

Chapter 2

Curves and Tangents

2.1 Curves

In her sourcebook of mathematics,¹ Jacqueline Stedall represents Euclid via facsimile reproductions of selections from Isaac Barrow's 1660 edition of the *Elements* written for undergraduates. The definition given of a curve reads

A Line is a longitude without latitude.²

The most familiar English translation, available in multiple editions, is due to Thomas Little Heath.³ In this translation the definition reads

A line is *breathless length*.⁴

The word “line” in these quotes is used to mean a curve, our line being called a “straight line”. The annotated edition of Heath's translation follows this statement with several pages of discussion,⁵ beginning with the attribution of this definition to Plato's school. He cites Aristotle's quibble about the negative form of the definition and also offers a couple of alternative definitions from Aristotle (384–322 B.C.) already cited in the commentary of Proclus, and then goes on to discuss classifications of curves and offering a number of examples of such. I quote Proclus (412–485):

¹Jacqueline Stedall (ed.), *Mathematics Emerging: A Sourcebook 1540–1900*, Oxford University Press, Oxford, 2008.

²*Ibid.*, p. 10.

³Thomas Little Heath, *The Thirteen Books of Euclid's Elements*, 3 volumes, Cambridge University Press, Cambridge, 1908. This translation has been reprinted a number of times. The edition put out by Dover Publications includes all the annotations. Two other editions currently in print but lacking the annotations are that in the series *Great Books of the Western World* and an attractively typeset single volume published by Green Lion Press.

⁴*Ibid.*, vol. 1, p. 158.

⁵*Ibid.*, pp. 158–165.

II. A line is length without breadth.

The line is second in order⁶ as the first and simplest extension, what our geometer calls “length,” adding “without breadth” because the line also has the relation of a principle to the surface. He taught us what the point is through negations only, since it is the principle of all magnitudes; but the line he explains partly by affirmation and partly by negation. The line is length, and in this respect it goes beyond the undividedness of the point; yet it is without breadth, since it is devoid of the other dimensions. For everything that is without breadth is also without depth, but the converse is not true. Thus in denying breadth of it he has also taken away depth, and this is why he does not add “without depth,” since this is implied in the absence of breadth.

The line has also been defined in other ways. Some define it as the “flowing of a point;” others as “magnitude extended in one direction.” The latter definition indicates perfectly the nature of the line, but that which calls it the flowing of a point appears to explain it in terms of its generative cause and sets before us not the line in general, but the material line. This line owes its being to the point, which, though without parts, is the cause of the existence of all divisible things; and the “flowing” indicates the forthgoing of the point and its generative power that extends to every dimension without diminution and, remaining itself the same, provides existence to all divisible things.⁷

As the quotation from Proclus makes clear, the Euclidean definition takes its one-dimensionality as the defining property of a curve, thus distinguishing it from the zero-dimensionality of a point, the two dimensionality of a surface, and the three-dimensionality of a solid. There is, of course, a certain intuitiveness to this definition, but as a basis on which to establish theorems it leaves a lot to be desired. It is not something for the Calculus course, but a matter of the more advanced field of Topology. The definition of the “flowing of a point”, i.e., a kinematic approach, which can be found in Aristotle, is more accessible and is abstractly what the standard definitions in the Calculus are based on. The “magnitude extended in one direction” also harks back to Aristotle. Heath explains

A line is, according to Aristotle, a magnitude “*divisible* in one way only”, in contrast to a magnitude divisible in *two* ways, or a surface, and a magnitude divisible “in all or in three ways”, or a body; or it is a magnitude “*continuous* one way (or in one direction),” as compared with magnitudes continuous in *two* ways or *three* ways, which curiously enough he describes as “breadth” and “depth” respectively, though he immediately adds that “length” means a line, “breadth” a surface, and “depth” a body.⁸

Without some more abstract topological concepts, one would be hard put to make these definitions mathematically precise. About the only thing that seems clear is that a line, i.e., curve, is to have only one dimension, whatever that might mean.

⁶Euclid first defined the point in Definition I as that which has no part.

⁷Proclus (Glenn R. Morrow, ed.), *A Commentary on the First Book of Euclid's Elements*, Princeton University Press, Princeton, 1970, pp. 79–80.

⁸Heath, *Elements, op. cit.*, vol. 1, pp. 158–159. I have omitted his parenthetical insertions of Greek terms and page references in Aristotle.

Our intuition, however, is often misleading. In an oft-cited paper published in 1933,⁹ Hans Hahn (1879–1934) discusses this point with reference to the notion of a curve. Our two intuitions of a curve as the “flowing of a point” and as a one-dimensional entity, though properties common to most things we consider to be curves, are not equivalent and, when given rigorous formal definitions, both can be shown not to agree completely with our intuition of what constitutes a curve. The best that one can usually hope for in mathematics is to replace a vague intuitive notion by a precise, formally defined one that agrees with intuition in all familiar cases and all future cases not too dissimilar to these. It can happen, after giving such a definition, that at the fringes of our experience there are objects our intuition might accept but our formal definition rejects or *vice versa*. When this happens we may amend our formal definition to accommodate or to exclude the new objects, or we may define a new class of objects. In the case of curves, Hahn offers two definitions based on the “flowing of a point” and one-dimensionality, respectively, and shows by example that each concept accepts some questionable curves and rejects some things we might intuitively accept as curves. The question is not one of giving a definition that precisely captures the intuitive notion — after all, different people have different intuitions — but to offer, as Hahn puts it, a “serviceable definition”:

Since the time-honoured definition of a curve fails to cover the fundamental concept, what other more serviceable definition can be substituted for it?¹⁰

The word “serviceable” is relative, not absolute. A definition of “curve” is serviceable in a given context if it applies to those curves we are likely to come across in that context and does not apply to the non-curves we are likely to meet. Our context is the first-year Calculus course, not an advanced Topology course, and we don’t need Hahn’s abstract definitions that we would be hard pressed to apply usefully in the Calculus course.

