Intermediate values and Continuity

In order to find the cube root of a number like 5, we talked about the "bisection method" which proceeds as follows.

- We take $f(x) = x^3 - 5$.
- We find a number, say $x_1 = 1$, such that $f(1) < 0$ and another number, say $y_1 = 2$ such that $f(1) > 0$.
- We define $z_n = (x_n + y_n)/2$ for $n \geq 1$ where x_n and y_n are iteratively defined by:
  - if $f(z_n) \leq 0$, then $x_{n+1} = z_n$ and $y_{n+1} = y_n$.
  - if $f(z_n) > 0$, then $x_{n+1} = x_n$ and $y_{n+1} = z_n$.

As can be seen easily, by an application of induction, $(y_{n+1} - x_{n+1}) = 1/2^n$. Moreover, $(x_n)_{n\geq 1}$ is non-decreasing and $(y_n)_{n\geq 1}$ is non-increasing. Hence we have a least upper bound α of (x_n) and a greatest lower bound β of (y_n). Since the two sequences converge *and* the distance between them goes to 0, we conclude $\alpha = \beta$. So we have a *candidate* α for a zero of $f(x)$.

However, how can we conclude that $f(\alpha) = 0$? In order to prove this, we need to know that $f(x_n)$ converges to $f(\alpha)$ and $f(y_n)$ converges to $f(\alpha)$. Since the first limit is *at most* 0 and the second limit is *at least* 0, we would get the required condition that $f(\alpha) = 0$.

To understand what kind of f will have this property, we must first explore more clearly what one means by the expression $f(x)$.

Functions

In modern terms, a function can be seen to be like a computer program that, given an input x returns a number $f(x)$. The simplest such functions what we can think of are:

- The *constant* function that returns the same fixed value c, whatever input is given to it.
- The *identity* function that, given a number x as input, returns x as its output.

There is a natural way to combine functions. Given two functions f and g we can:

- Create the *sum* $f + g$ of the two functions that, given input x, returns the sum $f(x) + g(x)$ as its output.
- Create the *product* $f \cdot g$ of the two functions that, given input x, returns the product $f(x) \cdot g(x)$ as its output.
- Create the *ratio* f/g of the two functions that, given input x, returns the ration $f(x)/g(x)$ *if* $g(x)$ is non-zero and is *undefined* otherwise.

Note that the last example is a *partially* defined function: it is only defined for *some* values x of the input.

Given a function f and a constant c, we can define the product $c \cdot f$ as the product of the constant function and f. In particular, we have $-f = (-1) \cdot f$ and so subtraction of functions also makes sense. Addition and subtraction of functions together with multiplication of functions exhibits the collection of functions as a *vector space*. This is a very useful way to look at the collection of functions and leads to the study of "functional analysis": the analysis of vector spaces of functions.

By combining the two natural examples of functions using the operations given above, we can make sense of "rational functions" which are given by a formula of the type:

$$f(x) = \frac{a_0 + a_1 x + \cdots + a_p x^p}{b_0 + b_1 x + \cdots + b_q x^q}$$

for some chosen fixed numbers $a_0, a_1, \ldots, a_p$ and $b_0 b_1, \ldots, b_q$. Clearly, for the formula to be meaningful we need b_j to be non-zero for some j. In fact, if the function is to be non-zero, we may as well assume that $a_p \neq 0$ and $b_q \neq 0$ by dropping all terms with coefficient 0.

When comparing functions f and g which are partially defined, we can *only* compare them when *both* $f(x)$ and $g(x)$ are defined for the *same* input x. For example, we have the functions $f(x) = x + 1$ and $g(x) = (1 - x^2)/(1 - x)$. The function g is only defined for $x \neq 1$. However, we *do* have $f = g$ whenever both sides are defined.

Occasionally, we will want to be more strict and insist that for $f = g$, *both* functions should be defined for the *same* inputs x. In those cases we will make this explicitly clear.

Approximation

As we have already seen, numbers are often only specified approximately. One can imagine that a computer program takes two inputs: the value of x *and* an "error bar" k. The computation of $f(x)$ actually gives us a number y with the property that it differs from $f(x)$ by at most $1/k$. Typically, this error bar is given as a power of 10, like 10^r so we can say that we have calculated the function "to r places of decimal".

We must therefore take into account what we can do with functions described by such approximation programs.

One more way to combine functions is "composition", which should not be confused with multiplication! Given a function f and a function g we can form the function $f \circ g$ that, given input x, returns the value $f(g(x))$ *provided* that f is defined for the input number $g(x)$. For example, we can see $h(x) = x^4 + 1$ as $f \circ g$ where $f(x) = x^2 + 1$ and $g(x) = x^2$.

Now, if g only returns an approximate answer, what can we say about h? Similarly, if the input of a function is only given approximately, how good is the approximation in the answer? These questions lead us to the definition of "continuity" of functions.

Continuous Functions

A function is said to be *continuous* if:

> The limit of the values of the function on the terms of a convergent sequence is the value at the limit of the sequence.

