From the visual to the logical in mathematics

John Mills & David Tall
Mathematics Education Research Centre
University of Warwick
COVENTRY CV4 7AL

Mathematical concepts are more than just the logical sequence of definitions and deductions which make up the formal framework of the subject. In mathematical research it is first necessary to develop a framework of linkages between ideas before they are sorted out into a precise deductive sequence. Yet, in teaching, the subject is so often presented in its final organized form in a way that the majority of undergraduates seem to find notoriously difficult. We believe that the a major source of the students' difficulty lies in the lack of an appropriate intuitive background to provide an environment within which the formal ideas can be interpreted and hence understood.

The arrival of more powerful computers offers the opportunity of manipulating concepts visually in real time, and this facility can be put to good use in gaining insight into the logical relationships involved. The recent introduction of the RISC (reduced instruction set) chip in the Archimedes computer gives such a gain in speed that it is now possible to draw highly irregular curves in an interactive manner to give new insight into subtle theorems of mathematical analysis.

As an example, let us cite the theorem

> "the integral function I of a Riemann-integrable function f is continuous, and where f is continuous, I is differentiable".

There are two ways in which this may be interpreted: as a sequence of purely logical deductions from the definitions of the concepts involved, and a broader interpretation which gives the individual an understanding as to why the theorem is true, not only in showing how the continuity of f meaningfully leads to the differentiability of I, but also how *discontinuity* of f can sometimes (but not always!) lead to I being *non*-differentiable.

In this article we shall begin by establishing some fundamental principles in the calculus, which are somewhat different from the *definitions* of the concepts, yet form the basis for a better *understanding* of the concepts. The fundamental principles emphasise the *modelling* aspects of the calculus, and are therefore highly relevant to real-life applications of the theory. They suggest that we should fundamentally reassess the way in which we view the structure of the calculus both from a practical and a theoretical viewpoint.

The generative idea we take in differentiation is the notion of "local straightness", this is the property that a differentiable function is one whose graph "looks straight" when viewed under a microscope of sufficient. This seems initially a crude idea that requires further precision, but it has two vital ingredients. First the idea is one which, in a learning context, appeals immediately to the beginner, and second, its implications stretch right through the theory to formal ideas of differentiable manifolds, leading to concepts in both standard and non-standard analysis.

We will use the power of the Archimedes RISC chip to show models of both differentiable and non-differentiable functions, to gain insight as to why theorems in analysis might arise as a result of these ideas. Then we will look at the Fundamental Theorem of Calculus to see how the formal idea of continuity arises as a natural ingredient, and see why *continuity* is the generative idea for this theorem rather than

Published in *Bulletin of the I.M.A.* **24** 11/12 Nov–Dec, 176–183 (1988).

differentiability. In particular we will see the area function I for a continuous non-differentiable function n growing in real time to see why it is that the area function will be quite smooth and have n as its derivative. In particular we will be able to visualize a function I which is differentiable everywhere once, but nowhere twice. Furthermore, if we look at a second-order differential equation of the form:

$$\frac{d^2y}{dx^2} = n(x)$$

then we will be able to visualize the solution as a function which is differentiable every twice, but nowhere three times.

1. Local straightness

The concept of local straightness is exemplified by magnifying the graph of $y=x^2$ at any point. For example the graph at $x=1/2$ magnifies to give a curve which appears virtually straight, with gradient 1. Given suitable software it is possible to superimpose the graph of $y=x-1/4$ to make the point clearly (figure 1).

Figure 1

Notice that the computer picture of the line is made up of pixels, so it is necessary to discuss how the computer represents a curve. The highlighting of pixels may be compared with trying to represent a straight line drawn across a chess-board by only highlighting squares. It illustrates that the computer picture is only a *model* of the situation and emphasizes the need, once the visualization has been achieved, for a closer analysis of the calculations. The graph and the tangent line may *look* coincident within the limited discrimination afforded by the computer model but they are only indistinguishable within the accuracy of the picture.

The “local straightness” on magnifying the graph centred at x=a can clearly be expressed in the form

$$\frac{f(a+h)-f(a)}{h} \approx \text{constant}$$

for *any* “suitably small” h, depending on the accuracy of the computer screen. From this idea can be drawn the beginnings of the ε–δ definition of the derivative.

