VISUAL ORGANIZERS FOR FORMAL MATHEMATICS

David Tall
Mathematics Education Research Centre
University of Warwick
COVENTRY CV4 7AL
UK

Introduction

Formal mathematics involves definitions and deductions in a manner which is quite different from the mental processes of school mathematics. Formal definitions of function, limit, continuity, differentiation and integration (both Riemann and Lebesgue) involve possibilities that often conflict with the students' previous experience, leading to confusion and alienation. Examples given to "motivate" definitions invariably have specific properties that do not follow logically from the definition itself. For instance, examples of sequences are usually given by formulae so that that the sequence "gets closer and closer" to the limit, without actually reaching it. Consequently, many students believe that this is an essential property of the limit concept. Functions are nearly always given by formulae whose graphs look "smooth" so that students have difficulty imagining anything different. When discontinuities are exemplified by drawing a graph, the picture is often represented as a number of curved pieces with a "jump" at the point under consideration. The result is a widespread belief that a typical function is given by a formula and is continuous except at occasional isolated points. In this way the student builds up a personal concept image of the concepts at variance with the theory.

The formal theory starts from formal definitions and builds up relationships and new concepts through various linear chains of deduction. These are phrased in a combination of words and symbols which can involve far too great a processing strain on the short-term memory of a naive learner. For instance, the ε–N definition of limit and the ε–δ definition of continuity prove initially difficult to commit to memory for meaningful use. After a short time with the general theory, the course usually resorts to specifics such as various tests for the convergence of series which no longer refer back to the definitions and the students cease to worry about them and concentrate on what they need to survive their current problems. In the words of Smith & Moore (1991), students develop *coping skills* to get through the course in a manner which appears on the surface to be successful but may fail to address the subtleties of mathematical reasoning.

The teacher introducing students to formal mathematics may consider a number of possible alternatives to minimise these difficulties. One might be to attempt to avoid all reference to imagery that conflicts with the formal notions. Some very great mathematicians have succeeded in this way. According to Poincaré, Hermite "never evoked a sensuous image", yet he reached a point where "the most abstract entitities were for him like living beings" with "some principle of internal unity" (Poincaré, 1913, p. 220). The mind of a student is not a *tabla rasa* – it already has conceptual imagery which eventually needs to be reconciled with the formal theory. An alternative approach therefore is to provide the student with imagery that challenges their limited conceptions and lays intuitive foundations for the ideas that develop within the formalism.

A computer can provide a rich interactive source of possible imagery, both visual and computational. However, it is essentially a *finite* machine which cannot encompass concepts such as the actual infinity of the real line. There is therefore a genuine conflict in attempting to model formal analysis with pictures drawn on a finite screen using finite processes. Paradoxically, this apparent difficulty can be turned to good effect because the student may be made explicitly aware of the finite limitations and use the imagery and its flaws to stimulate the imagination and conceive of mental concepts that stretch beyond the limitations of finite experience.

Based on a presentation given at the *Technology in College Mathematics Teaching* Conference, November 1992, Chicago, published in new version in In R. Sutherland & J. Mason (Eds.), *Exploiting Mental Imagery with Computers in Mathematics Education*. Springer-Verlag: Berlin (1995), pp. 52–70.

My own quest in recent years has been to use the computer to visualise mathematical concepts in helpful ways in calculus and analysis. Imaginative use of graph-plotters and graphic calculators has enabled students to cope more meaningfully with concepts such as differentiation through the notion of “local straightness”, integration through area summation, and solving (first order) differential equations by visually building up solution curves with given gradient. During this time I became increasingly aware of the limited concept imagery afforded by graph-plotters that only draw reasonably smooth graphs given by formulae. Some allow functions given by different formulae on different domains such as:

$$\mathrm{f}(x)=\begin{cases} x & (x \le 0) \\ 1-x & (x > 0). \end{cases}$$

I wanted to be able to draw more subtle graphs, such as

$$\mathrm{h}(x)=\begin{cases} x & \text{if } x \text{ is rational} \\ 1-x & \text{if } x \text{ is irrational} \end{cases}$$

to use these to motivate deeper thoughts about continuity and more sophisticated ideas of integration, such as the Lebesgue “area” under this curve.

Of course, I was bound to fail. With only rational computer numbers, it is clearly impossible to draw such curves. Or is it? To give up attempting something just because it is theoretically impossible ignores the capacity of the human mind to imagine concepts unachievable in finite time and space.

Instead I set about a more humble task – to use a computer to *simulate* the idea of rational and irrational numbers so that pictures can be drawn to stimulate the visual imagination. To do so requires some kind of distinction between rational and irrational that can be modelled in a computer algorithm. It is the development of such a model, and its use in visualising concepts in analysis that forms the basis of the remainder of this article.

Distinguishing rationals from irrationals

The Ancient Greeks used an algorithm to represent any (real) number x in terms of rational approximations. It begins by finding the integer part n, and decimal part d:

$$x = n + d \quad (\text{where } 0 \le d < 1).$$

If d=0, then x is a (rational) integer. If not, the subtle part is to note that its reciprocal $1/d$ is greater than 1, so we can take the integer part again and write

$$1/d = n_2 + d_2 \ (\text{where } 0 \le d_2 < 1).$$

By continuing this process, the equations can be unravelled to give closer and closer rational approximations to any number x. For instance,

$$\pi = 3 + d \ (\text{where } d = 0.14159...)$$

$$1/d = 7.0626...$$

and so a good approximation to π is

$$\pi = 3\tfrac{1}{7} = \tfrac{22}{7}\ .$$

If the process is applied to a rational number such as $^{22}\!/_{7}$, then the remainder eventually becomes zero:

$$^{22}\!/_{7} = 3 + {}^{1}\!/_{7}$$
$$1/({}^{1}\!/_{7}) = 7+0$$

and the process terminates.