So how are we to define a curve here?

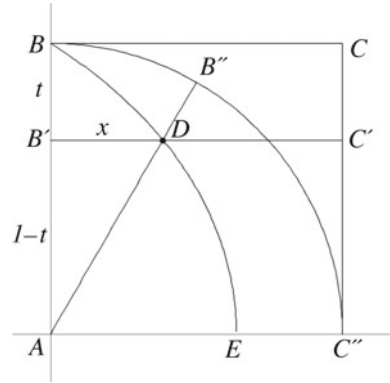
The first step toward isolating a formal notion of curve, one precise enough on which to base proofs, is to catalogue some curves and look for a commonality in their modes of definition. Traditional Geometry doesn’t have a lot of curves. If you look into Euclid’s *Elements*, you will find straight lines and circles. Other curves were known. However, Boyer, in his *History of Analytic Geometry*, says that the Greeks “did not discover more than half a dozen new curves in all of their enormous mathematical activity, and these were not systematically classified”.¹¹ The first of

⁹Hans Hahn, “Die Krise der Anschauung”, in: *Krise und Neuaufbau in den exakten Wissenschaften*, F. Deuticke, Leipzig and Vienna, 1933. An English translation, “The crisis of intuition”, appears in: Hans Hahn (Brian McGuinness, ed.), *Empiricism, Logic, and Mathematics: Philosophical Papers*, D. Reidel Publishing Company, Dordrecht, 1980.

¹⁰Hahn, “The crisis of intuition”, *op. cit.*, p. 88.

¹¹Carl B. Boyer, *History of Analytic Geometry*, The Scholars Bookshelf, Princeton Junction (NJ), 1988, p. 20. This work was originally published in 1956 as numbers 6 and 7 of *The Scripta Mathematica Studies*. Incidentally, the numerical estimate given here is figurative, not literal: Boyer cites at least half a dozen curves known to the Greeks and on page 35 announces, “yet scarcely a dozen curves were familiar to the ancients”.

Fig. 2.2 Quadratrix, a
Second View



i.e.,

$$x = (1 - t) \tan \frac{t\pi}{2}, \quad y = 1 - t, \quad \text{for } 0 \leq t < 1. \quad (2.1)$$

Formula (2.1) provides a *parametric definition* of the curve, defining x and y as functions $x(t)$ and $y(t)$ of the parameter t . Using this we can readily solve for x as a function of y :

$$\begin{aligned} x &= (1 - t) \tan \frac{t\pi}{2} = y \tan \frac{(1 - y)\pi}{2} = y \tan \left(\frac{\pi}{2} - \frac{y\pi}{2} \right) \\ &= y \cot \frac{y\pi}{2}, \quad \text{as } \tan \left(\frac{\pi}{2} - \theta \right) = \cot \theta. \end{aligned} \quad (2.2)$$

Here, $0 < y \leq 1$.

A glance at the graph informs us that y is also a function of x for some appropriate values of x . However, it is unlikely one will be able to find a closed form for expressing y in terms of x . Moreover, specifying the domain of such a function $y(x)$ is not trivial. The variable ranges over the interval $[0, a]$, where a is the x -coordinate of E . a should be $x(1)$ if we express x as a function of t à la (2.1) or $x(0)$ if we express x as a function of y . Unfortunately, for $t = 1$, (2.1) expresses x in the form $0 \cdot \infty$ and (2.2) expresses x in the form $0/0$. Either way it yields no value. Indeed, if we look to the definition of the quadratrix to see where the point E should be, namely at “the” point of intersection of $B'C'$ and AB'' at $t = 1$, we see the problem: the line segments intersect at all points of AC'' . We must determine E by some other means, as we shall do in the next chapter when we discuss L'Hôpital's Rule. Before doing this, however, we can see why the quadratrix as plotted in Figs. 2.1 and 2.2 does not quite reach the point E . The coordinate x not being calculable at $t = 1$ where $y = 0$, the curve stops short at the last calculated value of t . Depending on the resolution of one's display and the density of the values of t in one's table, one may or may not notice the gap. It is not visible on my graphing calculator, but is clearly visible with the computer software I used to generate the graphs.

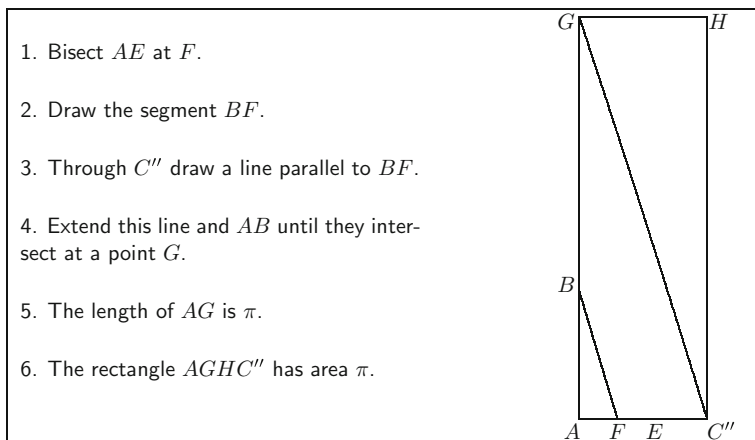


Fig. 2.3 Squaring the circle

Anticipating this calculation, I can report that the coordinates of the point E are $(2/\pi, 0)$, i.e., if AB is taken to be the unit, the length of AE is $2/\pi$. Standard ruler and compass constructions readily yield a segment of length π , whence a rectangle of area π from the segment AE . (See Fig. 2.3.) And Euclid's *Elements* shows how to find a square of area equal to that of any given rectangle. Thus we have used the quadratrix to square the circle of radius 1.

