A Graphical Approach to Integration and the Fundamental Theorem

David Tall

Mathematics Education Research Centre

University of Warwick

COVENTRY CV4 7AL

In 'Understanding the Calculus' [3] I suggested how the concepts of the calculus could be approached globally using moving computer graphics. The idea of area under a graph presents a fundamentally greater problem than that of the notion of gradient. Each numerical gradient is found in a single calculation as a quotient

$$\frac{f(x+h)-f(x)}{h}$$

but the calculation of the approximate area under a graph requires many intermediate calculations. Using algebraic methods the summation in all but the simplest examples becomes exceedingly difficult. A calculator initially allows easier numerical calculations but these can become tedious to carry out and obscure to interpret. Graduating to a computer affords insight in two ways: through powerful number-crunching and dynamic graphical display.

Algebraic methods

The area under a simple graph such as $f(x)=x^2$ may be approximated by dividing the interval into n equal width strips and adding together the "upper" and "lower" rectangular approximations in each strip. If one knows the appropriate formulae to simplify the sums, it is possible see what happens as n gets very large. The method for $y=x^k$ involves knowing the sum of the kth powers of 1, 2, 3, ..., n, which is plausible for sixth-form students when k=1 or 2, but becomes quite unmanageable for much larger values.

In 1635 the Italian mathematician Cavalieri demonstrated his computational facility by performing the calculations for all powers up to x^9. In those days Cavalieri was king, but even he would be hard put to cope with x^{100}!

Numerical methods with a calculator

The arrival of the calculator put power into the hands of ordinary mortals, enabling area approximations to be computed numerically. For instance, the approximations for the area under f(x)=x2 from 0 to 1 using 10 strips are:

Published in *Mathematics Teaching*, **113** 48–51 (1986).

lower sum: 0.285

upper sum: 0.385.

But this numerical information alone is not sufficient to determine the true value of the area and it would be very lengthy business to get more accurate approximations on an ordinary calculator.

A programmable calculator, suitably programmed, gives the area approximations using 100 strips as

lower sum: 0.32835

upper sum: 0.33835

and 1000 strips as

lower sum: 0.3328335...

upper sum: 0.3338335...

These still look a long way apart, but the average of lower and upper sums (equivalent to the trapezium rule):

10 strips ... 0.335

100 strips ... 0.33335

1000 strips ... 0.3333335,

suggests the true area is probably 1/3.

By patiently building up the areas over other intervals, say 0 to 1, 0 to 2, 0 to 3, and so on, it is possible to conjecture that the area from 0 to x is $x^3/3$, but the numerical work begins to get oppressive unless a computer is available.

To spread the load, Neill & Shuard[1] used a whole class of students with calculators to cooperate in producing a table of areas from 0 to 1 using the trapezium rule with 10 strips:

function f(x)	approximate area
x	0.50
x^2	0.33
x^3	0.25
x^4	0.20
x^5	0.17
x^6	0.15
etc	etc

By cunningly showing the area calculations to only two decimal places, the first four are visibly 1/2, 1/3, 1/4, 1/5,... and a check shows 0.17, 0.15 are approximately 1/6 and 1/7 respectively, leading to the conjecture that the area under $f(x)=x^n$ from 0 to 1 is $1/(n+1)$.

Numerical methods with a computer

A computer with a flexible computer language can allow much more powerful methods to be used. For example, Powell [2], shows how the average of a subtle mixture of the first ordinate, mid-ordinate and last ordinate rules can give the same result as Simpson's rule. This leads to far more accurate calculations and can be most helpful for inducing algebraic formulae from numerical results.

But there is far more to a computer than number-crunching. All the methods considered so far concentrate on the final result of an area calculation and neglect the information given by the intermediate sums. When the area is calculated, each partial sum from a to x gives the approximate area as a function of x. If we could only utilize this information in some sensible way, perhaps we could get to the area function more directly...

Graphical Insight with a Computer

Using the information from the partial sums is not as straightforward as it might be. However, if by guile or good luck one plots the successive cumulative area calculations for $f(x)=x^2$ over a suitable range, one gets an interesting picture (figure 1). The separate points represent the area under the graph from $x=0$ to the current x coordinate using the mid-ordinate rule. The area graph crosses the graph $f(x)=x^2$ at $x=3$, $y=9$. The area graph looks like a higher power of x, say kx^3 for some constant k. Substituting $x=3$, $y=9$ gives $k=1/3$ and suggests a possible area function $x^3/3$.