The process gives a sequence of fractions, r_1, r_2, ... which tend to the real number x. If x is rational, the sequence is eventually constant, equalling x expressed in lowest terms. If x is irrational, it is easy to see that the numerators and denominators of r_n must grow without limit. (For if the denominators were all less than an integer N, then the sequence $N!r_n$ would be a sequence of *integers* tending to $N!x$, so the terms must eventually be a fixed integer, implying $N!x$ is an integer, contradicting the fact that x is irrational.)

This gives a method of distinguishing between rationals and irrationals:

> Compute the continued fraction expansion of a number x, and if the rational approximations have denominators which grow without limit, then x is irrational, otherwise it is rational.

Working in the practical world of computers there are technical difficulties. Since the process involves taking reciprocals, if d is small, then $1/d$ is huge. If d should be zero, but errors make it tiny, then taking the reciprocal causes the method to blow up. The practical way out is to cease when the process gives a decimal part smaller than a specified error e, and check if the size of the denominator of the approximating fraction is bigger than a specified (large) number K.

We will therefore make the following technical definition to simulate the notions of rational and irrational in a finite computer world:

Definition: A real number x is said to be *(e,K)-rational* if, on computing the continued fraction approximation to x, the first rational approximation within e of x has denominator less than K, otherwise x is said to be *(e,K)-irrational*.

It is relatively straightforward to define a function $rat(x,e,K)$ to return the value TRUE if x is (e,K)-rational and FALSE if it is (e,K)-irrational. (See Mills & Tall, 1992 for details). When specific values of e, K, are used (say e=10^{-9}, K=10 000), one can define

$$rat(x) = rational(x,10^{-9},10\,000).$$

This partitions numbers into two disjoint subsets. Different values of e and K will, of course, give different partitions, but we will work with specific values of e and K and use the terms *pseudo-rational* and *pseudo-irrational* numbers in this implicit context.

In practice it is important to choose e and K judiciously to make best use of the model. Many computers store numbers to around 8 digit accuracy, so e=10^{-9} is a suitable size. By experiment, K=10 000 proves to be a suitable “large” number. With these values we find that π, $\sqrt{2}$, $\sqrt{3}+1$ are all pseudo-irrational and –1.5, 2/3, 22/7 are all pseudo-rational.

On the other hand, working with (approximately) 8 digit arithmetic, clearly $10^8\pi$ is (almost) a whole number, and registers as a pseudo-*rational*. For PI=3.14159265, it happens that the number 1000*PI is also pseudo-rational. By the opposite token, 1/20 000 turns out to be pseudo-*irrational* (because its denominator exceeds K=10 000).

Even though the partition fails to correspond to the theoretical partition into rationals and irrationals, the partition into two disjoint sets – pseudo-rationals and pseudo-irrationals –

still has useful properties. First the two disjoint sets are intimately intermixed. Then there are many more pseudo-irrationals than pseudo-rationals. Generating random computer numbers and counting those which are pseudo-irrational for e=10^{-9} and K=10 000 gives around 94% pseudo-irrational. In any given system it is best to choose e and K to maximise this percentage, providing a good model for the idea that the irrationals are far more numerous than the rationals (in fact of probability measure 100%).

Plotting highly discontinuous graphs

We now come to the nub of the problem. The following function

f(x) = if(rat(x), x, 1–x)

represents the statement “if rat(x) is true, then the value is x, else it is 1–x”. How can the graph be drawn on a computer screen in a way which truly represents the function? Traditionally the graph is sometimes conceived to look like figure 1.

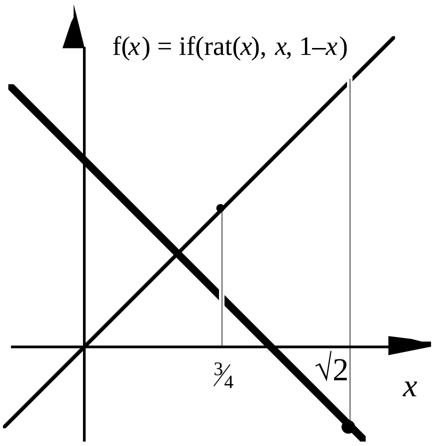


Figure 1 : plotting (¾, f(¾)) and (√2, f(√2)) on the graph of a function defined differently on rationals and irrationals

There seem to be two distinct lines, one with points where x is rational, the other irrational. Perhaps the irrational part looks denser than the rational. But it seems a crazy thing to try to draw in any practical sense. Every point on the x-axis has a unique point on the graph, so every vertical line meets only one of the two lines. For instance, the points x=¾ and x=√2 have corresponding points on the graph on different parts, as caricatured in the figure.

This graph is clearly impossible to represent in its full subtlety on the computer screen, yet by dynamically building up the graph from individual points – some (pseudo-) rational x, some (pseudo-) irrational – a sense of the nature of the graph can be suggested.

Binary Plot

One interesting plotting method is to make successive passes each of which doubles the number of new points – the first pass plots the mid-point, the second the quarters, the third the eighths, and so on, in such a way that points plotted on previous passes are not duplicated on successive passes. This can be done by the following general code.

```
s=(b–a) : x=a+s/2
    repeat
        repeat : plot (x, f(x)) : x=x+s : until x > b
        s=s/2 : x= a+s/2
    until satisfied
```

It starts with a step *s* equal to the interval width *w*=*b*–*a* and a point *x* in the centre of the interval, plots the point (*x*, f(*x*)), then successively adds the step *s* to *x* and repeats the plot until the point *x* moves beyond the end of the interval. The process is repeated, with a new step *s*=*s*/2, starting at *x*=*a*+*s*/2. On the second pass, this starts at point *a*+*w*/4, step *s*=*w*/2 and plots the remaining quarter points *a*+*w*/4, *a*+3*w*/4. The next pass starts at *a*+*w*/8, step *s*=*w*/4 and plots the remaining eighth points *a*+*w*/8, *a*+3*w*/8, *a*+5*w*/8, *a*+7*w*/8, and so on.

