Integration, a genetic introduction

Robert P. Burn

The genetic method of O.Toeplitz is a structuring of conventional instructional material using historical insights and historical problems, not for a history course, but for the better understanding of modern mathematics by students today. The paper which follows describes an introduction to integration suitable for a first course in analysis, in the style of Toeplitz. After two examples of the Greek method of exhaustion (intended for lecturers rather than students), historical developments from the 17th and 19th centuries pinpoint the conceptual steps which a student must take in a conventional first course in analysis up to the Riemann integral.

Because of previous misunderstandings exposed by the rigour of Weierstrassian analysis, university lecturers in analysis today generally try to keep to a deductive sequence for fear of creating the old misconceptions. However, a logical development may not coincide with a psychological or intuitive development. Logic prefers generality; intuition needs special cases.

Teaching experience, or psychological research, may well provide the appropriate examples for an intuitive development; but there are cases where the difficulties of a deductive sequence are well-known and seemingly perennial. A typical modern student, for example, comes to analysis knowing that integration is "anti-differentiation", and having used integration for practical purposes. From the student's perspective, the definition of the Riemann integral is not the obvious next step. In such cases, an examination of how mathematicians in the past have come to new understandings may well suggest a clearer developmental sequence and effective exercises for today's students.

In relation to the development of the theory of integration there are both progressive developments (from geometry to algebra, from curves to functions) and there are fluctuating methods (equal *or* unequal partitions, upper and lower sums *or* limits). If students experience these changes step by step, the development seems reasonable. If the changes come all at once, they may be indigestible. Using history to locate steps in development is the 'genetic method'.

Robert P. Burn *is professor of Mathematics Education at Høgskolen i Agder, Kristiansand, Norway.*

The 'genetic' method was so-termed by Otto Toeplitz in a lecture in 1926 in which he said

> *Regarding all these basic topics in infinitesimal calculus which we teach today as canonical requisites, e.g., the mean-value theorem, Taylor series, the concept of convergence, the definite integral, and the differential quotient itself, the question is never raised 'Why so?' or 'How does one arrive at them?' Yet all these matters must at one time have been goals of an urgent quest, answers to burning questions, at the time, namely, when they were created. If we were to go back to the origins of these ideas, they would lose that dead appearance of cut and dried facts and instead take on fresh and vibrant life again.*
> (*Jahresbericht der deutschen mathematischen Vereinigung, XXXVI (1927), 88-100,* reprinted in the preface to the German edition of Toeplitz (1963).)

The question which Toeplitz was addressing was the question of how to remain rigorous in one's mathematical exposition and teaching structure while at the same time unpacking a deductive presentation far enough to let a learner meet the ideas in a developmental sequence and not just a logical sequence. While the genetic method depends on careful historical scholarship it is not itself the study of history. For it is selective in its choice of history, and it uses modern symbolism and terminology (which of course have their own genesis) without restraint. Characteristically the history chosen is of special cases which did in the past and can for students today suggest fruitful generalisation. Those familiar with *The Calculus, a genetic approach*, by Toeplitz (1963) will recognise, in what follows, a reworking and extension of parts of chapters 1, 2 and 3 of that book. Many details have been modified, but the central idea belongs to Toeplitz. The outline here will repeatedly show mathematics developing from a special case to a generalisation. The development is presented for the most part in historical sequence. But since the prime intention is to provide a learning sequence, not to teach history, anachronisms, historical inversions and 'rational reconstructions' are to be found in all the sections except D12. An appropriate attention to historical development in any part of mathematics will give clues as to how knowledge has been and can be formed. Some of the history here has been retold many times with varying degrees of precision. To try to make the history that is used as accurate as possible I have cited primary sources wherever possible.

The initial motivation for most of the early development of integration was the measurement of areas bounded by curves. From Fermat onwards, the methods were algebraic. With the indefinite integrals of Newton and Leibniz the content of integration itself took on an algebraic character, and with the definitions of Cauchy and Riemann, the geometric nature of integration was reduced still further. As our story begins, area is a primitive notion, and areas may be compared only with other areas, with all measurements being essentially positive. With the Fundamental theorem the signing of areas becomes necessary, and with Riemann the definition of the integral is a formal one. But Riemann's definition results from stretching and consolidating the notion of area.

Alongside the shift from geometry to algebra was the increasing generality of the functions under discussion. With Fermat, powers of x (other than -1, but including radicals) and hence polynomials could be integrated. The functions which Newton integrated could all be expressed as power series. But in the nineteenth century, Fourier series provided a wider class still, and this was the context which provoked Riemann's widening of the notion of integrability.

A learning sequence could consist of sections B, C and D, restructured as a sequence of problems. Section A is background material for the lecturer; it is not for every student. It must be said that the 'genetic method' does not presuppose any particular teaching style or form of classroom organisation. It can be the basis of conventional lectures, or better, a course structured in the form of a sequence of problems to be solved. The point of the method is that students should work with a developmental sequence of ideas, not just follow the logic.

A. Ancient Greek mathematics - two special cases, the circle and the parabola. Approximation by triangles.

1. The areas of two circles

Euclid Book XII.2 (*c.* 300 BC). Translated in Fauvel and Gray (1987), p.136. The areas of two squares are in the ratio of the squares of their sides. This comparison extends to rectangles, parallelograms and triangles and hence to polygons. Does the proportion extend to circles? That is, is the ratio of the areas of two circles with radii r and R $r^2 : R^2$? The result is so familiar to us that it may seem hardly worth proving, but to doubt it and to address the possibility of a proof is to engage in an activity which is characteristic of any real analysis course. Euclid approached this problem by studying the areas of regular polygons inscribed in and circumscribed about a circle (a foretaste of the lower sums and upper sums in the later development of integration).