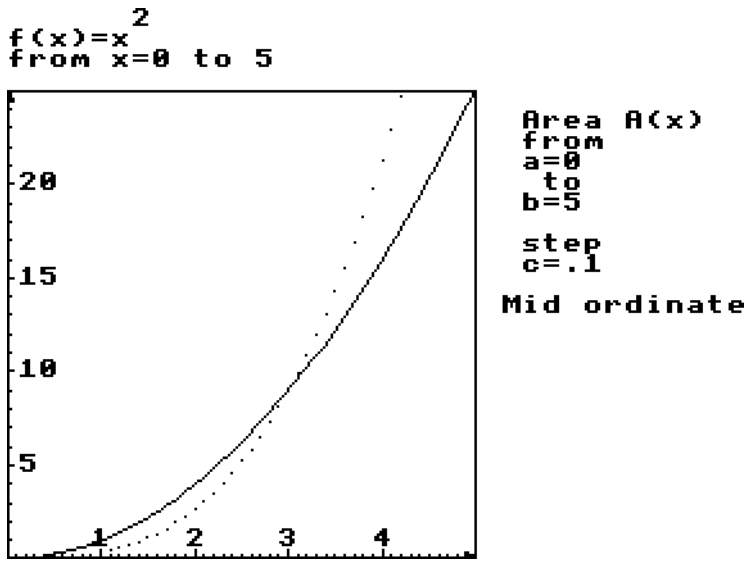

Figure 1 : The cumulative area under f(x)=x^2 from 0 to 5

This inspired guess can be given more credence if we are brave enough to try a more powerful approach.

Suppose we forget the obsession with calculating the area under $y=x^2$ from 0 to 1 and encourage students to look at more general examples. First they must know how to calculate an approximate area using, say, the first ordinate or mid-ordinate rule. This can be done in simple cases by hand, or by calculator, to get the idea. Then they can graduate to a piece of software that takes the drudgery out of the calculations so that they can concentrate on the ideas rather than the technicalities. The program AREA in Graphic Calculus II [5] draws any graph $y=f(x)$, then calculates the area between specified points a,b with any given step-width c, using either the first ordinate, mid-ordinate or last ordinate rule. (If the step-width c is not an exact divisor of the interval $b–a$, the last strip is reduced in width to make it fit.)

Notice that the program does not estimate upper and lower sums. These are easy enough if the graph is strictly increasing or decreasing but, if an interval encloses a maximum or minimum, a sophisticated estimation technique would be required to find the upper or lower value. The theory of "upper and lower sums" is built on an “existence” theorem that asserts these sums exist, without showing explicitly how they may be calculated. Employing the first or last ordinate in each interval is both more straightforward and more honest!

Using the software students may explore many possibilities in a short time. For example, can one estimate the area under the graph $y=\sin x$ from $a=0$ to $b=\pi$? By symmetry this is twice the area from $a=0$ to $b=\pi/2$ where the graph is increasing. Here first and last

ordinates give lower and upper sums sandwiching the true area in between. The mid-ordinate rule using a step c=1/4 gives the approximate area as 1.0024 (figure 2).