The process is repeated until a pre-determined condition is satisfied (say a specified number of passes, or a specified time, or until a pre-determined key on the keyboard is pressed). Each pass doubles the density of the plot. Note, however, that the points plotted are all of the form $a+mw/2^n$ where *m* and *n* are positive integers with $m<2^n$. In particular, when *a* and *b* are rational, the points plotted are all rational (figure 2).

In practice, if a computer plot is carried out with a rational *x*-range, say from *x*=–2 to 2, then the first dozen or so passes give all rational points. If the plotting mechanism is left to run for a long time, numerical errors eventually produce pseudo-irrationals.

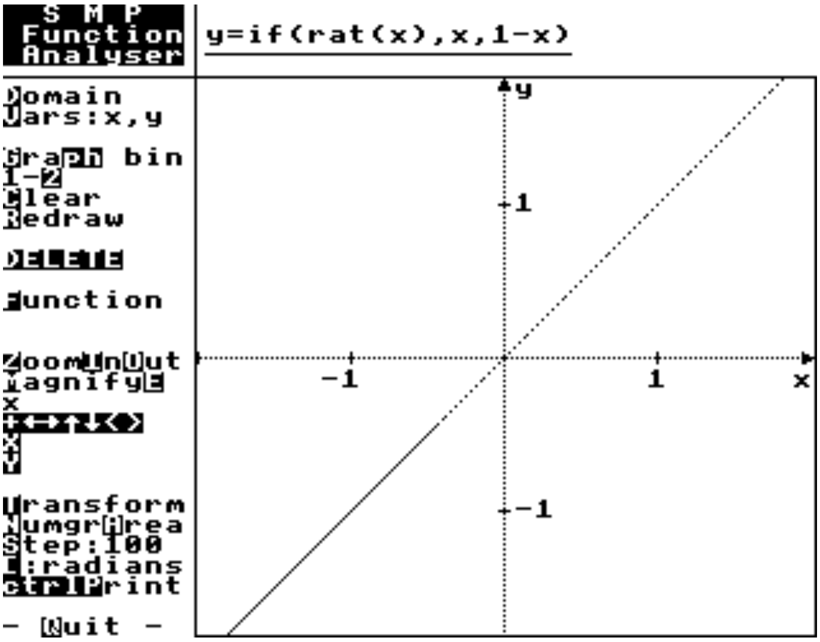

y=if(rat(x),x,1-x)

Figure 2 : a binary plot, initially producing only (pseudo-)rational points

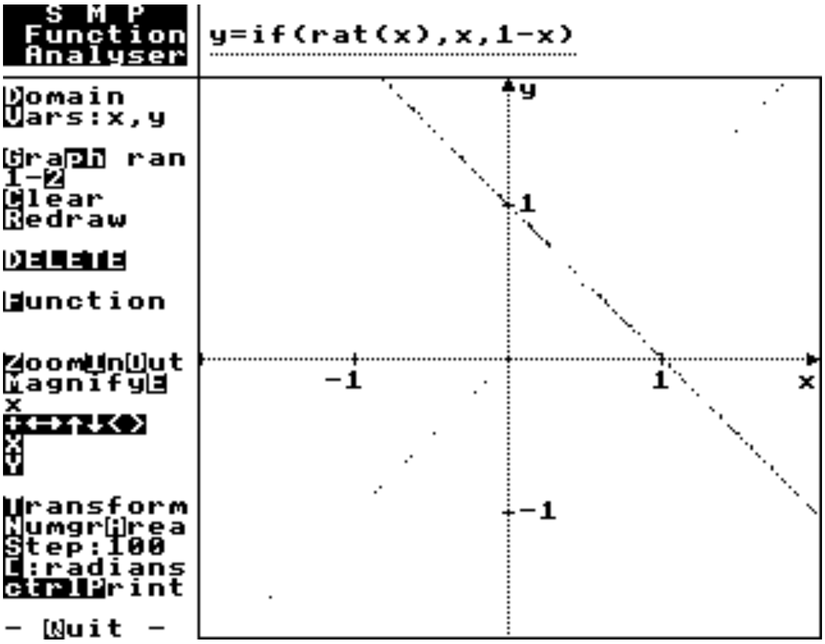
Random Plot

Another method is to calculate (pseudo-)random numbers in the interval [*a*,*b*] and plot the point (*x*, f(*x*)), until a pre-determined condition is satisfied (a certain number of points is plotted, a certain time has passed, or a selected key has been pressed).

```
repeat
    x=a+random*(b–a)
    plot(x,f(x))
until satisfied
```

This method simply sprays points on the graph where the *x*-value is calculated as *a*+*random**(*b*–*a*) where *random* is a random number between 0 and 1. *In theory*, such points have a high probability of having *irrational* values of *x*, based on the fact that a randomly chosen number has a low probability of having a repeating block of digits repeating in its decimal expansion.

In the context described, about 94% of the points plotted have pseudo-irrational *x*–values (figure 3).



y=if(rat(x),x,1-x)

Vars:x,y

Step:100

Figure 3 : a random plot producing mainly (pseudo-)irrational points

To get only pseudo-irrationals, the code could be modified to

```
repeat
      x=a+random*(b–a)
      if rat(x)=false then plot(x,f(x))
until satisfied
```

to exclude pseudo-rationals. However, in using the program to stimulate the imagination, it is important that the student *sees* that the model is only an approximate one. The existence of rogue values helps to focus on the difference between theory and practice.

Mixing plot routines

If the plotting program is set up with an easy switch between the two plot mechanisms then it is possible to draw points with both pseudo-rational and pseudo-irrational *x*. If the program includes a routine to specify *x* and plot (*x*,f(*x*)), then this shows that for each value of *x* there is just *one and only one* corresponding value on the graph. For instance, if *x*=3/2, then *y*=0.75, but if *x*=√2, then *y*=-0.4142135 (figure 4).

