# What is so wrong with thinking of real numbers as infinite decimals?

One of the early objectives of almost any university mathematics course is to teach people to stop thinking of the real numbers as infinite decimals and to regard them instead as elements of the unique complete ordered field, which can be shown to exist by means of Dedekind cuts, or Cauchy sequences of rationals. I would like to argue here that there is *nothing* wrong with thinking of them as infinite decimals: indeed, many of the traditional arguments of analysis become more intuitive when one does, even if they are less neat. Neatness is of course a great advantage, and I do not wish to suggest that universities should change the way they teach the real numbers. However, it is good to see how the conventional treatment is connected to, and grows out of, more `naive' ideas. I shall illustrate this later with a discussion of the square root of two and the intermediate value theorem.

---

## Constructing the real numbers as infinite decimals.

Recall that to `construct' the real numbers means to give an example of a complete ordered field. For this purpose one is allowed to assume the existence of the rational numbers (themselves constructed from the positive integers, which can, if you want, be constructed from sets) and certain notions, such as that of an infinite sequence, which are not entirely unproblematic (for reasons discussed here ) but which are in a sense more elementary.

Here is an indication of how infinite decimals give a perfectly good, and very easily understood, construction of the real numbers. Because of irritating difficulties such as the need to carry digits and to identify 0.999999.... with 1, and because there is something unnatural about the number 10 (or any other number one might choose, though 2 might be an improvement) this construction is less aesthetically pleasing than some. However, it has the advantage, for beginners, of being very close to the picture of real numbers they already have.

To begin with, one defines an infinite decimal in the obvious way, as a finite sequence of elements of the set {0,1,2,3,4,5,6,7,8,9} followed by a decimal point followed by an infinite sequence of elements of the set {0,1,2,3,4,5,6,7,8,9}. This isn't quite the whole definition since one must point out that some of these objects are equal: for example, 0124.383478... is the same number as 124.383478... (assuming of course that the sequences continue in the same way) and 1.9999999... is the same number as 2. (About this last example, by the way, there can be no argument, since I am giving a *definition.* I can do this in whatever way I please, and it pleases me to stipulate that 1.999999... =2 and to make similar stipulations whenever I have an infinite string of nines.)

That defines the set I am constructing. To make it a complete ordered field, I must now specify the ordering, explain how to add and multiply infinite decimals, and prove that the field axioms, order axioms and completeness axiom are all satisfied. First, then, how does one add two infinite decimals? The answer is simple, but to explain it let me introduce some notation. Given an infinite decimal x, I shall write x(n) for the *finite* decimal obtained by truncating x at the $n^{th}$ place after the decimal point. For example, if x is the square root of two, then x(1)=1.4, x(2)=1.41 and x(3)=1.414. It doesn't matter too much, but let us say that if x can be written either as a decimal ending in an infinite string of nines or as one ending in an infinite string of zeros, then we will go for the nines - just to remove the ambiguity from the definition of x(n) above.

To add x and y, the first step is to consider the sequence of finite decimals x(1)+y(1), x(2)+y(2), x(3)+y(3) and so on. Let us have a look at this when x=y=pi=3.141592653589793.... Then the sequence begins 6.2, 6.28, 6.282, 6.2830, 6.28318, 6.283184, 6.2831852, 6.28318530, 6.283185306, 6.2831853070, 6.28318530716, 6.283185307178, 6.2831853071794, 6.28318530717958, 6.283185307179586, ... . Now, given a term in this sequence, you cannot always get the next one by simply putting a new digit on the end, because sometimes you have to modify one or more of the earlier digits. For example, after 6.282 came not 6.282t for some t but 6.2830.

However, it is an easy exercise to show that *no digit is ever modified more than once* . Therefore, the rule for determining the infinite decimal for x+y is the following. The $n^{th}$ digit of x+y is the $n^{th}$ digit of x(n)+y(n), unless that is later modified, in which case it is the $n^{th}$ digit of x(m)+y(m), where m > n is the first (and only) time that the $n^{th}$ digit changes.

We can simplify the above description by defining a notion of limit appropriate for this context. First, let us note that it is easy to put infinite decimals in order: x is less than y if x(n) is less than y(n) for some n. (Here it is important that I define x(n) unambiguously.) Now suppose we have a non-decreasing sequence $(x_1, x_2, x_3, ...)$ of infinite decimals (if they are finite we can make them infinite by putting zeros on the end). Then for any n the sequence $(x_1(n), x_2(n), x_3(n), ...)$ is also non-decreasing. (This is the same sequence as before but with everything truncated at the $n^{th}$ decimal place.) Either this sequence changes infinitely often, in which case it is obvious that the sequence $(x_1(n), x_2(n), x_3(n), ...)$ is unbounded, and hence so is the sequence $(x_1, x_2, x_3, ...)$, or it changes only finitely many times, in which case it eventually stops changing and becomes some fixed decimal that terminates at the $n^{th}$ place. Hence, for any bounded non-decreasing sequence $(x_1, x_2, x_3, ...)$ all the sequences $(x_1(n), x_2(n), x_3(n), ...)$ are eventually constant. If we define y(n) to be the finite decimal that $(x_1(n), x_2(n), x_3(n), ...)$ eventually settles down at, then whenever m > n the sequence y(m) begins with the sequence y(n). Therefore, the finite decimals y(1), y(2), ... are the truncations of an infinite decimal y. This y we call the limit of $(x_1, x_2, x_3, ...)$.