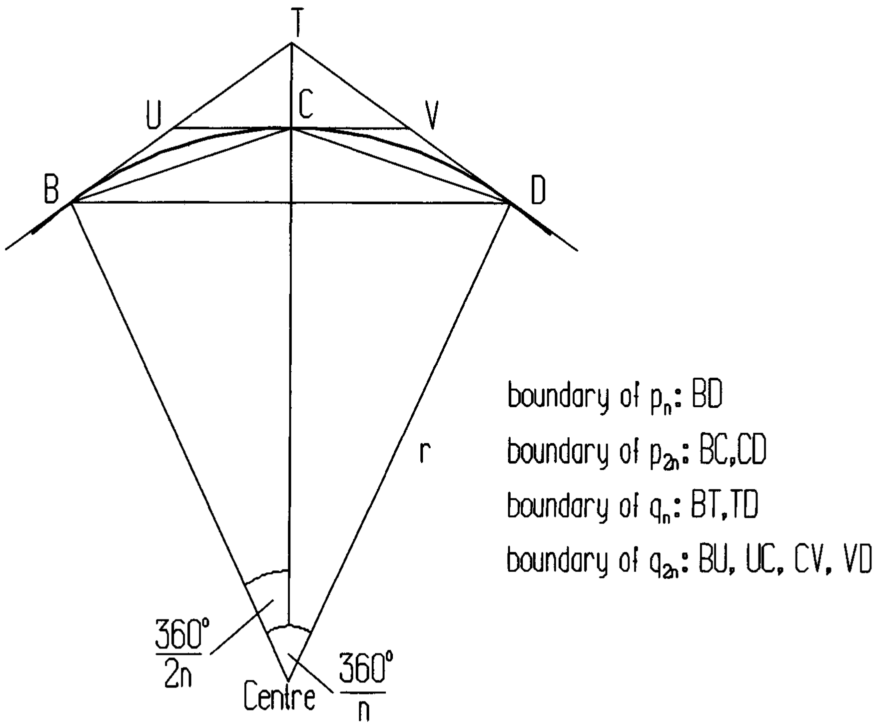


Figure 1

Consider a circle with radius r. Let the area of a regular polygon with n sides *inscribed* in the circle be p_n, and let the area of a regular polygon with n sides *circumscribed* about the circle be q_n. The geometric affect of doubling the number of sides of these polygons is to reduce the area between the polygons and the circle by more than ½, that is $a - p_{2n} < ½(a - p_n)$ and $q_{2n} - a < ½(q_n - a)$, where a is the area of the circle. This ensures that if $n = 1, 2, 4, 8, 16, ...$, the area $q_n - p_n$ eventually becomes as small as one might wish.

Now consider another circle with radius R, area A, and the areas of regular n-gons inscribed and circumscribed about the circle being P_n and Q_n.

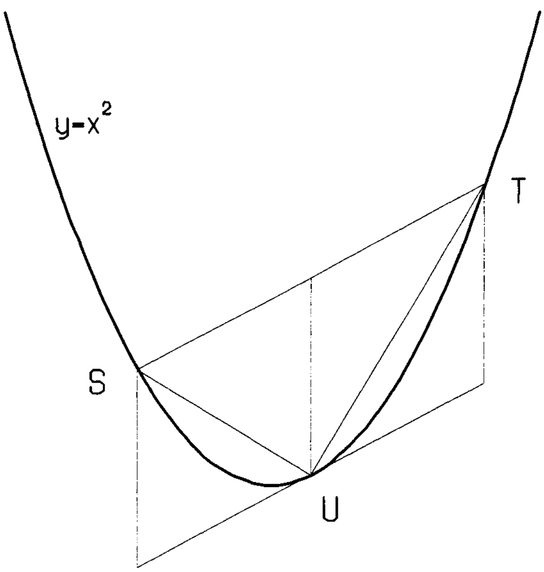
Then $\frac{P_n}{p_n} = \frac{Q_n}{q_n} = \frac{R^2}{r^2}$ and $p_n < a < q_n$ and $P_n < A < Q_n$.

So $R^2 p_n < R^2 a < R^2 q_n$ and $r^2 P_n < r^2 A < r^2 Q_n$.

Now $R^2p_n = r^2P_n$, and $R^2q_n = r^2Q_n$, and both $q_n - p_n$ and $Q_n - P_n$ become arbitrarily small as $n = 2^k$ increases, so the two numbers R^2a and r^2A are sandwiched between the same diminishing bounds. Thus either inequality ($R^2a > r^2A$ and $R^2a < r^2A$) is denied and we have $R^2a = r^2A$ or $\frac{A}{a} = \frac{R^2}{r^2}$ as required.

2. A segment of a parabola

Archimedes (*c.* 250 BC). Tranlated in Fauvel and Gray (1987), p.153. Archimedes was able to show that the area of a segment of a parabola was 4/3 times the area of the largest triangle which could be drawn in the segment. This was called the 'quadrature' of the parabola because it determined a square with area equal to the required parabolic segment. We can understand Archimedes' limiting argument, by dressing it in modern notation.



$$S = (s, s^2)$$
$$T = (t, t^2)$$
$$U = \left(\frac{s+t}{2}, \left(\frac{s+t}{2}\right)^2\right)$$
$$\Delta STU \text{ has area } \frac{1}{8}(t-s)^3$$

Figure 2

On the parabola $y = x^2$, let S be the point (s, s^2) and T be the point (t, t^2). If U is the point $(\frac{1}{2}(s + t), (\frac{1}{2}(s + t))^2)$, then the area of the triangle STU is:

$$\frac{1}{2}\begin{vmatrix} s & s^2 & 1 \\ t & t^2 & 1 \\ \frac{s+t}{2} & \left(\frac{s+t}{2}\right)^2 & 1 \end{vmatrix} = \left(\frac{t-s}{2}\right)^3$$

(The tangent at U is in fact parallel to the chord ST.) Because the triangle STU is exactly half the area of the parallelogram illustrated, triangle STU takes up more than half of the segment standing on ST. What it does not take up are the segments standing on SU and UT.