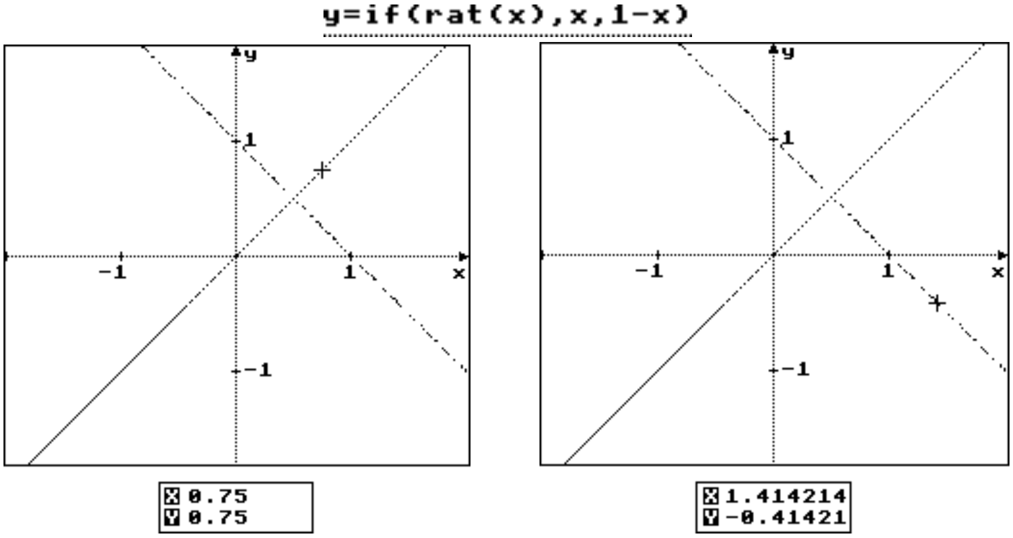


Figure 4 : plotting (*x*,f(x)) for *x*=3/4 and *x*=√2

Visual links between continuity, differentiability and integrability

There is much reference in the literature to the use of magnification to motivate the idea of gradient of a function through "local straightness". An overview is given in the MAA Notes on *Visualization in Mathematics* (denoted later as [VM]). What is less well understood is an appropriate intuitive notion of continuity (which tends to be linked with the connectedness of the graph with subtle inferences from the completeness of the reals). In the following sections I shall briefly recapitulate some of the central intuitive ideas from [VM] and extend them to deal with the wider visualisations possible using the plotting routines discussed in this article.

Differentiability

The notion of differentiable function can be motivated simply by magnifying its graph, *retaining the same relative scales on the axes*. The graph of such a function magnifies to look straight and the gradient of this straight line segment is the gradient of the graph. This visual image proves to be a powerful gestalt for the notion of gradient function, though care should be taken to link graphical, numerical and symbolic representations. As a bonus, it stimulates the imagination by allowing the student to conceive of a *nowhere differentiable function*, which is one which remains wrinkled, no matter how much it is magnified. Again this is well-represented in the literature and fractal graphs such as the blancmange function can be plotted which illustrate this principle (figure 5). (See [VM] for details.)

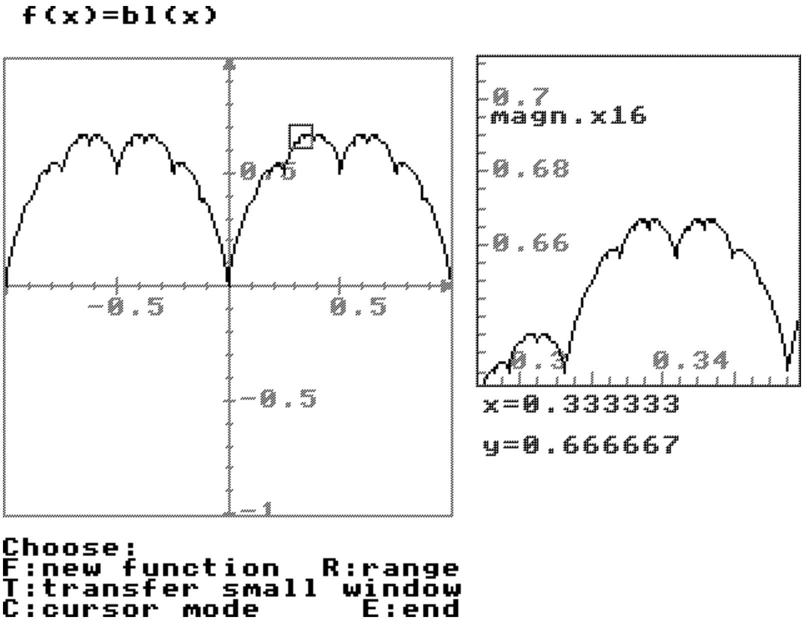


Figure 5: a highly wrinkled function, nowhere locally straight,
magnified at *x*=1/3 from the main window to the small window

It is also possible to motivate other strange looking graphs. For instance, figure 6 shows a graph whose derivative is 0 everywhere the derivative is defined except at one point where the derivative is 1. By magnifying the graph at *x*=0 it magnifies to look like a line of gradient 1 (despite the many jump discontinuities arbitrarily close to the origin). The veracity of this strange property, having been motivated visually, can, and should, be checked from first principles.

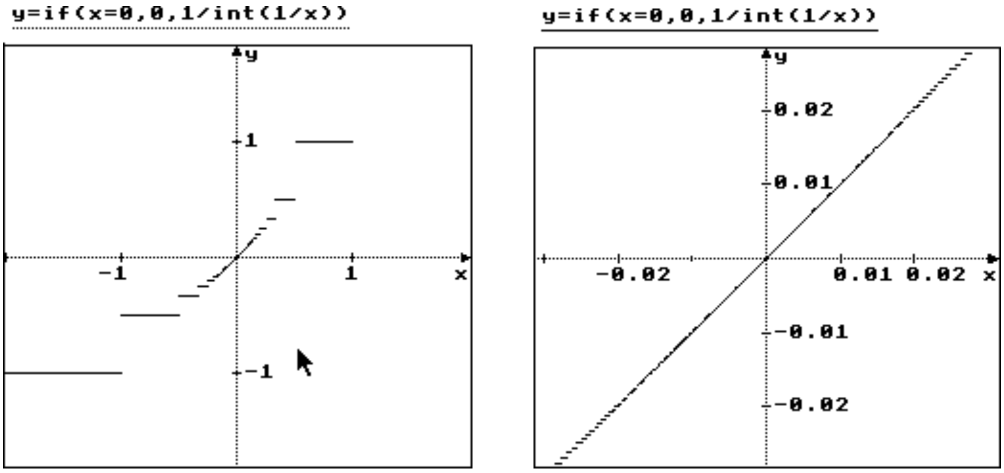


Figure 6 : A graph with derivative 1 at the origin, but zero everywhere else it is defined