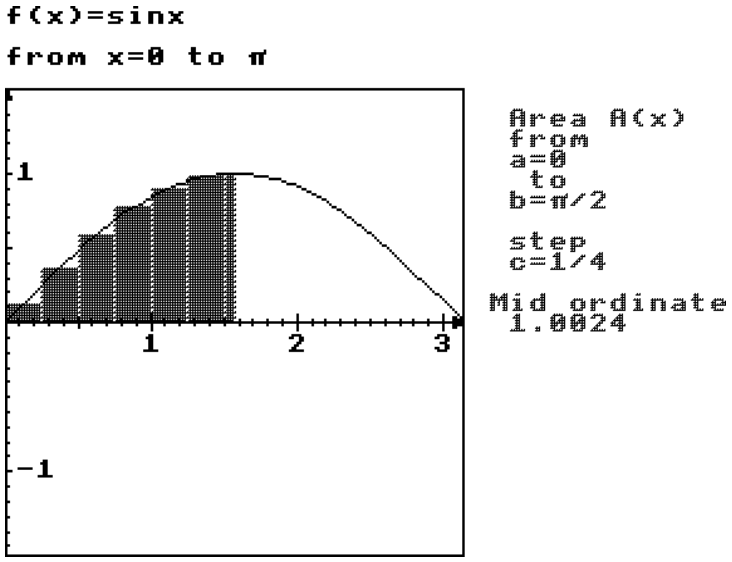

Figure 2 : The approximate area under sin$x$ from 0 to π

Alternatively, one may set the display to show the three calculations simultaneously; for x=1/50 the area approximations are:

first ordinate: 0.9900

last ordinate: 1.0010

mid-ordinate: 1.0000

Students are often intrigued to find that more accurate estimates for the area under a trigonometric graph seem to suggest the area is an exact whole number.

Subsequent investigations might be to see what answer the computer gives for the area from a=π to b=2π. Here the graph is below the axis and the calculation gives an estimate of the area from the x-axis down to the graph. The ordinates are all negative and the calculations all give a negative result, represented in the picture by rectangles in outline rather than filled in (figure 3). The reason for the sign is not hard to see. The method of calculating the area takes the step-length times the ordinate in each rectangle. If the step-length is positive and the ordinate negative, then the product will be negative.

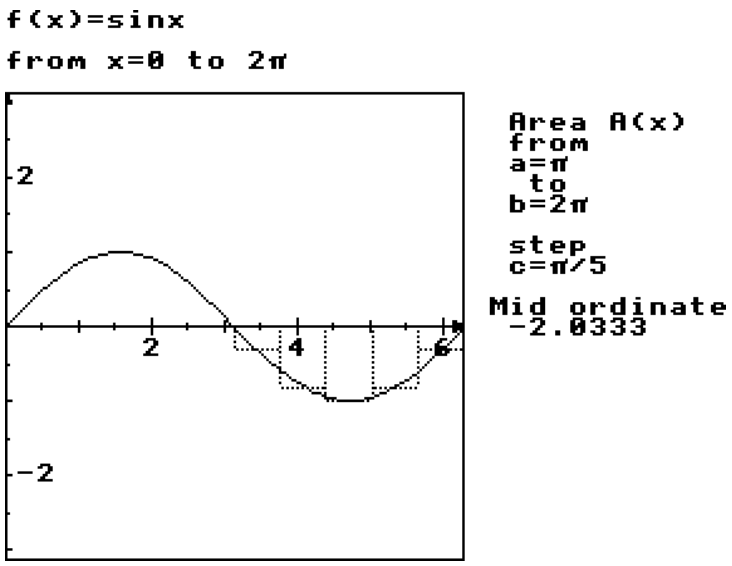

Figure 3 : Area from π to 2π

At this stage it is instructive to ask the students what might happen if one took *b* to the left of *a*, which would require a negative step. This is rarely done in traditional courses and teachers often suggest to me that it is too difficult for their students. However, the students I have worked with seem to find it no problem at all: a negative step and positive ordinate gives a negative product, a negative step and a negative ordinate gives a positive product. The ideas may be immediately tested using the program to perform the area calculation from $a=\pi$ to $b=0$ or from $a=0$ to $b=-\pi$, in each case with negative step, say, $c=-1/10$.

The computer simulation is rather better than a static picture (figure 4); using a negative step the picture of the area approximation builds up from right to left. One can see the negative step and sense the growth of the area as the picture develops.