Since the area of triangle STU is proportional to the cube of the difference of the parameters at S and T, the area of each of the triangles constructed similarly in the two segments standing on SU and UT respectively is 1/8 of the area of STU. If we denote the area of triangle STU by A, then the two further triangles have a combined area of $A/4$. The portion of the area of the segment standing on ST not covered by the area $A + A/4$ now appears as four smaller segments. Triangles fitted in these four segments have a combined area of $A/16$.

Continuing in this way, triangles within the parabolic segment ST may give a combined area of $A + A/4 + A/16 + \ldots + A/4^n$. This can be made as close to $4A/3$ as we wish, so the area of the parabolic segment is not less than $4A/3$.Moreover each step, for geometric reasons, measures more than half of the area remaining, so if P is the area of the parabolic segment,

$P < A + A = 4A/3 + 2A/3,$
$P < A + A/4 + A/4 = 4A/3 + A/6,$
$P < A + A/4 + A/16 + A/16 = 4A/3 + A/24,$
...
$P < A + A/4 + A/16 + \ldots + A/4^n + A/4^n = 4A/3 + 2A/3{\cdot}4^n.$

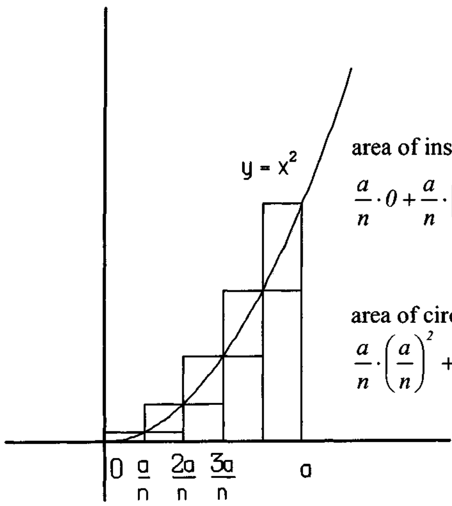
Each of these sums is greater than $4A/3$. But if the parabolic segment had area greater than $4A/3$, eventually one of these sums would be less than that supposed area.

Equality is the only remaining option so the parabolic segment has area $4A/3$.

B. Parallel rectangular strips of equal width

3. Fermat and Roberval find areas between the curves $y = x^2$ and $y = x^3$ and the x-axis.(1636)

Translated in Mahoney (1994), pp. 220-1.
Having invented coordinate geometry about the same time as Descartes, Fermat tried a new method for solving an old problem to which he already knew the answer. This method was to dominate thinking about these problems for more than 200 years. He sought the area bounded by the x-axis, the line $x = a$ and the curve $y = x^2$. Computing from Archimedes' results (A2, above) would have enabled Fermat to determine this area to be $a^3/3$ (= area of the triangle $(0, 0)(a, 0)(a, a^2)$ less the area of the parabolic segment standing on $(0, 0)(a, a^2) = ½a{\cdot}a^2 - (4/3)(a/2)^3$). We interpret the methods which Fermat and Roberval used, from their correspondence, in which the following seems to be implicit.



area of inscribed rectangles:

$$\frac{a}{n}\cdot 0+\frac{a}{n}\cdot\left(\frac{a}{n}\right)^2+\frac{a}{n}\cdot\left(\frac{2a}{n}\right)^2+\cdots+\frac{a}{n}\cdot\left(\frac{(n-1)a}{n}\right)^2$$

area of circumscribed rectangles:

$$\frac{a}{n}\cdot\left(\frac{a}{n}\right)^2+\frac{a}{n}\cdot\left(\frac{2a}{n}\right)^2+\cdots+\frac{a}{n}\cdot\left(\frac{na}{n}\right)^2$$

Figure 3

The area has been divided into n rectangular strips, parallel to the y-axis, each of width a/n. With the longest strips which could be inside the area, these together had area

$\frac{a^3}{n^3}(1^2 + 2^2 + 3^2 + ...+ (n - 1)^2)$. Calculating with the shortest strips which just covered the area, these together had area

$$\frac{a^3}{n^3}(1^2 + 2^2 + 3^2 + \ldots + n^2).$$

Using the formula $1^2 + 2^2 + 3^2 + \ldots + n^2 = n(n + 1)(2n + 1)/6$, simplifying, and denoting the area to be found by A,

$$\frac{a^3}{6}(1-\frac{1}{n})(2-\frac{1}{n}) < A < \frac{a^3}{6}(1+\frac{1}{n})(2+\frac{1}{n}).$$

Now these inequalities hold for all values of n, and this leaves $a^3/3$ as the only possible value for A. This exclusion of impossible answers by inequalities closely echoes the Greek method of exhaustion. Neither Fermat nor Roberval gave details of how they considered the limiting process, but in C6 Fermat refers to Archimedes' method of inscribed and circumscribed polygons.

The application of this method to show $\int_0^a x^3 dx = \frac{a^4}{4}$ using $1^3 + 2^3 + 3^3 + \ldots + n^3 = (\frac{1}{2}n(n + 1))^2$ follows quite naturally. Both Fermat and Roberval, in their correspondence, recognised that their method could be extended to give $\int_0^a x^k dx = \frac{a^{k+1}}{k+1}$ for any positive integer k, provided they had sufficient information about the sum $1^k + 2^k + \ldots + n^k$.
Such a result leads also to the value of the integral $\int_0^a \sqrt[k]{x}dx$, since the graph of $y = \sqrt[k]{x}$ is the mirror image of the graph of $y = x^k$ about the line $y = x$.

4.The logarithmic property of areas under $y = 1/x$.

Gregory of St.Vincent (1647) interpreted by Antonio de Sarasa (1649) and translated in Katz (1993), p. 449-450. The use of rectangles of equal width here is a suggestion of Toeplitz (1963, p. 55-57). Gregory of St.Vincent used indivisibles.

Neither the algebra of Fermat used in B3 above nor that which he used in C6 below can be extended to $k = -1$.