Continuity

The intuitive idea (which proves to be flawed) that a continuous function is “one whose graph can be drawn *continuously* with a pencil, without taking the pencil off the paper” can be used to motivate the formal definition. Given a graph drawn “continuously” in this intuitive sense, simply by stretching it horizontally, keeping the vertical scale constant, pulls the picture of the graph in a window out flat (figure 7).

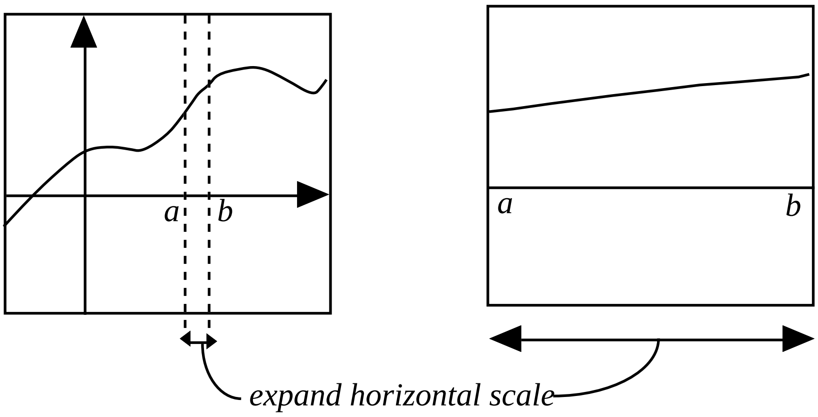


Figure 7 : stretching a graph horizontally

This can be performed on a computer to give a horizontal graph. (figure 8).

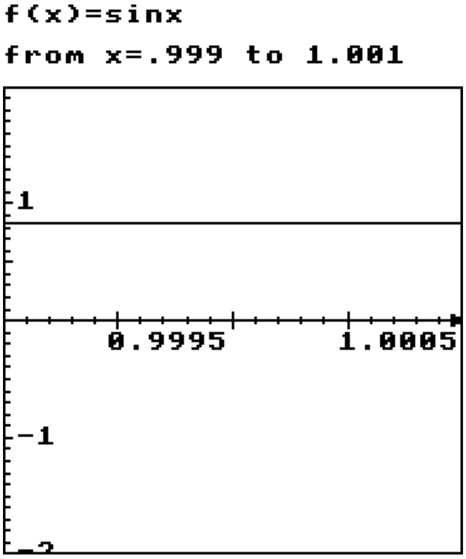


Figure 8 : horizontal stretching of the sine graph

The graph is captured in a horizontal line of pixels which can be thought of as being height f(x_0)±ε (figure 9).

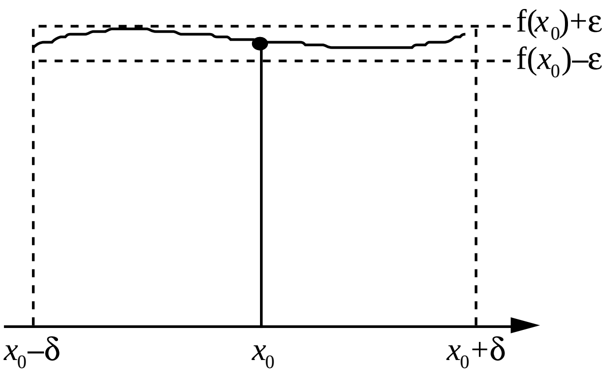


Figure 9 : the concept of continuity through horizontal stretching

So the fact that a small enough x-interval width x_0±δ can be found to draw such a picture is enshrined in the property:

> Given ε>0, there exists δ>0 such that $x_0 - \delta < x < x_0 + \delta$ implies $f(x_0) - \varepsilon < x < f(x_0) + \varepsilon$.

The relationship between (visual) continuity and (visual) differentiability is clearly that a graph which is locally straight can be stretched horizontally to look flat. So *continuity implies differentiability*. But the image of wrinkled graphs shows that *there are continuous functions which are not differentiable anywhere*.

It is also possible to motivate other seemingly strange ideas. The graph of the function f(x)=if(rat(x),x,1–x) seems highly discontinuous. Yet on keeping the y-range fixed and stretching a smaller and smaller x-range either side of x=½ to fit the fixed window eventually draws the graph flat. The graph is *continuous* at x=½, but *nowhere else* (figure 10)