It was this application of the quadratrix to the problem of quadrature that gave the curve its name. Boyer,¹² admittedly not the most up-to-date reference, but a convenient one, tells us this was a later development due to Dinostratus (*fl. c.* 350 B.C.); Hippias himself seems to have invented it to solve the trisection and other multisection problems. This is an easy matter: Given an angle $\angle ABC$ of less than 90° , copy it to one's diagram of the quadratrix as in Fig. 2.4. Let D be the point of intersection of the elevated ray of the angle with the quadratrix. Draw a horizontal line through D and let it intersect the vertical line through the vertex B of the angle at a point E . Let BF equal a third of BE and let G be the point of intersection of the horizontal line through F with the quadratrix. The angle $\angle GBC$ is one third the angle $\angle ABC$, this last because of the uniform motions of the lines generating the quadratrix.

In the larger scheme of things, the applications of the quadratrix to the solution of two of the three classical problems are no more than amusing asides. However, these problems seem to have been the inspiration behind the next family of curves to arrive on the scene — the conic sections. Menæchmus, often cited as the mathematical tutor of Alexander the Great, was also the brother of Dinostratus, a connexion of greater immediate relevance here. One of Plato's contemporaries, Archytas (*fl. c.* 395 B.C.) had solved the problem of duplicating the cube through application of a cone, a cylinder, and a torus. Menæchmus realised that conic sections alone could do the trick.

¹²Boyer, *op. cit.*, p. 11.

2.1.1 Definition Let a line L , a point F not lying on L , and a positive real number e be given. Consider the *locus* (i.e., set) of all points P such that the distance from P to F equals e times the distance from P to L :

$$\gamma = \{P | \text{dist}(P, F) = e \cdot \text{dist}(P, L)\}.$$

γ is called an *ellipse* if $e < 1$, a *parabola* if $e = 1$, and a *hyperbola* if $e > 1$. F , L , and e are called the *focus*, *directrix*, and *eccentricity*, respectively, of γ .

The advent of symbolic algebra in 16th century Europe and the growing shift in emphasis from geometry to algebra made inevitable the invention of Analytic Geometry, as evidenced by the fact that the invention occurred twice at the hands of Fermat and Descartes around 1630. Both men studied conic sections and their analytic expression.

Equational representations for the conic sections are readily derivable. Suppose F , L , and e are given. We can take L to be the x -axis, the normal connecting F to L to be the y -axis, and the distance from F to L to be the unit. With respect to these axes, F has coordinates $(0, 1)$. A point P with coordinates (x, y) lies on γ just in case

$$\text{dist}(P, F) = \sqrt{(x-0)^2 + (y-1)^2} = e \cdot |y| = e \cdot \text{dist}(P, L).$$

Squaring the terms of the central equation yields

$$x^2 + (y-1)^2 = e^2 y^2, \quad (2.3)$$

i.e.,

$$x^2 + (1 - e^2)y^2 - 2y + 1 = 0. \quad (2.4)$$

For $e = 1$, (2.4) becomes

$$y = \frac{1}{2}x^2 + \frac{1}{2},$$

which we can quickly enough graph to obtain a recognisably parabolic shape.

For $e \neq 1$, we can divide by $1 - e^2$ to get

$$\begin{aligned} \frac{x^2}{1 - e^2} + y^2 - \frac{2}{1 - e^2}y &= -\frac{1}{1 - e^2} \\ \frac{x^2}{1 - e^2} + \left(y - \frac{1}{1 - e^2}\right)^2 &= -\frac{1}{1 - e^2} + \frac{1}{(1 - e^2)^2} = \frac{e^2}{(1 - e^2)^2}. \end{aligned} \quad (2.5)$$

If $e < 1$, one has $0 < 1 - e^2 < 1$, and Eq. (2.5) assumes the form

$$ax^2 + b(y - \beta)^2 = c$$

with a, b, c all positive. And, if $e > 1$, $1 - e^2 < 0$ and (2.5) assumes the form

$$ax^2 + b(y - \beta)^2 = c$$

with a negative and b, c positive. In either case one can solve for y to get

$$y = \beta \pm \sqrt{\frac{c - ax^2}{b}}, \quad (2.6)$$

graph the two resulting functions, and recognise the familiar elliptical shape when $e < 1$ and hyperbolic one when $e > 1$.

2.1.2 Exercise Carry out the above derivation for the following values of e :

- i. $e = 1/2$
- ii. $e = 1$
- iii. $e = 2$.

Graph the resulting curves to verify they are of the appropriate forms.