It should be noted that, during the course of the magnification,

equal scales must be used to see the true gradient. There is a school text-book, which shall remain nameless, that asks the student to draw the graph of sinx from 0 to 180 (in degrees) and then to obtain a gradient of 1 at the origin. Using unequal magnifications produces other interesting ideas, as we shall see later when we come to consider the concept of continuity.

More exotic examples to strengthen the learner’s grasp will need to show not only non-differentiability, but also such things as the possible non-continuity of the derivative.

The well-known function $f(x)=x^2\sin(1/x)$ exhibits a number of interesting properties, for example, the shape for *large* x (and how ‘large’ is ‘large’). Or the more usual study of what happens *near*, and *at* the origin. Zooming in near, but not including, zero reveals a sine curve (with gradient varying between ±1), but magnifying the graph centred on zero (taking $f(0)=0$) shows the curve to be flat (figure 2). The function is differentiable everywhere but the gradient visibly has a discontinuity at the origin.

Figure 2

A practical method to sketch a tangent to a curve at a specific point is to place a ruler ‘inside the curve’, in such a position that equal portions of the curve are just visible on both sides of the point (figure 3).

Figure 3

This suggests that the very reasonable expression

$$\frac{f(a+h)-f(a-h)}{2h}$$

could be used as an approximation to the derivative. Graphs with ‘corners’ (different left and right gradients at a specific point) show the inherent fallacy of this attempted definition. This is clearly seen from tbe simple example $|x|$, or the more interesting function, the maximum of sinx and cosx, defined as

$$d(x)=\max(\sin x,\cos x).$$

In this case, a line drawn from $x-h$ to $x+h$ for $x=\pi/4$ in no way represents local straightness: zooming in shows the graph is not locally straight at this point (figure 4).

Hence the necessity for the one-sided traditional limit $\frac{f(x+h)-f(x)}{h}$ is established together with the fact that the limit must be the same for h both positive and negative. Students (and many teachers) have an implicit belief in the existence of the tangent, even at a corner, as a line that “touches the graph at one point only”. Examples such as these show that a graph has a tangent at a given point if and only if it has a (two-sided) derivative there.

Figure 4

Initial experiences with the computer might suggest that it is only necessary to visualize the gradient using a “small h” rather than use the formal limit. However, a specific scale may fail to exhibit even tinier variations in detail which are revealed by a higher magnification. The function

$$h(x)=\cos(1000x)/1000$$

produces tiny oscillations and, although the graphs of d(x) and d(x)+h(x) look identical on the screen when drawn from 0 to 2π, a much higher magnification reveals the tiny differences between the curves which lead to considerable differences in their gradient. For the graph of d(x)+h(x) a value of h several degrees of magnitude smaller than before is necessary to get a reasonable estimate for the gradient. Meanwhile, magnifying the curve at x=π/4 shows the “corner” still persists (figure 5). The only way to be sure that the gradient is correctly calculated is via “an appropriately small value of h”, which will depend on the nature of the graph and possibly vary at different parts of a given graph, leading to the necessity to introduce the formal derivative.

Figure 5

The examples so far described can give the student the impression of non-differentiability as an occasional, localized phenomenon, much as it was viewed in the nineteenth century before the full range of possibilities became apparent.

A function with an infinite number of discontinuities is

$$g(x)=1/[1/x]$$

(where square brackets denote the “integer part” function, and we also define g(0)=0).

The derivative, everywhere it exists, is seen to be zero by zooming in, except at the origin, where it clearly looks to be 1 (figure 6) Here is motivation to calculate the gradient of the function at x=0 from first principles (not too difficult a task). The result is confirmation of a very nasty beast indeed, with a derivative of 1 at the origin and a derivative of 0 “almost everywhere else”, i.e. except at a sequence of points where the function is defined but discontinuous. Magnifying the graph at the origin gives a picture which *looks* like a straight ine of gradient 1, but anywhere else, even near the origin, the graph either has a jump discontinuity or has derivative zero.

Figure 6

Given an adequate drawing package on a computer it is possible to look at some really awkward cases. The Function Analyser for the Archimedes computer encourages the use of function notation to look at difficult functions such as g(d(x)), giving real challenges to sketch and interpret such graphs, on occasion testing the computer software beyond its limits!