In symbols, this is said as follows. Suppose that $(x_n)_{n\geq 1}$ is a sequence of numbers converging with $\lim(x_n)_n \geq 1 = x$. Assume that the function f is defined at all these numbers. Then the sequence of values $(f(x_n))_{n\geq 1}$ converges to $f(x)$.

From the arithmetic properties of limits and the arithmetic operations on functions as defined above, we see that the sum, difference and product of continuous functions yield continuous functions. Moreover, if f and g are continuous functions, then we see that f/g is continuous in the region where $g(x) \neq 0$.

Thus, we see that rational functions as introduced above are continuous functions in the regions where they are defined.

If f and g are continuous and $\lim(x_n)_{n\geq 1} = x$ is a point where $g(x)$ is defined, then $\lim(g(x_n))_{n\geq 1} = g(x)$. So, if $g(x)$ is a point where $f(g(x))$ is defined then, by the continuity of f we see that $\lim(f(g(x_n)))_{n\geq 1} = f(g(x))$. Thus, we see that $f \circ g$ is continuous.

This shows us that continuous functions solve the problem raised in the previous sub-section regarding approximation of functions.

Intermediate Value Property

An important property of continuous functions can be stated by saying that there is "no break" in such a function. More precisely, if $f(x) = a$ and $f(y) = b$ and $a < c < b$, then, provided that f is continuous for all numbers z such that $a \leq z \leq b$, we can show that there *is* such a z for which $f(z) = c$. We note that the special case where $c = 0$ is the question that we started this section with. This property is called the "Intermediate Value Property"; it justifies our intuitive sense of "continuity" in the sense of "without a break".

We follow the bisection method and put $x_1 = x$ and $y_1 = y$. We define $z_n = (x_n + y_n)/2$ for $n \geq 1$ where x_n and y_n are iteratively defined by:

- if $f(z_n) \leq c$, then $x_{n+1} = z_n$ and $y_{n+1} = y_n$.
- if $f(z_n) > c$, then $x_{n+1} = x_n$ and $y_{n+1} = z_n$.

As seen earlier, the sequences $(x_n)_{n\geq 1}$, $(y_n)_{n\geq 1}$ and $(z_n)_{n\geq 1}$ all converge to the *same* limiting value which we call z. We note that (x_n) is an increasing sequence, so $z \geq x_1 = x$. Similarly (y_n) is a decreasing sequence so $z \leq y_1 = y$. Hence, $x \leq z \leq y$.

Next, by continuity of f, we have $\lim(f(x_n))_{n\geq 1} = f(z)$. By the above choice of x_n, we have $f(x_n) \leq c$ for all n. It follows that $\lim(f(x_n))_{n\geq 1} \leq c$; so $f(z) \leq c$.

Similarly, by the choice of y_n, we have $f(y_n) \geq c$ for all n. It follows, that $\lim(f(y_n))_{n\geq 1} \geq c$. By continuity of f, we have $f(z) = \lim(f(y_n))_{n\geq 1}$. So we get $f(z) \geq c$.

Combining the two steps, we see that $f(z) = c$.

Uniform continuity

Given a function $f(x)$ that is defined and continuous for all x satisfying $a \leq x \leq b$ we would like to give approximate values with a chosen error bar. To do so, given any positive integer k, we would like to produce a positive integer n_k so that, for every pair of points x and y lying between a and b that satisfy $|x - y| < 1/k$, we have $|f(x) - f(y)| < 1/n_k$. If this is so, then we can break up the segment of the numbers between a and b into (a finite number) of segments of length at most $1/n_k$. The value of f at any two points of this segment would be close enough to each other, so we only need to determine the value at *one* of these points for each sub-segment to get a table of good approximations of f.

We now show that if f is *any* continuous function, then it has the above property. However, the proof is indirect and uses "proof by contradiction". As a result, it is not easy to determine n_k by going through the proof!

Suppose that there is *no* such n_k. (This "suppose not" is the first step in any proof by contradiction.) In that case, for *every* integer n (which is a candidate for n_k), there would be a pair of points x_n and y_n which lie between a and b and have the property that $|x_n - y_n| < 1/n$, *but* $|f(x_n) - f(y_n)| \geq 1/k$. Since $(x_n)_{n\geq 1}$ is a bounded sequence (all of them lie between a and b), there is a convergent subsequence $(x_{n_p})_{p\geq 1}$. (For example, we know that $\limsup(x_n)$ exists and a subsequence of (x_n) converges to it.)

Since $|x_{n_p} - y_{n_p}| < 1/n_p$ it follows that $(y_{n_p})_{p\geq 1}$ *also* converges and has the same limit as $(x_{n_p})_{p\geq 1}$. Moreover, since x_n's and y_n' lie between a and b, so do the limits. Now, f is given to be continuous, so

$$\lim(f(x_{n_p}))_{p\geq 1} = f\left(\lim(x_{n_p})_{p\geq 1}\right) = f\left(\lim(y_{n_p})_{p\geq 1}\right) = \lim(f(y_{n_p}))_{p\geq 1}$$

On the other hand, we have $|f(x_{n_p}) - f(y_{n_p})| \geq 1/k$ for *all* p. This is not possible. This means that our original supposition that there is no such n_k is wrong!