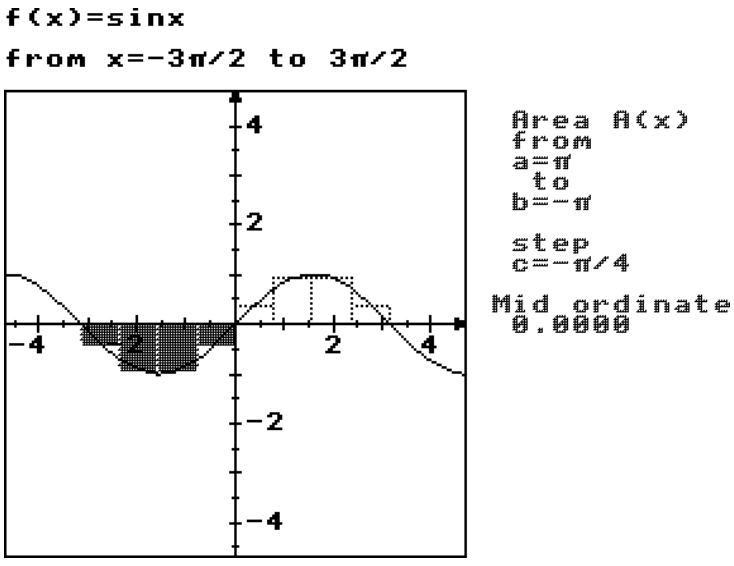

Figure 4 : using a negative step

By allowing students to investigate negative steps, all four possible combinations of signs are covered: positive or negative step with positive or negative ordinate. Thus they see that calculations from left to right produce a positive result above the axis and a negative result below, but going from right to left reverses the signs.

It is known that students doing calculus the traditional way have difficulty understanding the reasons for positive and negative areas; often they are simply told the rule in an instrumental way. With the computer graphic approach the concept is given a meaningful interpretation which relates to their other mathematical experiences. It is hard to think of many other contexts in mathematics where the multiplication of signed numbers is given such a powerful and meaningful interpretation.

Notice that this discussion of the sign of the area can occur early on in the encounter with the area concept. By exploring the software described, students gain a meaningful awareness of two fundamental ideas:

(1) when smaller strips are taken, the area approximation gets closer to the true area,

(2) the area calculation takes the sign of step and ordinate into account in a meaningful way.

This experience can now be applied to the problem of finding the area between a graph and the x-axis from a fixed point a to a variable point x, and considering it as a function of x.

For example, the cumulative area function for the graph $f(x)=x$ starting at $a=0$ may be plotted. Moving to the right of the origin requires a positive step and has a positive ordinate, so the result is positive. But moving to the left has a negative ordinate and a negative step, which again gives a positive result. Using the program to plot the area calculations starting from the origin, first moving to the right, then starting again at the origin and moving to the left gives figure 5. The graph now takes on a recognisable shape: probably a quadratic function, in the form kx^2 for some k. As the area graph crosses the original graph at $x=2$, $y=2$, it is easy to conjecture that the area function is $x^2/2$.

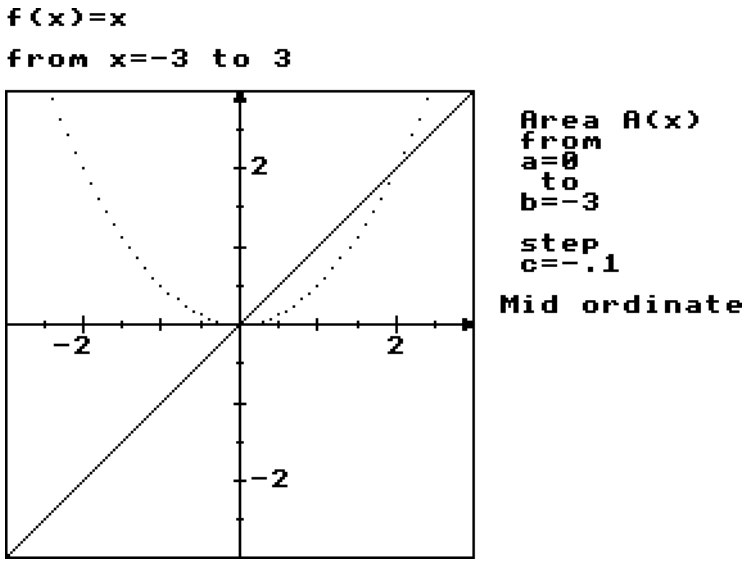

Figure 5 : The area under f(x)=x calculated from x=0

The program allows a formula to be typed in and compared with the original: it fits like a glove.

The same process applied to f(x)=x^2 also gives an increasing graph to the right of the origin, but moving to the left with a negative step and positive ordinate, gives a negative result. The area graph looks like a cubic curve (figure 6). The earlier argument that the area may be $x^3/3$ is then enhanced. It can be tested by superimposing the latter graph.