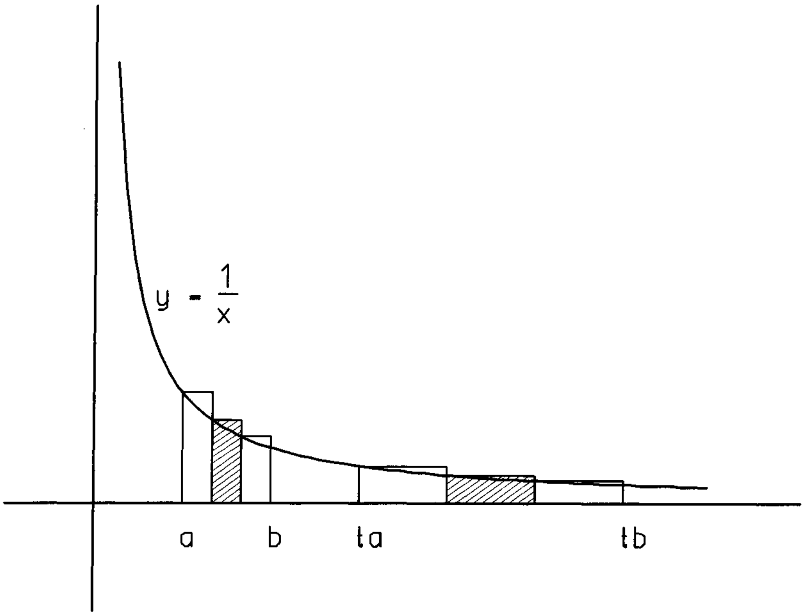


Figure 4

Without knowing the value of $\int_a^b \frac{dx}{x}$, it is possible to compare $\int_a^b \frac{dx}{x}$ with $\int_{ta}^{tb} \frac{dx}{x}$, by dividing both the intervals [a, b] and [ta, tb] on the x-axis into the same number, n, of equal parts at a,

$a+\frac{b-a}{n}, a+2\frac{b-a}{n}, \ldots, a+p\frac{b-a}{n}, \ldots, b$, and

ta, $ta+\frac{tb-ta}{n}, ta+2\frac{tb-ta}{n}, \ldots, ta+p\frac{tb-ta}{n}, \ldots, tb$, respectively.

If rectangular strips are constructed on these intervals just covering the curve, the corresponding rectangles for the two integrals are equal in area. The areas of matching strips being

$$\frac{b-a}{n} \cdot \frac{1}{a+p\frac{b-a}{n}} = \frac{tb-ta}{n} \cdot \frac{1}{ta+p\frac{tb-ta}{n}}$$

Corresponding rectangles just inside the curve are also equal in area. Letting n tend to infinity we have

$\int_a^b \frac{dx}{x} = \int_{ta}^{tb} \frac{dx}{x}$, and so $\int_1^{ab} \frac{dx}{x} = \int_1^a \frac{dx}{x} + \int_a^{ab} \frac{dx}{x} = \int_1^a \frac{dx}{x} + \int_1^b \frac{dx}{x}$,

which is the logarithmic property.

5. Generalisation of B3 and B4 to monotonic functions

The graphs of discontinuous functions do not appear to bound an area, so their integration was not considered until they appeared as the limits of Fourier series in the nineteenth century. The analytic definition of the Riemann integral (1854, but published 1867, see D11 below) opened up this possibility, and Darboux (1875) and others determined necessary and sufficient conditions for integrability in Riemann's sense (see D12 below). For one particular class of functions, Newton's (1687) generalisation of the method of Fermat and Roberval, translated in Fauvel and Gray (1987), p.391, lemma II) is adequate to establish the integrability of any real monotonic function, f, whether continuous or not, because it establishes the arbitrary closeness of upper and lower sums.

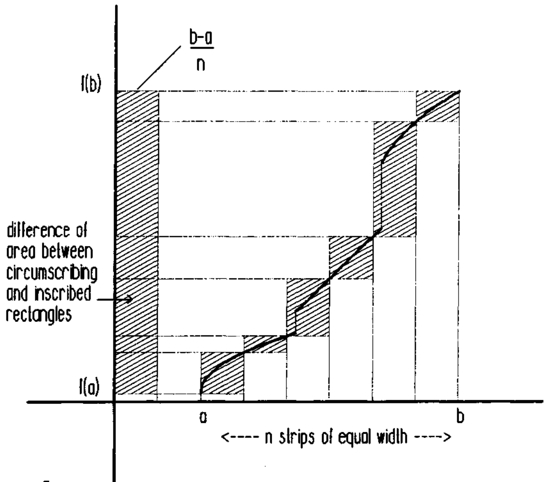


Figure 5

If the domain of such a function $[a, b]$ is divided into n equal segments, the difference between an *upper sum* (of areas of minimal circumscribing rectangles) and a *lower sum* (of areas of maximal inscribed rectangles) is $(f(b)-f(a))\dfrac{b-a}{n}$. This difference tends to 0 as n tends to infinity, as in the Greek method of exhaustion. The upper sums are evidently greater than the area being sought and the lower sums less than the area being sought. Since the upper and lower sums are arbitrarily close, they define a Dedekind cut, and the integral is well defined.

Unlike sections A2 and B3, where the rational numbers suffice, the argument here, as also in A1 and B4, presumes the completeness of the real numbers, and indeed such a proof belongs properly within Weierstrassian analysis.

C. Parallel rectangular strips of unequal width

6. A geometrical dissection to find $\int_a^\infty \frac{dx}{x^2}$.

Fermat (c. 1640), translated i Struik (1969), p. 219-222, and Mahoney (1994), p. 244-254.
Fermat acknowledges at the end of his correspondence with Roberval in 1636 that the method of section B3 could not be used to obtain $\int_0^a x^k dx$, unless k were a positive integer or its reciprocal. In Fermat's unpublished *Treatise on Quadrature* (1658), he exhibited a new method to tackle the other possibilities.