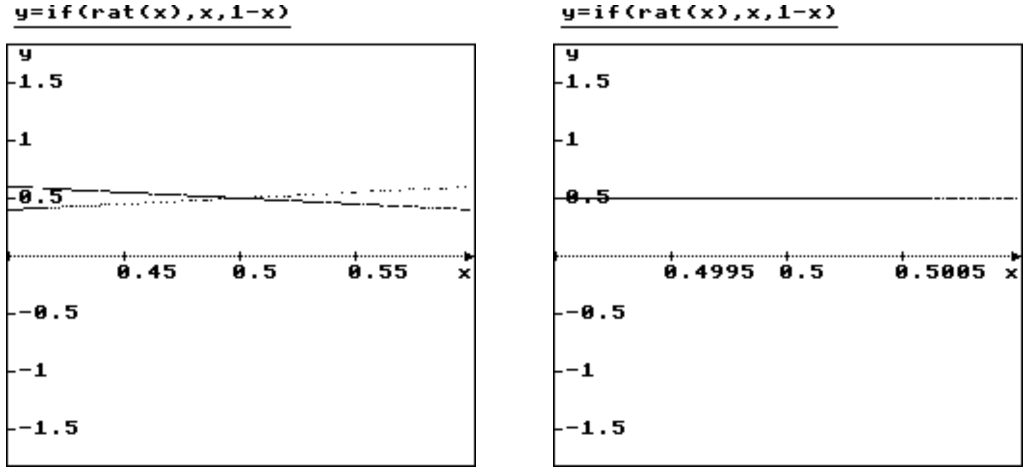
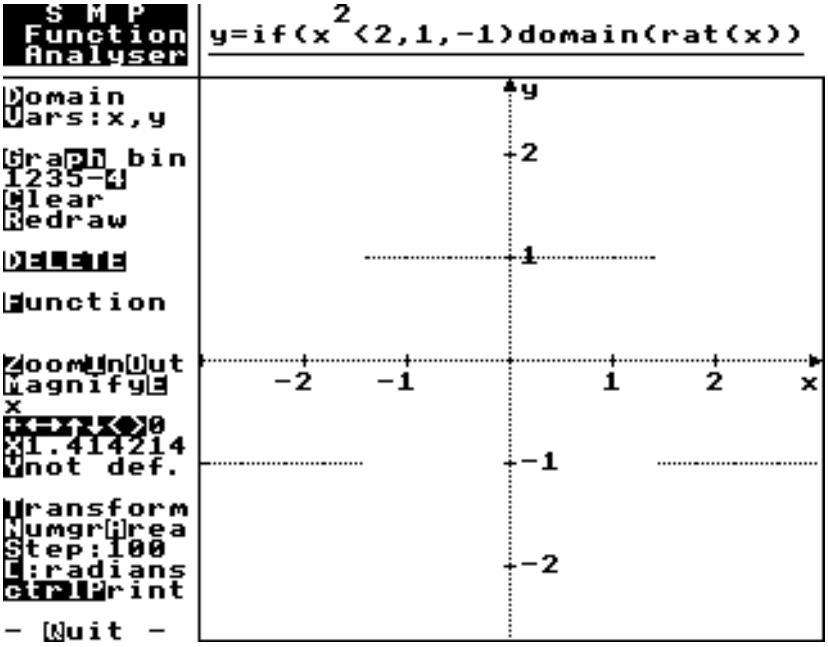


Figure 10 : stretching horizontally to see if(rat(x),x,1-x) is continuous at x= ½

Figure 11 shows a graph whose domain is restricted to the pseudo-rationals which *appears* to have a discontinuity. But on inputting x=√2, the value is declared *not defined* because √2 is not a pseudo-rational. The graph is not defined on the very point where the jump occurs. On the other hand it *is* continuous at every point in the domain.



$y=\text{if}(x^2<2,1,-1)\text{domain}(\text{rat}(x))$

X1.414214
Ynot def.

Figure 11: a graph defined on a (pseudo-) rational domain with an apparent discontinuity

With this latter graph the intuition is stimulated to focus on the notion of continuity as a pointwise definition *on the domain of the function concerned*. The pictures are now stretched to their practical limits and the human mind must take over to contemplate the underlying formal mathematics.

Integrability

When integration is seen as the reverse of differentiation, the role of continuity becomes obscured. However, when integration is given an independent meaning through summation, the role of continuity becomes more focused. Calculating the area A(x) from a fixed point x_0 to a variable point x under a continuous function f(x) may be seen as giving an area function satisfying $A'(x)=f(x)$. This is because the area from x to $x+h$ is $A(x+h)–A(x)$ as in figure 12 and, for small h, looking at a stretched picture which pulls the graph flat gives a visual approximation of $A(x+h)–A(x) \approx f(x)\times h$. So $\frac{A(x+h)–A(x)}{h}$ approximates to f(x) and motivates the fundamental theorem $A'(x)=f(x)$ only on the assumption that f(x) is *continuous*.

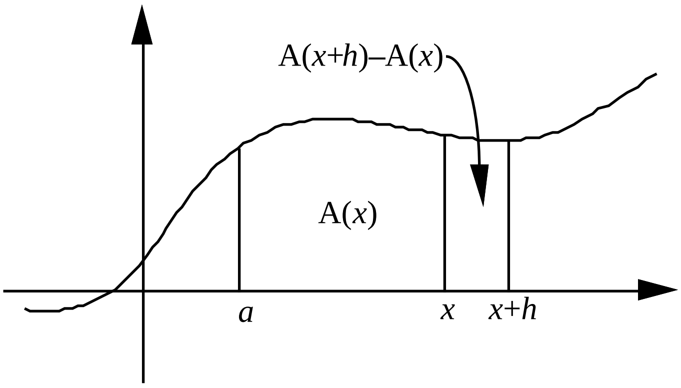
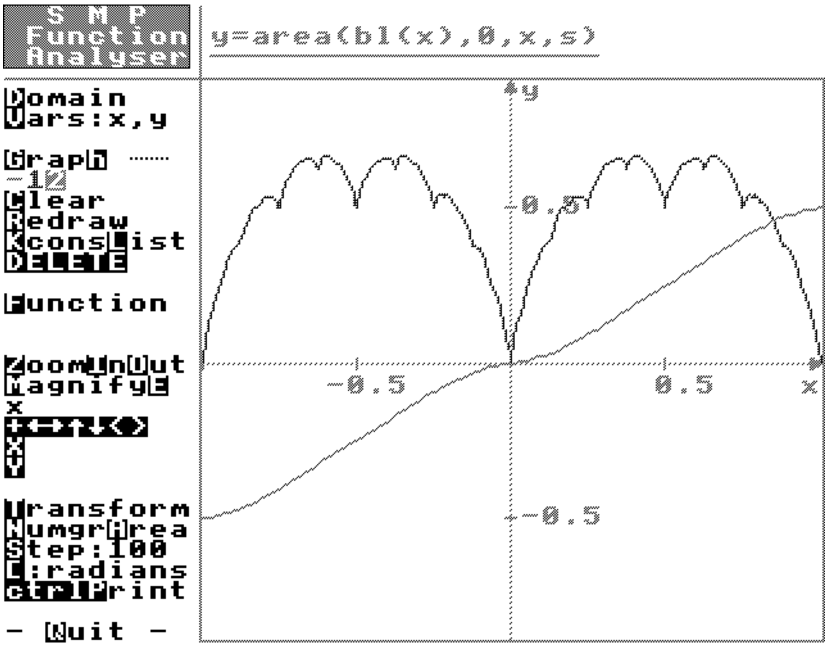