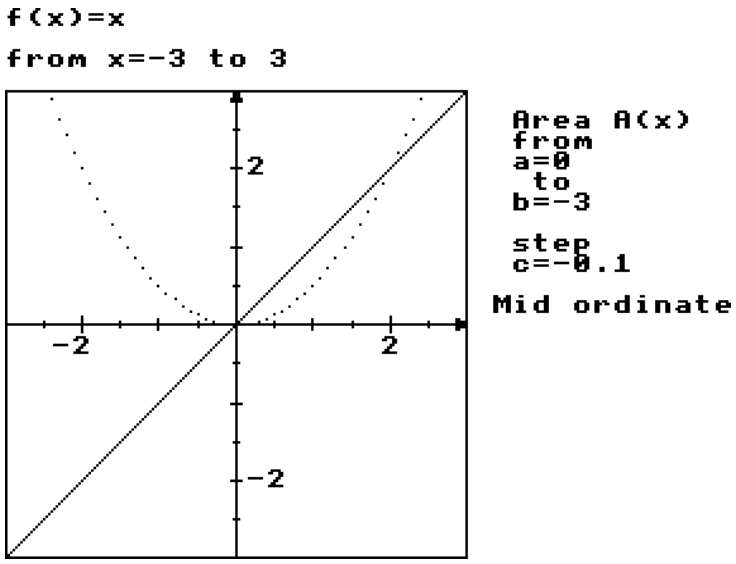

Figure 6 : The area under f(x)=x^3 calculated from x=0

The two area functions calculated from the origin:

function	area
x	$x^2/2$
x^2	$x^3/3$

suggest the general pattern: the area under $f(x)=x^n$ from 0 to x may be conjectured as $I(x)=x^{n+1}/(n+1)$. The formulae may be typed into the program in this form to see if it works for other values of n.

As the area from $x=a$ to $x=b$ is the difference between the areas $I(a)$, from $x=0$ to $x=a$, and $I(b)$, from $x=0$ to $x=b$, the area from $x=a$ to $x=b$ may be conjectured to be the difference $I(b)–I(a)$. The program displays $I(b)–I(a)$ during area calculations, so one may see if the numerical approximations get close to this value for various values of a,b,n. It does so for all positive n (even though x^n is only defined for positive x when n is not a whole number).

At a later stage one may test what happens with other values of n. For instance, if n is negative and not a whole number, the graph is only defined for $x>0$ and the formula works for any positive values of a,b. But when n is a negative whole number the situation gets far more interesting. The formula works for $n=–2,–3,...,$ but there is a "hiccup" crossing the origin which we shall see also shows up in the case $n=–1$.

For $n=–1$ the area from $x=a$ to $x=b$ can be calculated numerically (provided one does not hit the point $x=0$). But the formula $I(x)=x^{n+1}/(n+1)$ is nowhere defined because it involves dividing by $n+1$ which is now zero.

There is room here for interesting investigative work into the property of the area function for $f(x)=1/x$. For instance, if $A(x)$ is the area under the graph from $a=1$ to $b=x$, one may explore the relationship between $A(x_1)$, $A(x_2)$ and $A(x_1+x_2)$ to discover the logarithmic property, or estimate the constant k such that $A(k)=1$, leading to the constant e.

The area function $I(x)=\ln(x)$ gives the area from a to b as $I(b)$-$I(a)$ provided that a,b are both positive. But $\ln(x)$ is not defined for $x<0$. The function $I(x)=\ln|x|$, typed as ln(abs(x)), is defined on both sides of the origin, but the area calculation $I(b)$-$I(a)$ only works if a and b have the same sign.

To highlight the problem, the program has an option to make each step a random size up to the nominated value of the step-width. When a,b are both on the same side of the origin, the numerical calculation of area using a random step does not vary much and for small steps it approximates to $I(b)$-$I(a)$ as expected.

However, if a,b are taken on opposite sides of the origin, several runs using a random step produce totally different results, depending on what large values of $1/x$ are picked up when x is near zero. For example, three runs of the program from $a=–1$ to $b=1$ with maximum random step $c=1/10$ produced the results: 4.3590, 0.7267, –11.6602.