2. Nowhere differentiable continuous functions

It is now possible to draw models of everywhere continuous, nowhere differentiable functions as integral parts of graph plotting programs. Appropriate functions can be calculated by algorithms which are simpler to calculate than familiar functions such as sine or cosine, and given sufficient computing power (available on many computers in compiled languages such as C and Pascal, and on the Archimedes in interpreted BASIC!) they may be drawn at a most satisfactory speed. This offers not only visual insight to stimulate the thinking about the logical construction of such functions, but also a flexible context in which the combinations of these functions may be drawn and investigated.

One such function was christened the “blancmange function” by John Mills (figure 7), because of its shape, (Tall 1982). We were amused, however, to find that the French do not recognize this Franglais term at all; their French word for this culinary creation is “pudding”. Other nations are very patriotic in their attributions: the blancmange function is called the “Takagi function” in Japan, after the Japanese mathematician who described it in (Takagi 1903) and in Holland graphs of this type are named after Van der Waerden, who used them to give a simple proof of non-differentiability.

Figure 7

The general case of a Van der Waerden function is constructed by starting with the saw-tooth

$$s(x) = \min(x-[x], 1-(x-[x]))$$

defining

$$s_n(x) = \frac{s(k^n x)}{k^n} ,$$

and setting

$$\text{van}(x,k,m) = s_0(x)+s_1(x)+...s_{m-1}(x).$$

(Figure 8.)

Figure 8

The kth Van der Waerden function is the limit of this as m increases. The blancmange function is the particular case when k=2.

The whole thing could be a formal nightmare to the uninitiated. But visually it is easy to see. To construct the kth Van de Waerden function, start with a saw-tooth s(x) which moves up to a half and down again in every unit interval. Now take a graph reduced in size to have k such teeth in each interval and add it on. Successively add on smaller teeth reduced in size by a factor k each time. The successive approximations to the Van der Waerden functions will soon stabilize to give a satisfactory picture. (Figure 8(c) superimposes van(x,3,n) for n=1,2,3,4 and the stabilization is already evident on-screen. The Archimedes computer holds “real numbers” only to an accuracy of 32 binary digits (to utilize the speed of its 32 bit processor), so the practical limit of computer accuracy will be reached adding on the mth tooth where $1/k^m$ is less than $1/2^{32}$. For k=2 this requires m larger than 32, k=3 requires m>21 and k=10 requires m>10. Thus for the most accurate models possible on the Archimedes require approximations such as van(2,x,32) or van(10,x,10). Provided that they are not pressed too close to the limitations of numerical accuracy, they give most satisfactory visual presentations.

Logically these pictures can motivate ideas concerning the limits of sequences in analysis. For each x, $s_m(x)$ is a positive number between 0 and $1/(2k^m)$, so the sum of m terms van(x,k,m) does not exceed

$$\tfrac{1}{2}(1+1/k+...+1/k^{m-1}) ,$$

which is less than $\frac{1}{2}$ /(1-1/k) and is certainly less than 1. For any x,k the sequence van(x,k,m) is a monotonic increasing sequence in m, which tends to a limit less than 1. Hence it is (mathematically) straightforward to give a formal proof of the result, and once more link the visual idea of the limiting process to the formal notion.

3. The fundamental theorem of calculus and continuity

The Riemann integral may be simulated on the computer by drawing the area under the graph using strips of appropriate height. Consider the simple example of sinx from 0 to 2π, using a fairly coarse strip-width, say c=0.5, with the mid-point approximation for the height of each strip. This gives a surprisingly close estimate.

But what happens when we add thin extra strips to an area already calculated? Suppose we have the area from 0 to x calculated approximately as A(x) and look at the area calculation around x=2.4. The generative idea here is to stretch the graph in the horizontal direction, without changing the vertical scale (figure 9).

Figure 9

This pulls the graph out horizontally and clearly shows that

$$A(x+h) \approx A(x)+f(x)*h,$$

so

$$f(x) \approx \frac{A(x+h)-A(x)}{h} \quad .$$

What can be motivated by this idea is that the more one stretches this particular graph horizontally, the more it pulls out flat. The essential property is that, to get the y-value onscreen within a pixel, say to get f(x+h) within a tolerance ±ε of f(x), requires the choice of a suitably small value of h, say h less than some specified δ. Thus the idea of being able to stretch a graph out flat corresponds to the symbolic idea:

Given $\varepsilon>0$, one may specify $\delta>0$ so that when $|h|<\delta$ so $|f(x)-f(x+h)|<\varepsilon$,

which is precisely the definition of continuity. In other words *continuity arises as a natural ingredient for the fundamental theorem of the calculus.*

This can be tested out on other continuous functions, for example, although the blancmange function looks very steep at each integer value of x, stretched horizontally it still pulls out flat! (Figure 10.) Even √x, which has a vertical tangent at the origin, pulls out flat for given y-scale by choosing the x-range suitably small...