Figure 12 : the change in area

This means that if a function f(x) is continuous but not differentiable, then its area function A(x) is differentiable *once* (to get f(x)) *but not twice*. This can be illustrated by computing the area under the blancmange function. In figure 13 the increasing function is the numerical area under the blancmange calculated from the origin (for both positive and negative steps), using the mid-ordinate rule with steps width 0.1. See [VM] for details.

It is even possible to compute the numerical gradient of the area function. The blancmange looking function in figure 13 is in fact the graph of this gradient, motivating the fact that A′(x) is the original function.



y=area(bl(x),0,x,s)

Vars:x,y

Step:100

Figure 13 : the area function for the blancmange and its derivative

Integrating discontinuous functions

What happens when we attempt to compute the area under a function which is *not* continuous? The numerical calculations to compute the area approximations work in the usual way, but can produce insightful pictures. A graph with simple discontinuities such as f(x)=x–INT(x) proves to have an area graph (the dotted curve in figure 14) which is locally straight, except at the points where f is discontinuous.

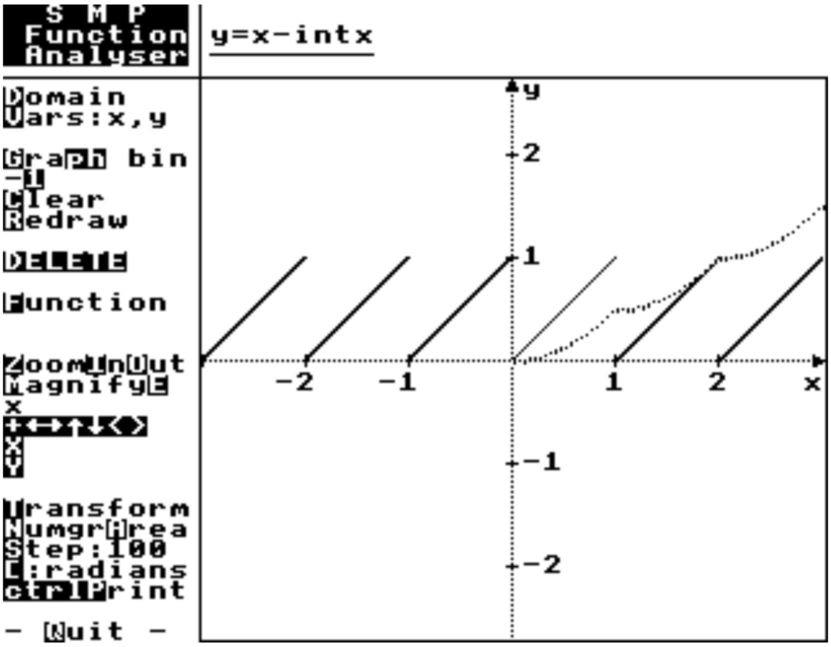


Figure 14 : The area under a curve with isolated jump discontinuities

Many early area plotters simply calculated the area and plotted the cumulative values *as a set of points*. In figure 15 the area is calculated *as a function* area(*f*,*a*,*b*,*h*) of *f* from *a* to *b* with step *h* (using the mid-ordinate rule, here with *h*=0.05). It is then possible to zoom in on the graph at an appropriate points to see where it has different left and right gradients. Note that at *x*=1, the graph of *x*–int(*x*) has left and right limits which are different (one and zero respectively); these are the gradients of the area graph to the left and right.

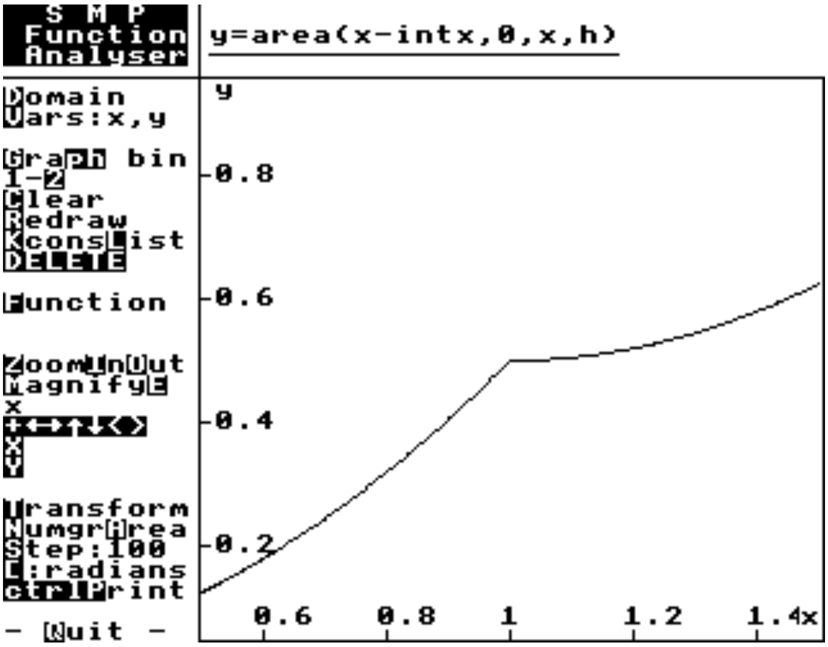


Figure 15 : zooming in on the area curve at *x*=1

What happens when we attempt to compute the integral under a highly discontinuous graph such as f(*x*)=if(rat(*x*),*x*,1–*x*) ?