Experiences such as these help the student to understand why one should not attempt to calculate areas across a place where the function gets arbitrarily large.

The Fundamental Theorem

The experiences of the preceding sections naturally lead to the idea that the area function I(x) differentiates to give the original function f(x). Thus one may conjecture the "theorem" that if I′(x)=f(x) for all values of x from a to b, then the area between the graph y=f(x) and the x-axis is I(b)-I(a). But is this always so?

My suggestion here is to stretch the graph horizontally whilst keeping the vertical scale the same. To do this in a picture one takes the y-range to be a normal size, say y=-2 to 2 and chooses the x-range very small, say from x=0.999 to 1.001. If these ranges are used to give a square picture, the result is an x-scale greatly stretched compared with the y-scale. The graph comes out much flattened. For instance the graph of y=sin$x$ over these ranges looks like a horizontal straight line! (Figure 7.)

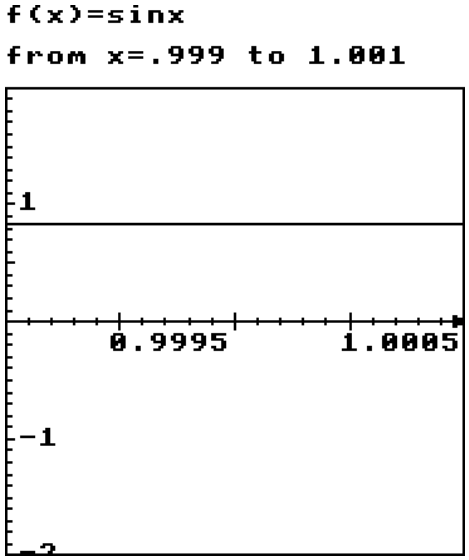

Figure 7 : Horizontal stretching of the sine graph

The area under a horizontal graph is the width times the height of the graph. Thus the area under this stretched graph from x to x+h (where h is small) is approximately h times f(x). But if I(x) is the area from a fixed point a to a variable point x, the area from x to x+h is then approximately

$$\mathrm{I}(x+h)-\mathrm{I}(x) \approx h\,\mathrm{f}(x)$$

which gives

$$\frac{\mathrm{I}(x+h)-\mathrm{I}(x)}{h} \approx \mathrm{f}(x). \qquad (*)$$

As h gets very small the left-hand side gets close to $I'(x)$, suggesting that $I'(x)=f(x)$.

Of course this is a flawed argument which causes students some problems, particularly in deducing an equality from an inequality. A useful exercise which may help to set the argument in perspective is to draw the graph over a number of ranges, starting with the same *x*- and *y*-ranges, then successively reducing the *x*-range. As each graph is drawn, with a smaller *x*-range stretched to fit into a square, the graph gets flatter and flatter. It is possible to suggest the idea that, as *h* gets smaller in (*) the graph gets closer to a straight line and the approximation gets better. Thus it is possible to believe that, as $h\to 0$, so the inequality (*) tends to the equality $I'(x)=f(x)$.

The role of continuity

Continuity is very rarely discussed to any great extent in a first course in calculus. But it arises naturally in the fundamental theorem. To get the theorem to work requires the idea that the graph can be stretched horizontally so that a small part looks flat. What does this mean? The graph is not genuinely flat. What happens is that the variation in height is so small that it occurs within a pixel height on the television screen. To simplify matters, suppose that the point x_0 is in the middle of the *x*-scale and the point $(x_0,f(x_0))$ on the graph is in the middle of a pixel representing an actual height $2e$. We want to know that we can find an x-range from x_0-d to x_0+d so that for *x* in this range the value of $f(x)$ lies in the pixel. Thus we need to know that, given *e*, we can find *d* such that for *x* in the range:

$$x_0-d < x < x_0+d$$

we have $f(x)$ in the pixel height:

$$f(x_0)-e < f(x) < f(x_0)+e.$$

Translating the letters *e*,*d* into Greek to make it look more mathematical gives the well-known definition of continuity:

The function $f(x)$ is continuous at x_0 if:

given $\varepsilon>0$ there exists $\delta>0$ such that

$x_0-\delta< x < x_0+\delta$ implies $f(x_0)-\varepsilon < f(x) < f(x_0)+\varepsilon$,

or, making it even more abstruse with modulus signs:

given $\varepsilon>0$ there exists $\delta>0$ such that

$|x-x_0|<\delta$ implies $|f(x)-f(x_0)1<\varepsilon$.