Figure 10

4. Insights into theorems about integration

Looking more specifically at the fundamental theorem we have a program which plots the area calculation as a sequence of dots and simultaneously draws the straight line through the last two dots at each stage. The gradient of this straight line is an approximation to the gradient of the area function. But as the area of the last strip (say from x to x+h) is h*f(x+h/2) using the mid-ordinate approximation, the gradient is this value divided by h, which is f(x+h/2). For small values of h this is close to f(x), the ordinate on the original curve, again giving a visual representation of the fundamental theorem, though one which contains a lot of information and is therefore more difficult to comprehend in real time. Taking the strip-width to be very small and using the power of the RISC chip to draw the moving picture at maximum speed can give interesting insights.

Consider the integral of the discontinuous function $[x]$ from $x=-1$ to 3 with a small step, say 0.02 (figure 11). Over each horizontal strip of the graph the area function increases linearly, but as each new step is higher, the area function increases faster. At the discontinuities the different left and right gradients match the growth rates on the two successive horizontal parts of the original graph...

Figure 11

Rather more interesting is $x-[x]$, using the fast speed of drawing to actually *see* the sudden change in gradient of the area function as it occurs in real time. Figure 12, which becomes truly meaningful only when seen changing dynamically, has the graph $f(x)=x-[x]$ drawn and is building up the picture of the area function, momentarily drawing the gradient of the area function as it does so. The successive pictures show the sudden change in gradient of the area function as the calculation passes over a discontinuity of $f(x)$ at $x=3$.

Figure 12

A discontinuity in the original function does not always lead to a sudden change in gradient. Try drawing the graph of $o(x)=|\mathrm{sgn}(x)|$, the absolute value of the signum function, which gives $o(x)=1$ for $x\neq 0$ and $o(x)=0$. The graph-drawing algorithm is unlikely to pick up the odd point where the function is zero so it will just draw a horizontal line. However, if the software will calculate the value of individual points on the graph, if $x=0$ is given specifically, then one obtains $o(0)=0$.

Clearly calculating the numerical area for this graph over an interval including the origin will like as not fail to pick up the value of the function at the origin unless we select the intervals very carefully, and even then the effect of this will be diminished by taking very thin strips. The area function from $x=a$ in this case is $A(x)=x-a$, with derivative $A'(x)=1$. Thus the area function can be everywhere differentiable, but differ from the original function if that has isolated discontinuities.

What would the integral of $1/[1/x]$ look like? It is an interesting challenge to try to sketch it... Where is it continuous? Where is it differentiable? (Figure 13.)

Figure 13

The integral of a non-differentiable function is surely a very odd creature. As a challenge, what does the area function look like for the 3rd van der waerden function? The resulting graph is of a function which is differentiable everywhere *precisely once*. (Figure 14.)

Figure 14

Again, the drawing of the area function in real time produces interesting insights, particularly when one looks at the changing gradient. For example, near the point x=1 the graph is near zero, so the change in area is small too, leading to the area graph being nearly flat. In fact the area function for this non-differentiable function is a very boring looking function indeed, just lumpily increasing in a relatively smooth manner, with gradient varying between 0 and 3/4 (the lower and upper bounds of the Van der Waerden Function).

5. Differential Equations

A first order differential equation

dy/dx=f(x,y)

can be visualized by drawing an array of short line segments through points (x,y) with gradient f(x,y). A solutions then follows the direction of these line segments.

The value of drawing a direction field for differential equations to visualize solutions is well known (Neill & Shuard 1982, Hubbard & West 1985, Tall & West 1986). By allowing a wider repertoire of functions, graphical programs can begin to address broader problems. For instance, under what circumstances does the differential equation dy/dx=f(x,y) have solutions? Limited experience might suggest f(x,y) would need to be some kind of “nice function”, but looking at possible solutions of a differential equation such as

dy/dx=bl(x)

(the blancmange function) suggests that continuity of the function is more essential than differentiability. Drawing a direction field through an array of points in the diagram, with each line segment having gradient bl(x) produces a rather uneventful looking picture (figure 15). As bl(x) is everywhere between 0 and 1, the gradient of the line segments are everywhere between 0 and 1 and a solution curve which follows the given directions gently ambles along increasing in a steady but lumpy way we saw earlier. Since a solution y=I(x) of the differential equation satisfies I'(x)=bl(x), we once more find a function I(x) which is everywhere differentiable once and nowhere twice.