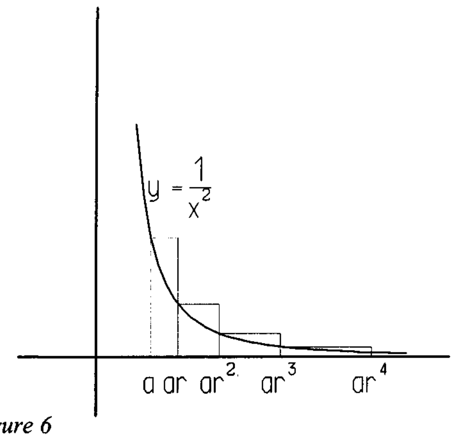


Figure 6

Taking $r > 1$, and circumscribing the curve $y = 1/x^2$ with rectangles based on the x-axis on the segments $[a, ar]$, $[ar, ar^2]$, $[ar^2, ar^3]$, ... ,

$$\text{the upper sum} = a(r - 1)\frac{1}{a^2} + a(r^2 - r)\frac{1}{a^2r^2} + a(r^3 - r^2)\frac{1}{a^2r^4} + \ldots$$

$$= \frac{r-1}{a}\left(1 + \frac{1}{r} + \frac{1}{r^2} + \ldots\right) = \frac{r-1}{a}.\frac{1}{1-\frac{1}{r}} = \frac{r}{a}.$$

Now let r tend to 1 and the area under the curve is $1/a$.
Fermat recognised that this method could be applied to any integral of the form $\int_a^\infty \frac{dx}{x^k}$ when k is a rational number greater than 1, and (using $r < 1$) to any integral of the form $\int_0^a x^k dx$ when k is a positive rational number.

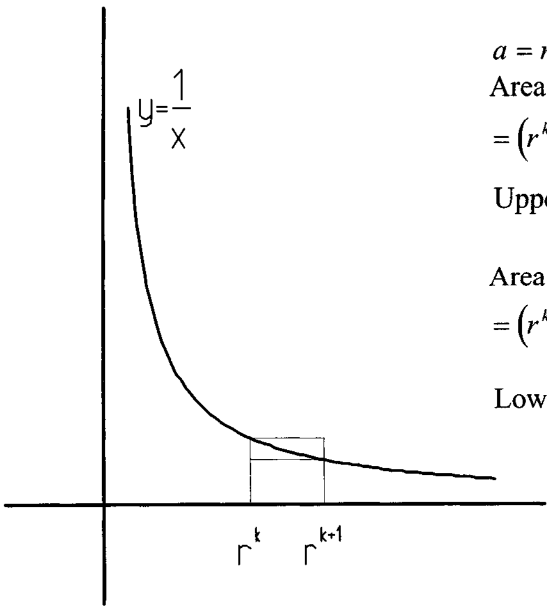
7. A geometrical dissection applied to $\int_1^a \frac{dx}{x}$.

At the end of chapter VII on logarithms and exponentials in *Introductio ad analysin infinitorum* (1748) Euler obtained the limit

$n(\sqrt[n]{a} - 1) \to \ln a$ as n tends to infinity by inverting the limit

$$\left(1+\frac{x}{n}\right)^n \to e^x.$$

In Gregory of St. Vincent's study of $y = 1/x$ (1647) he had shown that if the area between the curve and the x-axis is divided by lines parallel to the y-axis, but spaced out on the x-axis in a geometric progression, the areas marked off are equal. (Boyer, (1949) p 160) This is justified in B4 above.



$a = r^n$

Area of upper rectangle

$= \left(r^{k+1} - r^k\right)\frac{1}{r^k} = r - 1$

Upper sum $= n(r-1)$

Area of lower rectangle:

$= \left(r^{k+1} - r^k\right)\frac{1}{r^{k+1}} = \frac{r-1}{r}$

Lower sum $= n\frac{r-1}{r}$

Figure 7

If the segment $[1, a]$ is partitioned into n pieces in this way we obtain:

$$\text{lower sum} = \frac{n(\sqrt[n]{a}-1)}{\sqrt[n]{a}} < \int_1^a \frac{dx}{x} < n(\sqrt[n]{a}-1) = \text{upper sum},$$

which gives the logarithmic limit as $n \to \infty$, since $\sqrt[n]{a} \to 1$. This particular derivation is not old.

8. The fundamental theorem of calculus, with Newton/Leibniz assumptions

Before Newton's period away from Cambridge (1665-1666), independent methods of computing the slopes of tangents and areas under curves had revealed that, for polynomials, these two kinds of computation were algebraic inverses of each other. Newton, with his dynamic view of curves formed by a moving point, found a reason why, which he described in an unpublished tract of 1666, translated in Whiteside (1968), p. 427 and 430. Newton saw that the rate of change of the area under a curve was given by the y-coordinate at the point of change. If o is a small amount of time, and the rate of change of x denoted by $\dot{x}$, then, as in figure 8, $o\dot{A} \approx y \cdot o\dot{x}$, and, in the limit, $\dot{A} / \dot{x} = y$. More obviously the slope of the tangent is $\dot{y} / \dot{x}$, and so the relationship between area and the y-coordinate is matched by relationship between the y-coordinate and the tangent slope. This is the fundamental theorem.

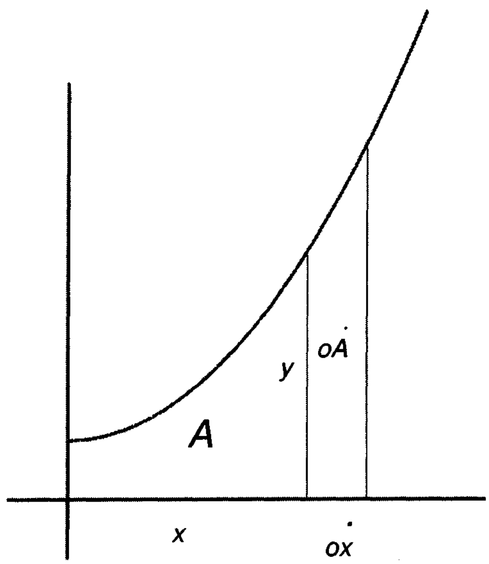


Figure 8

Newton illustrated rather than proved this, in 1669, (translated in Whiteside, 1968, p 207) by taking $A = \dfrac{n}{n+m} a x^{\frac{m+n}{n}}$ and obtaining $\dot{A} = a x^{\frac{m}{n}} \dot{x}$, so $y = a x^{\frac{m}{n}}$. He then claimed the converse.