Definition 2.1.1 does not capture all conic sections. Missing are the degenerate conics — points, lines, and circles. There are two exceptions for $e = 0$ and $e = \infty$: For fixed F and L , if one graphs (2.6) for successively smaller values of e , one gets ellipses that become more and more circular. But they also become progressively smaller and at $e = 0$, the graph consists solely of the focus F . Indeed, plugging 0 in for e the Eq. (2.3) results in

$$x^2 + (y - 1)^2 = 0,$$

the equation of the circle of radius 0 centred at $\langle 0, 1 \rangle$. At the other extreme, larger and larger values of e give graphs of hyperbolas hugging more and more closely to their asymptotes, which themselves are closing scissors-like towards the x -axis. And, indeed, plugging¹³ ∞ in for e in (2.5) results in the equation,

¹³One can go a long way calculating with ∞ taking ∞ as an ideal element and applying rules like

$$a \pm \infty = \pm\infty, \quad a \cdot \infty = \infty, \quad a/\infty = 0$$

for real a . Terms like $0 \cdot \infty$, $\infty - \infty$, and ∞/∞ are indeterminate and simple algebra doesn't apply. In fact, I have cheated in writing $\infty^2/\infty^4 = 1/\infty^2 = 0$. One should first manipulate (2.5) to express

$$\frac{e^2}{(1 - e^2)^2} = \frac{1}{(1 - e^2)^2/e^2} = \frac{1}{(1/e - e)^2}$$

and only then plugging ∞ in for e :

$$\frac{1}{(1/\infty - \infty)^2} = \frac{1}{(0 - \infty)^2} = \frac{1}{\infty^2} = 0.$$

$$\begin{aligned}
\frac{1}{1 - \infty^2}x^2 + \left(y - \frac{1}{1 - \infty^2}\right)^2 &= \frac{\infty^2}{(1 - \infty^2)^2} \\
\frac{1}{-\infty^2}x^2 + \left(y - \frac{1}{-\infty^2}\right)^2 &= \frac{\infty^2}{\infty^4} \\
0x^2 + (y - 0)^2 &= \frac{1}{\infty^2} \\
0x^2 + y^2 &= 0,
\end{aligned}$$

i.e., $y = 0$. Thus $e = \infty$ yields the directrix itself as the resulting conic section.

Relative to a pre-existing pair of coordinate axes, the focus will not necessarily have as simple a pair of coordinates and the equation of the directrix will be more complicated than $y = 0$. The computation becomes more involved, but it follows roughly the same lines.

2.1.3 Example Let F be $\langle 1, 2 \rangle$, L be given by $x + 2y = 3$, and $e = 2$. The first step is to determine the distance from a point $\langle \alpha, \beta \rangle$ to L . To this end note that a line perpendicular to L has an equation $2x - y = c$ for some constant c . For $\langle \alpha, \beta \rangle$ to lie on the perpendicular in question one must have $c = 2\alpha - \beta$. The point on L closest to $\langle \alpha, \beta \rangle$ is the point $\langle x, y \rangle$ of intersection of the lines:

$$\begin{aligned}
x + 2y &= 3 \\
2x - y &= 2\alpha - \beta.
\end{aligned}$$

Doubling the first of these and subtracting the second from the result yields

$$\begin{aligned}
5y &= 6 - 2\alpha + \beta \\
y &= \frac{6 - 2\alpha + \beta}{5} \\
x &= 3 - 2\frac{6 - 2\alpha + \beta}{5} = \frac{15 - 12 + 4\alpha - 2\beta}{5} \\
&= \frac{3 + 4\alpha - 2\beta}{5}.
\end{aligned}$$

The distance from $\langle \alpha, \beta \rangle$ to L is thus the square root of

$$\begin{aligned}
\left(\alpha - \frac{3 + 4\alpha - 2\beta}{5}\right)^2 + \left(\beta - \frac{6 - 2\alpha + \beta}{5}\right)^2 \\
&= \left(\frac{-3 + \alpha + 2\beta}{5}\right)^2 + \left(\frac{-6 + 2\alpha + 4\beta}{5}\right)^2 \\
&= \frac{\alpha^2 + 4\alpha\beta - 6\alpha + 4\beta^2 - 12\beta + 9}{5}. \tag{2.7}
\end{aligned}$$

And the square of the distance from $\langle \alpha, \beta \rangle$ to $\langle 1, 2 \rangle$ is

$$\begin{aligned}
 (\alpha - 1)^2 + (\beta - 2)^2 &= \alpha^2 - 2\alpha + 1 + \beta^2 - 4\beta + 4 \\
 &= \alpha^2 + \beta^2 - 2\alpha - 4\beta + 5.
 \end{aligned} \tag{2.8}$$

Combining (2.7) and (2.8) we see that the equation of the hyperbola in question is thus

$$x^2 + y^2 - 2x - 4y + 5 = 2^2 \left(\frac{x^2 + 4xy + 4y^2 - 6x - 12y + 9}{5} \right),$$

i.e.,

$$5x^2 + 5y^2 - 10x - 20y + 25 = 4x^2 + 16xy + 16y^2 - 24x - 48y + 36,$$

i.e.,

$$x^2 - 16xy - 11y^2 + 14x + 28y - 11 = 0.$$

In general every conic section will have an equation of the form

$$ax^2 + bxy + cy^2 + dx + ey + f = 0, \tag{2.9}$$

and, conversely, every Eq. (2.9) will define a (possibly degenerate) conic section. In my student days a goodly portion of an Analytic Geometry course was devoted to graphing conic sections and determining the type and basic parameters of the curve defined by (2.9) from the coefficients. The first step was to transform the equation into one with no mixed term,

$$Au^2 + Cv^2 + Du + Ev + F = 0, \tag{2.10}$$

by performing a substitution,

$$\begin{aligned}
 x &= u \cos \theta - v \sin \theta \\
 y &= u \sin \theta + v \cos \theta,
 \end{aligned}$$

where $\theta = 45^\circ$ if $a = c$ and $\tan 2\theta = \frac{2b}{a - c}$ otherwise. This represented a simple rotation of the xy -axes into a new pair of uv -axes. The type was then easily determined: if A and C had the same sign one had an ellipse; opposite signs meant the curve was a hyperbola; one of the two coefficients being 0 indicated a parabola; and both being 0 made for a straight line. The exact details are easily forgotten and books of mathematical tables and formulæ would include a table outlining the classification.