This definition more or less contains a proof of the completeness axiom for infinite decimals, in the monotone-sequences form. (What I have not done is show that this notion of limit agrees with the usual one. That is, however, not very difficult.) And now I can simply define x+y to be the limit of the bounded non-decreasing sequence x(1)+y(1), x(2)+y(2), x(3)+y(3), ... . (Actually, it may sometimes decrease if one of x and y is negative. For negative numbers one could add two integers to make them positive, add those numbers together and then subtract the two integers again. See below for a different convention.)

How might one now prove that addition is commutative and associative? It is very easy: for example, the sequences (x(1)+y(1), x(2)+y(2), x(3)+y(3), ...) and (y(1)+x(1), y(2)+x(2), y(3)+x(3), ...) are equal and therefore have the same limit. A similar remark does associativity. It is clear that 0 is an identity. Inverses are a little bit more complicated because of the problems with negative numbers experienced above. One could define limits for more general sequences than increasing ones. I prefer to define infinite decimals as follows: start with an integer k (positive or negative) and follow it with a decimal point and then an infinite string s of numbers from 0 to 9. This represents the number k + 0.s. For example, by (-2).386... I mean the number that would normally be written -1.613... . Then the inverse of $n.s_1s_2s_3...$ is $(-(n+1)).t_1t_2t_3...$ where each $t_i$ is $9-s_i$. It is not hard to check that these two numbers add up to (-1).999999... which equals zero.

Multiplication can be treated in a similar way. For positive x and y define xy to be the limit of the bounded non-decreasing sequence x(n)y(n), and then define (-x)y to be -(xy) and so on. To make this definition more vivid, let us imagine trying to work out pi squared. The definition above tells us to work out pi(n) squared for larger and larger n, and watch as more and more digits of the decimal expansion gradually settle down and remain constant. Here are the first four values: $pi(1)^2=3.1^2=9.61$, $pi(2)^2=3.14^2=9.8596$, $pi(3)^2=3.141^2=9.865881$ $pi(4)^2=3.1415^2=9.86902225$. Since $3.1416^2=9.86965056$, we already know that the first three digits after the decimal point have reached their final values. These final values are defined to be the digits of pi squared.

It is not hard to prove that multiplication thus defined is commutative and associative and that 1 is an identity element. Once again, inverses present more of a problem, though not much more. Again, let us imagine trying to do the calculations for a specific example - say pi again. One natural way to do it would be to take the reciprocals of pi(n), which would be infinite decimals, and watch them gradually converge. One would obtain a decreasing sequence rather than an increasing one, but it would be easy to extend the definition of limit to accommodate this. Alternatively, one could define pi[n] to be the smallest decimal which is larger than pi and terminates at the $n^{th}$ place so that the sequence 1/pi[n] would now be increasing. Yet another possibility would

be to define r(n) to be the largest decimal that terminates at the $n^{th}$ place and satisfies the inequality r(n)pi < 1. Any one of these will work, though a little bit of effort is needed to show that the resulting number really does give 1 when multiplied by pi.

---

## The existence of the square root of two.

I hope, even if the above discussion falls short of a complete proof, that I have given enough detail to convince you that there are natural definitions of addition and multiplication and ordering of infinite decimals, and that with these definitions they form a complete ordered field. Now let us look at a `naive' approach to the existence of the square root of two (about which further discussion can be found here ).

One reason it seems obvious that the square root of two exists is that one can calculate it as an infinite decimal. If you have worked out the first few digits, then the next one will be the largest number between 0 and 9 such that the resulting number squares to less than 2. For example, $1.4^2=1.96 < 2$ and $1.5^2=2.25 > 2$, so the first digit after the decimal point is 4. Then, since $1.41^2 < 2$ and $1.42^2 > 2$, we find that the next digit is 1, and so on.

I shall now show that this idea forms the basis for a perfectly good and correct proof of the existence of the square root of two. First of all, if we have defined real numbers as infinite decimals, then the procedure just outlined really does unambiguously define a real number. In fact, it is the unique infinite decimal x such that, for every n, $x(n)^2 < 2$ and $(x(n)+10^{-n})^2 > 2$. This doesn't quite prove that 2 has a square root though, since it might be that x, though well defined, doesn't actually square to 2.