The (implied) intermediate steps are

1) $A + o\dot{A} = \dfrac{n}{n+m} a(x + o\dot{x})^{\frac{m+n}{n}}$,

2) an expansion by the binomial theorem,

3) division by o,

4) regarding o as negligible.

These are exactly the same algebraic steps as when finding slopes.

Leibniz came to the same conclusion after examining how sums and differences combined with each other. Our integral sign is Leibniz' drawn out S for sum, and the d in our dx is Leibniz' initial for difference.

Both Newton's and Leibniz's notation were geared to indefinite integrals and the generality of their claims made it necessary to treat areas as having a sign (±). Both of these steps are steps away from a purely geometric notion of integral.

All Newton's functions were (at least locally) both continuous and monotonic, though that is not the way either Newton or Leibniz would have referred to them. But it is just such conditions on functions which makes it natural to view them as curves, and this in turn makes the fundamental theorem easily accessible.

This theorem was so fruitful that throughout the eighteenth century integration became synonymous with anti-differentiation. Integrals were habitually taken to be indefinite integrals, and definite integrals were evaluated by finding the difference between two values of an anti-derivative. Leibniz's account is given in Struik (1969), (p. 282).

D. From geometry to algebra

9. Cauchy's definition of the integral (1823)

Throughout the eighteenth century, the power of the fundamental theorem meant that integration was regarded as the opposite of differentiation. The functions Cauchy considered (and it must be said that Cauchy looked at functions rather than curves) included sin $(1/x)$ (Cauchy 1821, p.14) which, though continuous away from the origin, had no familiar anti-derivative and was not monotonic in any neighbourhood of the origin. In his *Résumé des Leçons sur le Calcul Infinitésimal*, (Cauchy 1823), Cauchy says that his first concern is to establish the existence of integrals before studying their properties, and it was he who insisted on the primary status of the definite integral. So in *Leçon* 21, he considered a function f continuous on a closed interval $[a, b]$ and worked from an arbitrary dissection of the interval into n pieces: $a = x_0 < x_1 < x_2 < \ldots < x_n = b$.

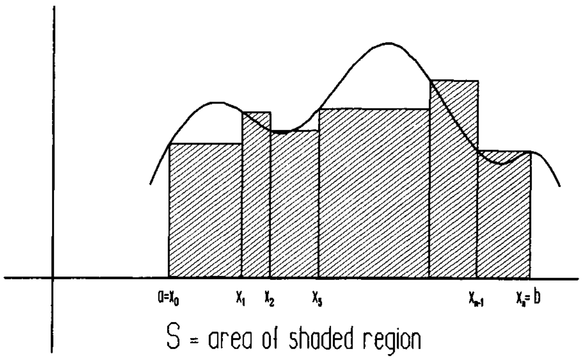


Figure 9

He examined the sum

$$S = f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_1) + f(x_2)(x_3 - x_2) + \ldots + f(x_{n-1})(x_n - x_{n-1}).$$

(Look and see the connection between this sum and the terms in the computations in B3, B4, B5, C6 and C7 above, to see how Cauchy's definition is a generalisation of earlier ideas.) He then conceived a limit of S as the dissection becomes finer and the greatest of the $(x_{i+1} - x_i)$ tends to 0.

Cauchy did not distinguish between 'continuity' and 'uniform continuity' in the modern sense, so he was able to show that when the dissection was fine enough the different possible values of S were close together, but the proof of the existence of a limit for S depended on a clearer notion of the completeness of the real numbers than Cauchy had.

10. Cauchy's version of the fundamental theorem (1823)

We will illustrate the novelty and precision in Cauchy's version of the fundamental theorem of the calculus in two steps, firstly with continuous and monotonic functions (Newton's curves) and then as Cauchy did, with arbitrary continuous bounded functions.

In was only from 1820 that Fourier proposed the notation $\int_a^b f(x)dx$ for a definite integral, and it is this notation which assists in the definition of the integral as a function of x

$$F(x) = \int_a^x f(t)dt,$$

whether there is an algebraic formula for F or not. We then have

$$F(x + h) - F(x) = \int_x^{x+h} f(t)dt.$$

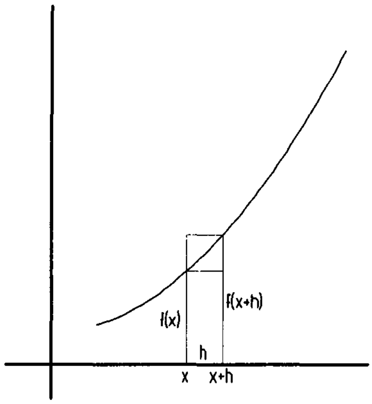


Figure 10

To see the structure of Cauchy's argument, we will first see it working in Newton's context of a continuous and monotonic increasing function f.

$$h{\cdot}f(x) \le \int_x^{x+h} f(t)dt \le h{\cdot}f(x+h),$$

$$\text{so } f(x) \le \frac{F(x+h)-F(x)}{h} \le f(x+h).$$

Now let $h \to 0$. Then $\dfrac{F(x+h)-F(x)}{h} \to f(x)$, since f is continuous, and so $F'(x) = f(x)$.