Thus the fundamental theorem is the natural place in the calculus for the notion of continuity to arise. If f(x) is continuous, its area function I(x) is differentiable and I′(x)=f(x). In this way intuitive ideas may be laid down to form a basis for the later formal theory.

Integrating continuous functions

The fundamental theorem does not require the function f(x) to have a derivative, it only requires continuity. Thus one can consider a function such as f(x)=|x| (typed as f(x)=abs(x)) which is continuous but not differentiable at the origin. Its area function starting from a=0 is

I(x)=x^2/2 (for x>0)

and

I(x)=-x^2/2 (for x<0).

Being devious, one can type this in as

I(x)=x(abs(x))/2.

Thus one gets a function I(x) which is differentiable everywhere to give I′(x)=f(x), but f(x) is not differentiable at the origin (figure 8).

Integrating discontinuous functions

The program AREA can be used to investigate slightly more off-beat functions. One of my favourite examples is

f(x)=x–int(x)

which is the BASIC function x-INT(x), subtracting the integer part of x from x, leaving the decimal part.

Thus

f(2.345) = 2.345-INT(2.345) = 2.345-2 = 0.345.

The graph is not continuous at any integer. (Try distending it horizontally to see if it will stretch to look flat there!) But an attempt to draw the area "under" the graph produces an interesting result.

Using, say, the first ordinate approximation, an interval is divided into strips and the area approximated by a sequence of rectangles, each of which has its height as the first ordinate in the strip (figure 9). Visibly the area can be calculated in this way, and for very small strips it approximates to a triangular area in each unit interval.

It is an interesting problem-solving exercise to find the formula for the area function (see [4]). Even without the formula, the approximate area function may be represented graphically, starting at a=0 and moving in steps, first to the right with step c=0.1, then to the left with c=–0.1. The resulting picture (figure 10) is not very accurate but, with a little graphic licence, one may see that the area function graph is smooth except at each integer point, where it clearly has different gradients to the left and the right. One may imagine that magnified up near these points the graph of the area function will never look straight, it will look like a "corner". Thus this area function has the property that it is not differentiable at each point where the original function is discontinuous. This gives a foretaste of a powerful theorem in analysis not usually met until the first or second year of a university mathematics course:

- If a function y=f(x) is (Riemann) integrable in the interval [a,b] to give an area function I(x), then I(x) is continuous.
- At every point x_0 in [a,b] where f(x) is continous, I(x) is differentiable and I'(x_0)=f(x_0).

Integrating non-differentiable functions

The program AREA only allows functions to be typed in which are expressed as standard formulae. But the fundamental theorem applies to any continuous function. For example, it applies to the blancmange function b(x) [4], which is everywhere continuous and nowhere differentiable. Thus the area function $b_1(x)$ for the blancmange function has the property that it is differentiable everywhere and $b_1'(x)=b(x)$. The function $b_1(x)$ is a function everywhere differentiable once, but nowhere differentiable twice. Similarly the area function $b_2(x)$ for $b_1(x)$ satisfies $b_2'(x)=b_1(x)$, so $b_2(x)$ is everywhere differentiable twice but nowhere three times. Repeating the process gives an example of a function differentiable everywhere n times, but nowhere n+1 times. The mind boggles!

A graphical approach to calculus is thus not just a "simple way in" for beginning students, it also provides insight into powerful theorems that occur much later in formal mathematical analysis.

References

1. H Neill & H Shuard : *Teaching Calculus* (Blackie) 1982
2. M Powell: A numerical approach to integration, *Mathematics in School*, 14, 5, 44-46 1985
3. D O Tall: Understanding the calculus, *Mathematics Teaching* 110, 48-53 1985
4. D O Tall: The gradient of a graph, *Mathematics Teaching* 111, 48-52 1985
5. D O Tall: *Graphic Calculus II* (Integration) for the BBC Computer, Glentop Publishers, 1986