This proves to be highly interesting. Attempting to use a mid-ordinate rule to compute the area starting at a rational point with a rational step, clearly (mainly) pseudo-rationals will be encountered. Thus the area under f(x) from 0 to 5 using this method is an upward facing parabola, of the approximate form $x^2/3$. The area under $y=x$ from 0 to 5 is 25/3 = $8\frac{1}{3}$. The difference between this and the result 7.51875 is due to errors in arithmetic giving a few pseudo-irrational points in the plot.

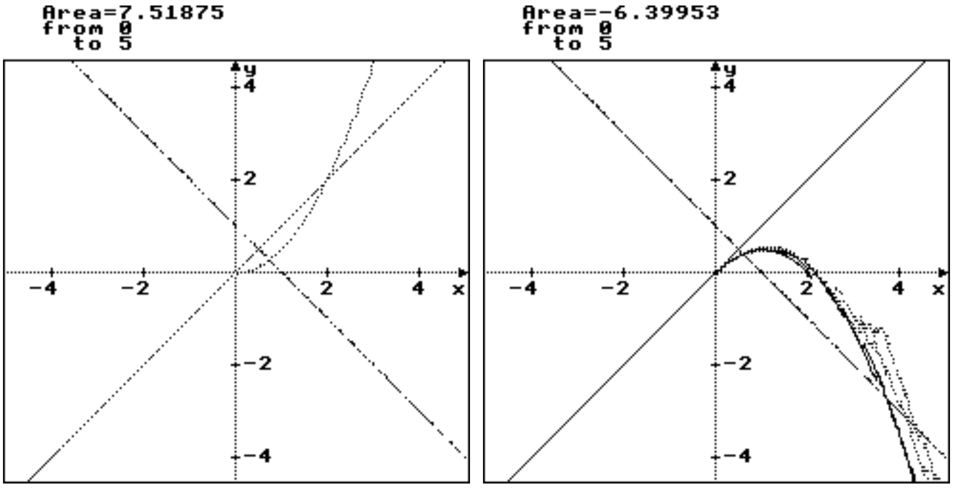


Figure 18 : calculating the area under the graph with rational and random steps

But a calculation with a *random* step-length, using strips whose height is computed at a *random* point in the strip produces a very different picture. The second picture shows the result after plotting the area from 0 to 5 several times over. There are several distinct but fairly close curves. A random value is highly likely (probability about 0.94) of being a pseudo-rational with the specific parameters for computation used in the program. Thus *most* of the values lie on f(x)=1–x. Since $\int_0^t (1-x)\, dx = t-\frac{1}{2}t^2$, the graph will lie roughly on this curve but may diverge as the occasional pseudo-rational distorts the value. The actual value of –6.59953 compares approximately with the value of 5–12.5=–7.5. Some of the other computations were closer.

The interesting stimulus arising from the latter is that, in theory, irrationals are highly more likely to arise from random partitions than rationals. Using such a partition therefore may give an area which approximates to $x-\frac{1}{2}x^2$. This can motivate the possible development of a more powerful theory of integrations – the theory of Lebesgue in which domains such as the irrationals inherently contribute more to an integral than the smaller domain of rationals. The step is, of course, a great one. It involves questions of cardinality and infinite series, which requires a leap in the imagination of the human mind. But it is one which is better taken by a mind prepared for the journey.

Classroom experiences

The ideas in this article were piloted in a 60 hour analysis course for mathematics education students planning to be teachers. Traditionally such students have struggled with the subject. For instance, the previous year's students were asked, after two weeks instruction in the ε-N definition of the limit of a sequence, to write down the definition. *None* of them were able to do so. Discussion on the mathematics was rare and there was sometimes a sense of alienation from the subject, as if it belonged to a different universe.

The decision was taken to revamp the course, using the ideas outlined here to make it more visual, and to encourage discussion techniques to build up the concepts. It was not expected that students would obtain greater facility with the formal aspects, but that they were more likely to be able to visualise and verbalise the concepts. This proved to be the case. Formal definitions (even if remembered!) are long and complex and usually need to be *written down* to be able to grasp them as a whole. Visual ideas prove to be easier to discuss in everyday language. The formal concepts were based on more meaningful conceptualizations.

The level of discussion was mature and often insightful. For example, on seeing the graph of the area function of the blancmange, a student observed that the area function must be “differentiable once but not twice”, before this was formally discussed. Few traditionally taught students have any idea what such a function might look like.

On another occasion, when the random plot was being used, another student suggested that random points were far more likely to be irrational, and so the values of the function on rational numbers are less important when random methods were used to calculate the area for the function in question. This led to a reconsideration by the group of cardinal concepts that had caused them so much trouble two years ago when they had a formal course on sets and foundational topics.

The discussions indicated a level of visualisation and verbalisation far greater than had been traditionally expected in earlier courses and they were by no means limited to a minority of the students.

Dealing with these visual concepts requires careful focusing and guidance to distinguish between the theoretical mathematics and the finite images on the computer. Given a supportive environment, students may confront the conflict to produce a more meaningful foundation for the theory.

References

Smith, D. A. & Moore L. C., (1991). Project CALC: An Integrated Laboratory Course. In Leinbach *et al* (Eds.), *The Laboratory Approach to Teaching Calculus*, MAA Notes Volume 20, Mathematical Association of America, 81–92.

Mills, J. T. S. & Tall, D. O., (1992). “Modelling Irrational Numbers in Analysis using Elementary Programming”, *The Mathematical Gazette*, 76, 243–250.

Tall, D. O., (1991). “Intuition and rigour: the role of visualization in the calculus”, *Visualization in Mathematics* (ed. Zimmermann & Cunningham), MAA Notes Volume 19, 105–119, (designated as [VM] in the text).

Poincaré, H., (1913). *The Foundations of Science* (Trans. G. B. Halsted), reprinted by University of America Press, 1982.