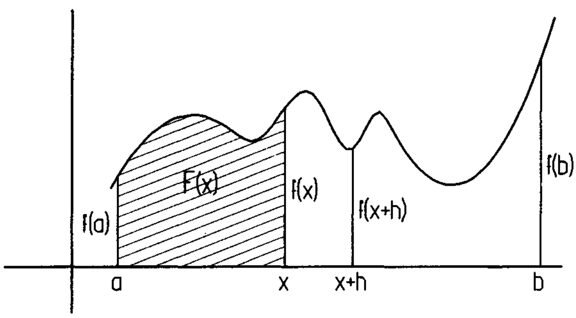


Figure 11

Cauchy's argument, in Cauchy (1823), of course requires that the function f is continuous, but not necessarily monotonic. The definition of the integral, as a function of x, given above, still holds.

$$F(x) = \int_a^x f(t)dt\text{, for } a \leq x \leq b.$$

And then $F(x + h) - F(x) = \int_x^{x+h} f(t)dt = h{\cdot}f(x + \theta h)$, for some θ lying between 0 and 1, appealing to his Intermediate Value Theorem for continuous functions, since the value of the integral lies between the greatest and least values of $h{\cdot}f(t)$ for $x \leq t \leq x + h$.

Now $\dfrac{F(x+h)-F(x)}{h} = f(x + \theta h)$, and as $h \to 0$,

we have $F'(x) = f(x)$, since f is continuous.

This proof appears in *Leçon* 26 of Cauchy (1823), and this, or something very similar to it, appears in every first course in real analysis, today.

11. The Riemann integral (1854)

Translated in Birkhoff (1973), p.21-23.
In attempting to characterise those functions which are the limits of their own Fourier series, Riemann (1854) found that Dirichlet's (1829) condition of piecewise continuity could be replaced by a condition of integrability. This raised the question of what functions might be integrable, in a new way. Cauchy had defined infinite integrals with a definition such as $\int_a^\infty f(x)dx = \lim_{b\to\infty} \int_a^b f(x)dx$.

Riemann recognised that Cauchy's definition of the integral (which had only been stated for continuous functions) might accomodate $f(x)$ becoming infinitely large at a point, b, by defining

$\int_a^c f(x)dx$ as $\lim_{\varepsilon\to 0}\left(\int_a^{b-\varepsilon} f(x)dx + \int_{b+\varepsilon}^c f(x)dx\right)$.

He used this example to prepare his readers for something more general.
Riemann, like Cauchy, examined functions which were defined on a closed interval $[a, b]$, but, unlike Cauchy, without continuity as a precondition. He considerd an arbitrary dissection of the interval into n pieces: $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. He examined the sum

$$S = f(t_0)(x_1 - x_0) + f(t_1)(x_2 - x_1) + f(t_2)(x_3 - x_2) + \ldots + f(t_{n-1})(x_n - x_{n-1}),$$

where $x_i \leq t_i \leq x_{i+1}$ and said that the function f was *integrable* when S has a limit as max $(x_{i+1} - x_i) \to 0$. Riemann constructed a function which was integrable in this sense but which was discontinuous on a dense set of points. He proved its integrability by controlling the size of the intervals on which the function oscillates finitely. Riemann had taken the sum S, which Cauchy had constructed to prove that continuous functions on closed intervals were integrable, and used it to define integrability itself. Freed from Cauchy's condition of continuity, Cauchy's and Riemann's definitions are equivalent. But with Riemann, the notion of area under a curve had become algebraic, with the function f integrable on [0, 1] defined by f(irrational) = 0, $f(p/q) = 1/q$. (This example is due to Thomæ (1875), page 14.)

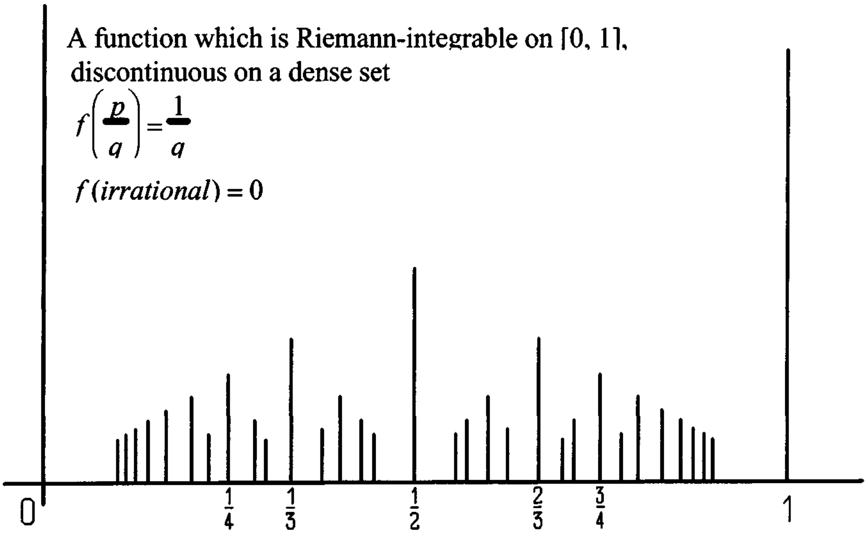


Figure 12

Riemann's own example in 1854 was S$((nx))/n^2$, where $((nx)) = 0$ when $nx - [nx] = ½$, and $((nx)) = nx - [nx + ½]$ otherwise. This is hard to graph but can be computed.

12. Upper and lower sums

Translated in Darboux (1875).

What was the difference between Riemann's discontinuous but integrable function, and Dirichlet's non-integrable f given by f(rational) = 1, f(irrational) = 0? Riemann described the difference in terms of the intervals on which finite oscillations take place. Another description was proposed in 1875 by both Darboux and Du Bois Reymond which has since become standard. It builds on the clearer understanding of the completeness of the real numbers which had come with Dedekind and others by 1872.

With completeness, a function which is bounded on an interval has both a least upper bound and greatest lower bound on that interval. So, if $f(x)$ is bounded on $[x_i, x_{i+1}]$ it has a least upper bound M_i on that interval and a greatest lower bound m_i, and
$m_i(x_{i+1} - x_i) \le f(t_i)(x_{i+1} - x_i) \le M_i(x_{i+1} - x_i)$, for all t_i such that $x_i \le t_i \le x_{i+1}$.