Note that, in this proof, we crucially used the fact that all the points lie in a *bounded* region of the number line and that f is continuous at *all* these points.

Intervals

The following kinds of regions in the number line will occur often. Hence, it is convenient to give them names.

- The *closed interval* $[a, b]$ consists of numbers x which satisfy $a \leq x \leq b$.
- The *open interval* (a, b) consists of numbers x which satisfy $a < x < b$.
- The *left-open right-closed interval* $(a, b]$ consists of numbers x which satisfy $a < x \leq b$.
- The *left-closed right-open interval* $[a, b)$ consists of numbers x which satisfy $a \leq x < b$.

Henceforth, we will use this terminology to shorten our statements.

ϵ-δ definition of continuity

We note that if f is continuous in some interval $[x_0 - 1/p, x_0 + 1/p]$ around a point x_0, then we can apply the result of the previous subsection to conclude the following. Given any positive integer k there is a positive integer n_k (which must be larger than p!) so that if x satisfies $|x - x_0| < 1/n_k$, then $|f(x) - f(x_0)| < 1/k$.

Note that, unlike the previous sub-section, we are *fixing* x_0. This is why the condition in the previous sub-section is called *uniform* continuity — the *same* n_k works for *all* choices of x_0.

In words, we can put this as follows.

> Given any error bar, the values of $f(x)$ are within that error bar distance from $f(x_0)$ *provided* that x lies within a small enough interval centred at x_0.

It is sometimes inconvenient to state error bars in the form $1/k$, so it is considered convenient to use the Greek symbol ϵ to denote the error bar. The deviation of x from x_0 is required to be bounded by a number specified by the Greek symbol δ. This leads to the "traditional" definition of continuity that will be found in many books.

> A function f is said to be continuous at x_0 if, given any $\epsilon > 0$, there is a $\delta > 0$ so that, for all x such that (f is defined at x and) $|x - x_0| < \delta$, we have $|f(x) - f(x_0)| < \epsilon$.

The equivalence of the earlier definition with this definition can be seen as an application of the Archimedean principle.

Piecewise continuous functions

The Heaviside function is defined by

$$H(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

Now, the sequence $((-1/n))_{n \geq 1}$ converges to 0, but clearly $H(-1/n) = -1$ whereas $H(0) = 1$. So this function is *not* continuous. This is the problem with the continuity of functions defined "piecewise" by using "if-then-else" programming constructions. However, consider the function

$$a(x) = \begin{cases} -x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

On the one hand it is clear that this function is just the function $-x$ in any interval $[-a, -b]$ on the negative side of the number line and it is the function x in any interval $[a, b]$ on the positive side of the number line. What about continuity at 0?

Given any $\epsilon > 0$, we note that if $|x| < \epsilon$, then $|a(x)| < \epsilon$, since $a(x) = |x|$! So, this function satisfies the condition for continuity at 0 given above.

More generally, suppose that f is continuous on the interval $[a, b]$ and g is continuous on the interval $[b, c]$. In addition, *assume* that $f(b) = g(b)$. Then we can define the "join" of f and g by

$$h(x) = \begin{cases} f(x) & \text{if } a \leq x \leq b \\ g(x) & \text{if } b \leq x \leq c \end{cases}$$

It is clear that h is continuous at any point of $[a, b)$ since it is just f in an interval around that point. Similarly, h is continuous at any point of $(b, c]$ since it is just g in an interval around that point. What about continuity at b?

For all given $\epsilon > 0$, we need to find a $\delta > 0$ so that if $|x - b| < \delta$, then $|h(x) - h(b)| < \epsilon$. Now, f is continuous at b and defined on $[a, b]$. So, there is a δ_1 so that if $|x - b| < \delta_1$ *and* x lies in $[a, b]$, then $|f(x) - f(b)| < \epsilon$. Clearly, the condition on x says that $b - \delta < x \leq b$. Similarly, by the continuity at b of g which is defined on $[b, c]$ results in δ_2 so that, $b \leq x \leq x + \delta_2$, then $|g(x) - g(b)| < \epsilon$. Now, take $\delta = \min\{\delta_1, \delta_2\}$. Since $f(b) = g(b) = h(b)$, we easily check that $|h(x) - h(b)| < \epsilon$ for $b - \delta < x < b + \delta$ as required.

What can we do if $f(b)$ and $g(b)$ are different? We can "shift" one of the functions. For example, we can take $g_1(x) = g(x) + (f(b) - g(b))$ where the second term is treated as a constant function. Now, we can apply the above construction to f and g_1 to produce a continuous function by "patching" as above.

This method of patching together continuous functions gives one more method to create continuous functions.

Pitfalls

All of the above discussion *might* give the impression that continuity is *equivalent* to the intermediate value property. However, that impression is **wrong**! There are *pathological examples* of functions that satisfy the intermediate value property but are not continuous. Analysis is full of such counter-intuitive pitfalls. They underline the need to be careful in making our statements accurate and our proofs checkable.