Figure 15

Of course this solution is the integral

$$I(x) = \int_a^x bl(t)\, dt \ ,$$

which is the area under the blancmange curve from a fixed point a to a variable point x. We do not know a simple way of calculating this function in general, although the reader might like to check that the area from 0 to 1 is exactly $\frac{1}{2}$. Thus we know that a solution exists and be able to calculate it numerically, but are not be able to give it in terms of a formal algorithm or formula.

Differential equations such as

$$dy/dx=x^2+y^2$$

may have solutions, but these solutions may not be given by combinations of known functions (in “closed form”). Hubbard and West (1985), teaching differential equations to students have noted that students have difficulties distinguishing between the case that “a differential equation has no solution”, and that “a differential equation has no solution in closed form”. The problem lies in the similarity of the verbal descriptions and the lack of suitable examples. Graphical models enable the student to *see* that a solution exists, even though it has no simple formula.

Looking at higher order differential equations, such as

$$d^2y/dx^2=bl(x)$$

can lead to even more powerful ideas. Solutions to this equation can be easily drawn and have the property that they are *everywhere differentiable twice, but nowhere differentiable three times*. Once more the graphical solution, far from being exotic, is really rather prosaic (figure 16).

Figure 16

As we have indicated elsewhere (Tall 1986), the visual approach to first order differential equations through the direction diagram gives a far more powerful generative idea than the hotchpotch of symbolic methods, which are only of value in specific cases. It gives insights into the conditions under what solutions might exist and the qualitative shape of such solutions, simply by looking at the picture. See, for example, how figure 17 intimates the possible behaviour of solutions of the differential equation dy/dx=x-y+1. As x increases, no matter what the initial starting values, the solutions home in on the line y=x.

Figure 17

Solving differential equations is best done by a combination of symbolic methods interpreted using visual insight and dynamic computer graphics have a fundamental role to play.

6. Conclusion

Nature has few straight lines - one might say that the only straight lines in nature are made by man, or by motion. Yet we model the real world of fluctuation and change by the calculus, which itself is based on the idea of local straightness. Moreover, the real world models are by no means always amenable to the symbolic methods of the calculus. In the real world, computer models of the economy, aircraft manufacture, weather prediction, space travel and a myriad other applications use numerical approximations honed by carefully guided theory.

Does this not intimate to us that the traditional approach to the calculus, heavily biased towards symbolic methods, might not benefit from a greater use of numerical methods, backed by graphic visualization to give the general overview.

The once universally taught method of extraction of square roots is dead, long division is dying, we no longer double declutch when changing gear. Given the increasing power of visualization with the computer, the emphasis on symbolic manipulation for the majority of users stands in peril of being replaced. For the many who *use* mathematics, interactive computer graphics can begin to give the insights that has been so sorely lacking in purely formal methods. For those who pass on to study formal mathematical analysis the power of graphic packages is now being raised to a level where it is able to provide the cognitive foundations on which the theorems can be built.

References

Hubbard J.H. & West B.H. 1985: ‘Computer graphics revolutionize the teaching of differential equations’, in *Supporting Papers for the ICMI Conference on "The Effects of Computers & Informatics on Mathematics & its Teaching",* Strasbourg.

Neill & Shuard 1982: *Teaching Calculus*, Blackie.

Takagi 1903: ‘A simple example of a continuous function without derivative’, *Proc. Phys.-Math. Japan*, 1,176-177.

Tall D.O. 1982: ‘The blancmange function, continuous everywhere, but differentiable nowhere’, *Mathematical Gazette*, 66, 11-22.

Tall D.O. 1986: “Lies, damned lies and differential equations”, *Mathematics Teaching*, 114, 54-57.

Tall D.O. & West B.H. 1986: ‘Graphic Insight into Calculus & Differential Equations’, *The Influence of Computers and Informatics on Mathematics and its Teaching*, (ed. Howson & Kahane), 107-119, Cambridge University Press.