Why does this suggestion seem ridiculous? Is it not obvious that x squares to 2? Actually, it is not quite obvious, though it is not too hard either. Here is how an argument would go. If you look back to the definition of multiplication for infinite decimals, you will see that what we have to show is that, as n increases, $x(n)^2$ is a decimal consisting of 1 followed by a decimal point followed by a string of 9s whose length increases to infinity as n increases to infinity. That way, the square of x will be (by definition) 1.999999999.... which (by definition) equals 2.

Let us see, then, why one can choose n in such a way that the string of 9s in $x(n)^2$ has length at least 100. The rough argument goes like this. If you set n=101, then $x(n)^2 < 2$ while $(x(n)+10^{-101})^2 > 2$. However, x(n) and $x(n)+10^{-101}$ differ by only $10^{-101}$, from which it follows, after a brief calculation, that $x(n)^2$ and $(x(n)+10^{-101})^2$ differ by less than $10^{-100}$. From this it follows that $x(n)^2 > 2-10^{-100}$, as we wanted.

The brief calculation is the following: the difference between $(y+c)^2$ and $y^2$ is $2cy+c^2$. If $y=x(n) < 2$ and $c < 10^{-101}$, then this is less than $10^{-100}$. Clearly, a similar calculation shows that if you want a string of m 9s, then you can achieve it with x(m+1) (which, just to remind you, is the truncation at the $(m+1)^{st}$ place of the decimal expansion of the square root of two.)

---

## Continuous functions and infinite decimals.

One of the usual proofs of the existence of the square root of two goes as follows. The function $f(x)=x^2$ is continuous, f(1) < 2 and f(2) > 2, so by the intermediate value theorem there exists x such that f(x)=2. Examining the proof of the intermediate value theorem one sees that it specifies a value of x, namely the supremum of the set A of real numbers y such that $y^2 < 2$. Every step of this usual proof has its counterpart in the argument I have just given in terms of decimals. Let me spell this out.

1. To define the infinite decimal x, I made its decimal expansion as big as possible, taking care only that $x(n)^2$ was less than 2. This corresponds in a natural way to picking the supremum of A above. We can make the

correspondence even closer by redefining the notion of supremum in terms of infinite decimals. To do this, let B be any set of infinite decimals which is bounded above, and define a number x=sup B by setting x(n) to be the maximum of all y(n) such that y is in B. It is not hard to see that x(m) is an initial segment of x(n) when m < n, as is needed, and that the resulting infinite decimal x is the smallest possible upper bound of the set B (obviously x(n) has to at least as big as all y(n) for x to be an upper bound, and equally obviously if, for every n, x(n) is as big as all y(n), then x is an upper bound). Thus, it is easy to prove the least upper bound axiom for our construction. It is also not hard to see that if we let B be the set of all infinite decimals with squares less than 2, then sup B will be exactly the same as the decimal defined earlier.

2. When I showed that we could make the string of 9s as long as we liked, I used the fact that a change late on in a decimal expansion of a number x has only a small effect on $x^2$. This comes to the same thing as saying that the function $f(x)=x^2$ is continuous, and the calculation I did was the usual calculation needed to show this (at least in the range $1 < x < 2$).

From the point of view of decimals, what was the exact property of the function $f(x)=x^2$ that allowed our argument to work? It was the following. For any given m there exists an n such that, in order to work out f(x) to within an accuracy of $10^{-m}$ it is enough to know x up to the $n^{th}$ place after the decimal point - that is, to know x(n). This is a simple reformulation of the definition of continuity in terms of decimals. (See here for a further discussion of continuity.)

---

## The intermediate value theorem.

Using this definition, let us try to prove the intermediate value theorem in a hands-on way. For this purpose, suppose that f is a continuous function from the interval [a,b] to R, with f(a) < c and f(b) > c. How might we try to find x such that f(x)=c? Why don't we try to do it exactly as we did when f(x) was $x^2$ and c was 2? Thus, we should start by finding the largest possible x(1) such that f(x(1)) < c, and then the largest possible x(2) that starts like x(1) and satisfies f(x(2)) < c and so on. If we continue this process, then we end up by constructing the largest possible infinite decimal x with the property that f(x(n)) < c for every n. (Note that this may not be the supremum of the set of all y such that f(y) < c. With a slightly different argument, one can arrive at this x instead.) An easy inductive argument shows that for every n we have f(x(n)) < c and $f(x(n)+10^{-n})\geq c$. Since f is continuous, for every m we can choose n such that f(x(n)) and $f(x(n)+10^{-n})$ are both within $10^{-m}$ of f(x), from which it follows that f(x) is within $10^{-m}$ of c. Since we can do this for every m, f(x) must equal c.

The above argument is a slightly disguised form of another common proof of the intermediate value theorem, which uses repeated bisection. (We, of course, repeatedly split our intervals into ten parts rather than two.) The fact that it is also pretty similar to the proof via the supremum of $\{x:f(x) < c\}$ is an indication that the usual proofs of the theorem are all, fundamentally, based on the same simple and intuitive idea.