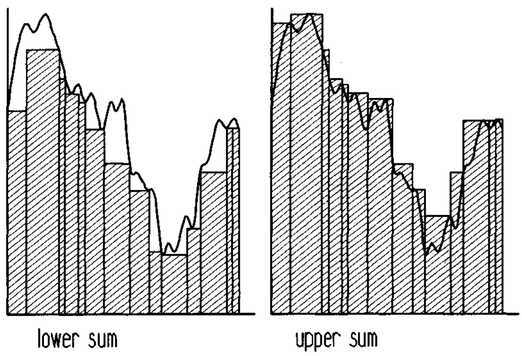


Figure 13

Adding the n inequalities relating to the parts of S as in D11 above, we get

$$\sum_{i=0}^{i=n-1} m_i(x_{i+1} - x_i) = \text{a lower sum} \le S \le \text{an upper sum} = \sum_{i=0}^{i=n-1} M_i(x_{i+1} - x_i)$$

When the dissection gets finer, the lower sum increases and the upper sum diminishes. But *any* lower sum can be shown to be less than *any* upper sum by uniting their two dissections. If the lower sums and upper sums can get arbitrarily close, they belong to the two sides of a Dedekind cut and define a unique limit, which is the integral. With Dirichlet's non-integrable function, the least upper bounds are all 1 and the greatest lower bounds are all 0. With Thomæ's integrable function the greatest lower bounds are again 0, and although none of the least upper bounds are 0, there are arbitrarily close upper and lower sums.

With completeness, *a continuous function on a closed interval* attains its bounds and can be shown to be uniformly continuous. So for a given positive $\varepsilon/(b - a)$ we can choose δ such that
$(x_{i+1} - x_i) < \delta$ implies $(M_i - m_i) < \varepsilon/(b - a)$, and then $max(x_{i+1} - x_i) < \delta$ implies (upper sum - lower sum) $< \varepsilon$, making f integrable. At last we have a full proof that continuous functions are integrable.

Toeplitz (1963) considered his 'genetic' introduction as the content of a lecture course. In Freudenthal's language, Freudenthal (1983) (p.ix), Toeplitz confronts and reverses the 'didactical inversion' that so commonly follows any major mathematical achievement (in this case, the definition of the Riemann integral). If the steps described here, in **B**, **C** and **D**, are problematised, and used to form a sequence of exercises it may well be that the last step, to what Freudenthal calls a 'genetic' structure may be taken, Freudenthal (1983), p.10. I suggest that the material here should not be *lectured* except to an audience already familiar with the theory. It can be broken down into exercise-sized chunks for study, as has been partly effected in Burn (1997).

An excellent account of the history of integration during the 19th century is given in Hawkins (1975).

—•—

Acknowledgement
I wish to acknowledge the advice of Otto Bekken and also of an anonymous referee for helping me to improve the historical quality of this paper.

References

Birkhoff, G. ed. (1973). *A Source Book in Classical Analysis*. Cambridge, US: Harvard University Press.

Boyer, C.B. (1949). *The history of the calculus*. New York: Dover reprint 1959.

Burn, R.P. (1997). *Numbers and Functions*. Cambridge University Press.

Cauchy, A-L. (1821). *Analyse Algébrique*. Éditions, Paris: Jacques Gabay. Reprint 1989.

Cauchy, A-L. (1823). *Résumé des leçons sur le Calcul Infinitésimal*. Paris: Ellipses reprint.

Darboux, G. (1875). Mémoire sur la théorie des fonctions discontinues. *Annales Sci. École Normale Supérieure. Sér.*2, **4**, 57-112.

Euler, L. (1748). *Introductio in analysin infinitorum*. Vol.1. English translation 1988. Berlin: Springer.

Fauvel, J. and Gray, J. (1987). *The History of Mathematics, a Reader*. Open University Press.

Freudenthal, Hans, (1983). *Didactical Phenomenology of Mathematical Structures*. Dordrecht: D. Reidel.

Hawkins, T. (2nd ed. 1975). *Lebesgue's theory of integration.* New York: Chelsea.

Katz, V.J. (1993). *A history of mathematics.* New York: Harper-Collins.

Mahoney, M.S. (1994). *The mathematical career of Pierre de Fermat.* New York: Princeton University Press.

Struik, D.J. ed. (1969). *A Source Book in Mathematics,* 1200 - 1800. Cambridge, US: Harvard University Press.

Thomæ, J. (1875) *Einleitung in die Theorie der bestimmten Integrale.* Halle.

Toeplitz, O. (1963) *The Calculus, a genetic approach.* English translation. Chicago: Chicago University Press.

Whiteside, D.T. ed. (1967). *The Mathematical Papers of Isaac Newton,* vol I. Cambridge, UK: Cambridge University Press.

Whiteside, D.T. ed. (1968). *The Mathematical Papers of Isaac Newton,* vol II. Cambridge, UK: Cambridge University Press.

Abstract (in Norwegian)

Den genetiske metoden til O.Toeplitz går ut på en strukturering av det fastsatte pensum ut fra historisk innsikt og historiske problemstillinger, ikke som et historiekurs, men for å styrke forståelsen av moderne matematikk hos dagens studenter. Denne artikkelen beskriver hvordan Toeplitz' methode kan brukes til å gi en passende innforing i integrasjon for et grunnkurs i analyse. Etter to eksempler på den greske utfyllingsmethoden (beregnet mer på lærere enn studenter) blir den historiske utviklingen i 17. og 19. århundre brukt til å belyse de begrepsmessige trinnene en student må ta i et tradisjonelt grunnkurs i analyse fram til Riemann-integralet.

Author

Robert P. Burn is professor of Mathematics Education at Høgskolen i Agder, Kristiansand, Norway

Research interests

Undergraduate mathematics education, history of mathematics, geometry.

Address

Prof Robert P. Burn, Department of Mathematics, Agder College, Tordenskjoldsgate 65, 4604 Kristiansand, Norway

Phone: (+47) 38 14 16 31

Fax: (+47) 38 14 10 11

E-mail: Robert.P.Burn@hia.no