But we need not stop here.

If $A = C = 0$, (2.10) is the equation of a straight line and is not very interesting. If $A \neq 0$, but $C = 0$, we have the parabola

$$Au^2 + Du + Ev + F = 0,$$

which, if not degenerate (i.e., a line or pair of lines when $E = 0$), can be solved for v in terms of u ,

$$v = \frac{-Au^2 - Du - F}{E},$$

thus yielding the parametrisation,

$$\begin{aligned} u(t) &= t \\ v(t) &= \frac{-At^2 - Dt - F}{E}, \quad t \in (-\infty, \infty). \end{aligned}$$

And this yields the following parametric equations for the original curve,

$$\begin{aligned} x(t) &= u(t) \cos \theta - v(t) \sin \theta \\ y(t) &= u(t) \sin \theta + v(t) \cos \theta, \end{aligned}$$

for θ as before and $t \in (-\infty, \infty)$.

The case $A = 0$ and $C \neq 0$ is treated similarly.

In the elliptic and hyperbolic cases, when $A \neq 0$ and $C \neq 0$, one first makes the substitution,

$$\begin{aligned} u &= U + \frac{D}{2A} \\ v &= V + \frac{E}{2C} \end{aligned}$$

to complete the squares and transform (2.10) into

$$AU^2 + CV^2 = \frac{D^2}{4A^2} + \frac{E^2}{4C^2} - F,$$

i.e.,

$$AU^2 + CV^2 = G,$$

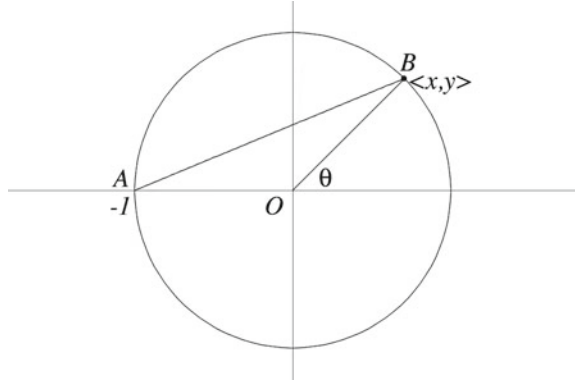
for some G . Taking A positive, another substitution,

$$\begin{aligned} U &= \sqrt{\left| \frac{G}{A} \right|} X \\ V &= \sqrt{\left| \frac{G}{C} \right|} Y, \end{aligned}$$

results in an equation in one of the forms,

$$X^2 + Y^2 = 1, \quad X^2 - Y^2 = 1, \quad X^2 - Y^2 = -1,$$

Fig. 2.6 Parametrisation of the circle



which define the unit circle, a standard left-right opening hyperbola, and a standard up-down opening hyperbola, respectively. These are easily parametrised.

The parametrisation of the unit circle most familiar in the literature is trigonometric:

$$\begin{aligned} x(\theta) &= \cos \theta \\ y(\theta) &= \sin \theta \end{aligned}, \quad \theta \in [0, 2\pi].$$

Some Calculus texts express $\cos \theta$ and $\sin \theta$ in terms of $t = \tan(\theta/2)$:

$$\begin{aligned} x(t) &= \frac{1 - t^2}{1 + t^2}, \quad t \in (-\infty, \infty] \text{ or } [-\infty, \infty). \\ y(t) &= \frac{2t}{1 + t^2} \end{aligned}$$

The second of these parametrisations is readily established. Consider Fig. 2.6. The angle $\angle BAO$ is half the angle θ and its tangent t is

$$t = \frac{y}{1 + x}.$$

This makes

$$y = t(1 + x). \quad (2.11)$$

Combining this with the equation $x^2 + y^2 = 1$ of the unit circle successively yields

$$\begin{aligned} x^2 + t^2(1 + x)^2 &= 1 \\ (1 + t^2)x^2 + 2t^2x + t^2 - 1 &= 0. \end{aligned}$$

The solution to the quadratic equation yields

$$\begin{aligned}
x &= \frac{-2t^2 \pm \sqrt{4t^4 - 4(t^2 + 1)(t^2 - 1)}}{2(1 + t^2)} \\
&= \frac{-2t^2 \pm \sqrt{4t^4 - 4(t^4 - 1)}}{2(1 + t^2)} \\
&= \frac{-2t^2 \pm \sqrt{4}}{2(1 + t^2)} \\
&= \frac{-t^2 \pm 1}{1 + t^2} = -1, \frac{1 - t^2}{1 + t^2}.
\end{aligned}$$

Now, $x = -1$ occurs only when $\theta = \pi$ and $t = \tan \theta/2$ is undefined. For other x we have

$$x = \frac{1 - t^2}{1 + t^2}.$$

If we now plug this value back into (2.11), we get

$$y = t(1 + x) = t \left(\frac{1 + t^2}{1 + t^2} + \frac{1 - t^2}{1 + t^2} \right) = t \frac{2}{1 + t^2} = \frac{2t}{1 + t^2},$$

as promised.

Thus, every point on the unit circle other than $(-1, 0)$ is given by

$$x(t) = \frac{1 - t^2}{1 + t^2}, \quad y(t) = \frac{2t}{1 + t^2} \quad (2.12)$$

for some $t \in (-\infty, \infty)$. Writing

$$x(t) = \frac{\frac{1}{t^2} - 1}{\frac{1}{t^2} + 1}, \quad y(t) = \frac{\frac{2}{t}}{\frac{1}{t^2} + 1},$$

and plugging $\pm\infty$ in for t yields $x(\pm\infty) = -1$, $y(\pm\infty) = 0$.

