le n{\cdot}\ S\ m\ i)\ /\ (n+1)$

Then the relative size of set A is

limit $(m\to\infty)$ *limit* $(n\to\infty)\ S\ m\ n$

The relative size of the set of naturals is 1 . The relative size of the set of even naturals is 1/2 . The relative size of the set of positive naturals is 1 . The relative size of the set of primes is 0 .

If there were an application area (outside mathematics!) that requires an infinite hierarchy of infinities, then for that application, the countability relation might be just right for comparing the sizes of sets. So

far, we have no such application. Without an application that requires higher cardinalities, we have no motivation for inventing higher cardinalities.

Perhaps higher cardinalities are just waiting for an application. Pure mathematics is sometimes praised as being ready to apply when the right application comes along. But I am skeptical of this justification. If you have a hammer and you're looking for an application, each potential application will look to you like a nail. An example is "Applications of Group Theory to Quantum Mechanics and Particle Symmetries" (I wrote a thesis with that title); you can apply group theory, but it would be better to design some mathematics to fit the application, rather than to bend the application to some existing mathematics.

Another example is the semantics of loops in programming. One can form a sequence of approximations to the semantics by unrolling a loop more and more, and then take the limit of the sequence of approximations. But this limit may not satisfy the loop's recurrence relation (may not be a fixed-point). So the sequence is restarted from a higher cardinality, as though the loop could be iterated more than an infinite number of times. But there are other, much simpler, methods for finding loop semantics that do not involve higher cardinalities [1]. So there have been applications (outside mathematics) that have used higher cardinalities, but as far as I know, there have been no applications that required higher cardinalities, and none that were best served by higher cardinalities.

Conclusion

It is popularly believed that Cantor's diagonal argument proves that there are more reals than integers. In fact, it proves only that there is no onto function from the integers to the reals; by itself it says nothing about the sizes of sets. Set size measurement and comparison, like all mathematics, should be chosen to fit the needs of an application domain. For all application domains that I know of, Cantor's countability relation is not the most useful way to compare set sizes.

References

[0] G.Cantor: über ein Elementare Frage der Mannigfaltigkeitslehre, *Deutsche Mathematiker-Vereinigung* v.1 p.75-78, 1890

[1] E.C.R.Hehner: *a Practical Theory of Programming*, first edition Springer 1993, current edition www.cs.utoronto.ca/~hehner/aPToP