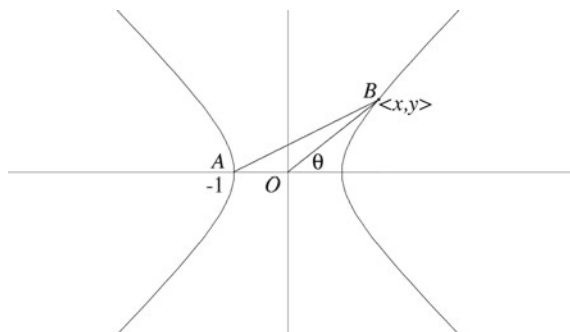
2.1.4 Exercise We have seen that every point (x, y) on the unit circle other than $(-1, 0)$ is of the form $(x(t), y(t))$ for $x(t), y(t)$ defined by (2.12). Complete the proof that these functions parametrise the circle for $t \in [-\infty, \infty)$ by showing the point $(x(t), y(t))$ to lie on the circle, i.e., show that

$$x(t)^2 + y(t)^2 = 1$$

for all $t \in (-\infty, \infty)$.

The hyperbola $x^2 - y^2 = 1$ is similarly parametrised. One starts with the analogous Fig. 2.7. Letting t be the tangent of the angle $\angle BAO$ which is no longer half the

Fig. 2.7 Parametrisation of the hyperbola



angle θ but is the slope of AB , we again have $t = y/(1 + x)$ and Eq. (2.11). If one now plugs $t(1 + x)$ into the equation $x^2 - y^2 = 1$ for the hyperbola, and performs the algebra, i.e., the simplification, one obtains

$$x(t) = \frac{1 + t^2}{1 - t^2}, \quad y = \frac{2t}{1 - t^2}. \quad (2.13)$$

2.1.5 Exercise Perform the algebraic derivation just described and show, for $t \neq \pm 1$, that $x(t)$, $y(t)$ defined by (2.13) do indeed satisfy

$$x(t)^2 - y(t)^2 = 1.$$

The domain of t is more complicated in the hyperbolic case than in the circular one. Consider first the right branch. The points on this branch are given by allowing t to range over the open interval $(-1, 1)$. Geometrically this is obvious because the slope of AB must lie between the slopes ± 1 of the asymptotes of the hyperbola. Algebraically we note that x is undefined, or infinite, for $t = \pm 1$, negative for $|t| > 1$, and positive for $t \in (-1, 1)$. For points B on the left branch, one must have $|t| > 1$, t negative for B on the upper half of the branch and positive for B on the lower portion. Once again, A corresponds to the choice $t = \pm\infty$.

2.1.6 Exercise One can explore this nicely on a graphing calculator, which, unlike a computer, is slow enough that one can see the curve as it is being drawn. On the TI-83 or TI-84 from Texas Instruments, I suggest setting the **MODE** to **Par**, entering

$$\begin{aligned} X_{1T} &= (1 + T^2)/(1 - T^2) \\ Y_{1T} &= 2T/(1 - T^2) \end{aligned}$$

in the equation editor, entering

$$\begin{aligned} T_{\min} &= -5 \\ T_{\max} &= 5 \\ T_{\text{step}} &= .1 \end{aligned}$$

in the **WINDOW** menu, and then choosing **ZDecimal** from the **ZOOM** menu. First the upper portion of the left branch will be drawn, starting just above the point $\langle -1, 0 \rangle$ and continuing in an upward-leftward direction for $t \in [-5, -1)$. Then the right branch, from lower right towards the centre and thence to the upper right, will be traced out as t covers the interval $(-1, 1)$. Finally, for $t \in (1, 5]$, the lower portion of the left branch will be drawn, proceeding from lower left toward the centre. (The drawing of the right branch being rather quick, one might prefer choosing $\text{Tstep}=.05$. One can then speed up the drawing of the left branch by choosing $\text{Tmin}=-4$ and $\text{Tmax}=4$. However, this does leave an even larger gap in the graph near $\langle -1, 0 \rangle$.)

I mentioned earlier that Menæchmus applied conic sections to duplicate the cube. This is actually quite simple. Let $\langle \alpha, \beta \rangle$ be the point of intersection of the two parabolas,

$$y = x^2 \quad (2.14)$$

$$2x = y^2. \quad (2.15)$$

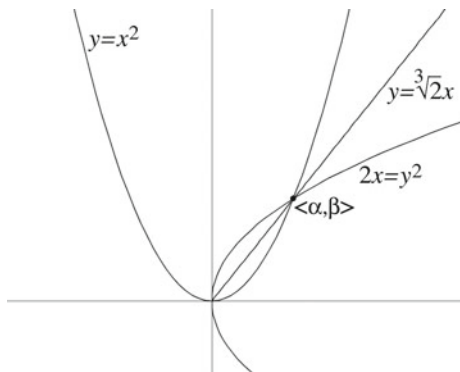
Then

$$\begin{aligned} \beta^3 &= \beta \cdot \beta^2 \\ &= \alpha^2 \cdot \beta^2, \text{ by (14)} \\ &= \alpha^2 \cdot 2\alpha, \text{ by (15)} \\ &= 2\alpha^3. \end{aligned}$$

The line connecting the origin with $\langle \alpha, \beta \rangle$ thus has slope $\beta/\alpha = \sqrt[3]{2}$ and choosing for x the length of the edge of any cube, the corresponding y will have a cube of twice the volume. See Fig. 2.8.

So all three problems — squaring the circle, trisecting the angle, and doubling the cube — were solved by the Greeks through the addition of new curves. Unlike the Chinese or Indian mathematicians who excelled in numerical methods, Greek

Fig. 2.8 Duplication of the cube



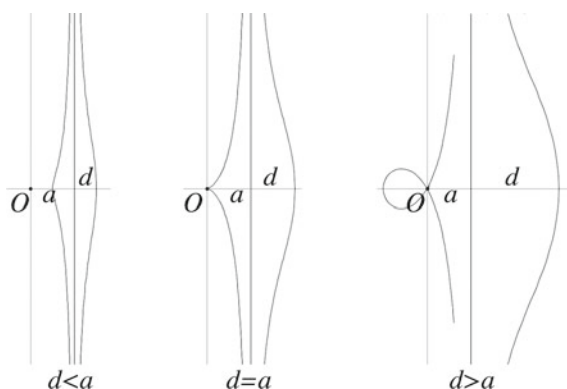
mathematicians were geometrically oriented. Curvilinear solutions were only good if one could graph the curves. None of the new curves could be graphed by ruler and compass alone, a fact that would only first be proven in the 19th century. The conic sections can be graphed mechanically using other tools, most crudely by slicing a cone. The string construction of the ellipse is known to every school child. And mechanical linkages can be constructed for all the conic sections. The same does not hold, however, for the quadratrix and we can say that, from a draughtsman's perspective, only the duplication of the cube had thus far been achieved.

The angle can similarly be trisected by appeal to conic sections, as would be done centuries later in Persia by 'Umar al-Khayyāmī (1048–1122). The reason for this is algebraic: the relation between the cosine of an angle and that of its tripled angle can be expressed algebraically as a cubic equation and al-Khayyāmī could solve cubic equations by intersecting conic sections. The Greeks were unfamiliar with this but they successfully solved the trisection problem by various other means. One solution can be had by means of the conchoid of Nicomedes, the next major curve to come along after the conic sections.

The conchoid, so named because of its resemblance to the curve of a conch shell, is defined kinematically. One starts with a line L and a point O not on the line. One takes another line L' anchored at O and rotates it around O . The locus of all points P at a fixed distance d from L as measured along L' gives the conchoid. It has two branches, one on either side of L . There are three types of conchoids determined by the relationship between d and the distance a of O from L . When $d < a$, the branch on the same side of L as O has a dip near O , but is smooth; when $d = a$, the dip reaches O in a cusp; and when $d > a$, the curve not only dips toward O , but passes through it as a loop. See Fig. 2.9.

Let us analyse the branch of the conchoid lying on the opposite side of L from O . To this end, draw two additional lines through O , one parallel to L to serve as the x -axis and one perpendicular to L to serve as the y -axis. The line L' is then completely determined by the angle θ it makes with the x -axis at the origin O . It intersects L in some point $P = \langle x, y \rangle$ for $0 < \theta < \pi$. (See Fig. 2.10, where I have rotated the

Fig. 2.9 Conchoids



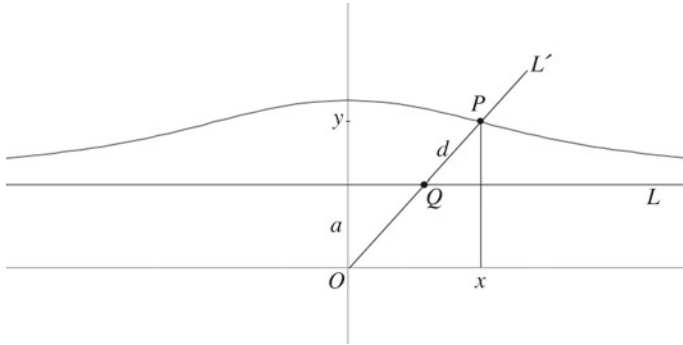


Fig. 2.10 Parametrisation of the conchoid

graph to better fit the allotted space.) From the figure we see that $y = a + d \sin \theta$. But $\tan \theta = y/x$, whence $x = (a + d \sin \theta)/\tan \theta$. Thus we have the parametric equation:

$$\begin{aligned} x(\theta) &= \frac{a + d \sin \theta}{\tan \theta}, \quad 0 < \theta < \pi. \\ y(\theta) &= a + d \sin \theta \end{aligned}$$

We can also find an algebraic parametrisation in terms of $t \in (-\infty, \infty)$ where $t = \cot \theta$. First rewrite

$$x = (a + d \sin \theta) \cot \theta = a \cot \theta + d \cos \theta = at + d \cos \theta. \quad (2.16)$$

Again, from $t = \cot \theta$, we have $t \sin \theta = \cos \theta$, and

$$1 = \sin^2 \theta + \cos^2 \theta = \sin^2 \theta + t^2 \sin^2 \theta = (1 + t^2) \sin^2 \theta,$$

i.e.,

$$\sin^2 \theta = \frac{1}{1 + t^2},$$

and

$$\sin \theta = \frac{1}{\sqrt{1 + t^2}}.$$

And

$$\begin{aligned} \cos \theta &= \sqrt{1 - \sin^2 \theta} = \sqrt{1 - \frac{1}{1 + t^2}} \\ &= \sqrt{\frac{1 + t^2 - 1}{1 + t^2}} = \frac{\pm t}{\sqrt{1 + t^2}}. \end{aligned}$$

Each choice of the plus or minus sign will yield a parametrisation of a branch of the conchoid. Recalling (2.16) and choosing the positive sign results in the parametrisation¹⁴

$$\begin{aligned}x(t) &= at + \frac{dt}{\sqrt{1+t^2}}, \quad -\infty < t < \infty, \\y(t) &= a + \frac{d}{\sqrt{1+t^2}}\end{aligned}$$

of the upper branch. The negative sign yields the corresponding parametrisation for the other branch:

$$\begin{aligned}x(t) &= at - \frac{dt}{\sqrt{1+t^2}}, \quad -\infty < t < \infty. \\y(t) &= a - \frac{d}{\sqrt{1+t^2}}\end{aligned}$$

2.1.7 Exercise As with the trigonometry-free parametrisation of the hyperbola, we can verify this on a graphing calculator. On the TI-83 or TI-84, in parametric graphing mode enter the functions

$$\begin{aligned}X_{1T} &= AT + \{1, -1\}DT/\sqrt{(1+T^2)} \\Y_{1T} &= A + \{1, -1\}D/\sqrt{(1+T^2)},\end{aligned}$$

with parameters A, D to be chosen later. Choose the window

$$\begin{aligned}T_{\min} &= -6.5 \\T_{\max} &= 6.5 \\T_{\text{step}} &= .1 \\X_{\min} &= -10 \\X_{\max} &= 10 \\Y_{\min} &= -3 \\Y_{\max} &= 5.\end{aligned}$$

Then graph the curves for the following A, D pairs:

- i. $A = 2, D = 1$
- ii. $A = 2, D = 2$
- iii. $A = 2, D = 3$.

In each case, the upper branch will be graphed first from left to right. Two brief periods of inactivity will occur before and after graphing this branch of the curve. This is because the points on this branch corresponding to values of T less than -3.9 and greater than 3.9 are offscreen. Following this, the lower branch of the curve will be drawn, again from left to right. [A final small warning: The axes are drawn to different scales and the usual calculator distortion will occur.]

The construction problems of antiquity are only remotely relevant to any discussion of the Mean Value Theorem, our central concern in this book; however, they have been a running thread throughout this section and I suppose I should comment on

¹⁴Apologies to the reader: dt here denotes multiplication by d , not the differential.

Fig. 2.11 Nicomedes's trisection

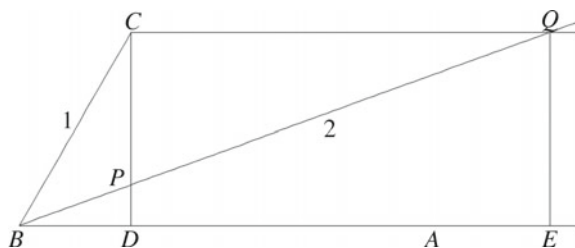
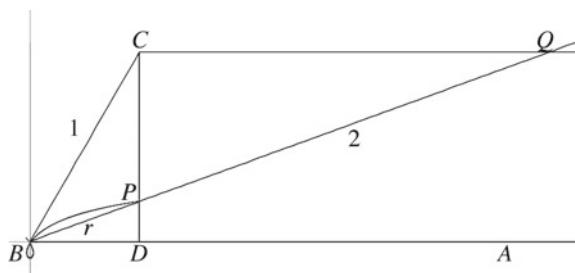


Fig. 2.12 Trisection via the conchoid



the use of the conchoid in trisecting the angle. The reader with no particular interest in the matter is invited to skip ahead to page 32. There is, however, a small paedagogical point illustrated by the construction that mirrors a criticism of the classroom presentation of the proof of the Mean Value Theorem. This is that the construction is given with no explanation for the choice of a crucial parameter.

Bunt, Jones, and Bedient¹⁵ begin their explanation with a diagram like Fig. 2.11. Their labelling is different and the perpendicular QE to AB isn't drawn, but overall their diagram agrees with this one. They explain that $\angle ABC$ is a given angle θ that is to be trisected. One performs the trisection by dropping a perpendicular CD to AB and drawing a line perpendicular to CD at C , i.e., a parallel to AB passing through C . If we choose BC as the unit, they suggest finding P as the intersection of the line CD and the conchoid given by choosing B as the pivot point O , CQ as the line L , $CD = \sin \theta$ as a (where we here use the geometric convention of writing XY for $\text{dist}(X, Y)$ for points X, Y), and finally 2 as d . Then $\angle ABP$ trisects $\angle ABC$. (See Fig. 2.12.)

Figure 2.12 is not the prettiest picture in the world, especially when graphed on the small screen of one's calculator, and some prefer to use the conchoid based on $O = B$, $L = CD$, $a = BD = \cos \theta$, and $d = 2$. Then Q is the point of intersection of the conchoid with the horizontal line passing through C .

2.1.8 Exercise I should include an illustration of the second conchoid for comparison with Fig. 2.12, but the image is so clear on the calculator and the graph can be redrawn for various choices of θ by using the variable θ on the calculator keyboard,

¹⁵Lucas N.H. Bunt, Phillip S. Jones, and Jack D. Bedient, *The Historical Roots of Elementary Mathematics*, Prentice-Hall, Inc., Englewood Cliffs (NJ), 1976, pp. 105–106.

so I choose instead to instruct the reader to do the diagram himself. Set the graphing mode to **Par** and enter the equation for the conchoid based on $a = \cos \theta$, $d = 2$ for a generic angle θ (noting that the rôles of x , y are reversed from those in our determination of the parametric equations for the conchoid):

$$\begin{aligned} X_{1T} &= \cos(\theta) + 2\sin(T) \\ Y_{1T} &= (\cos(\theta) + 2\sin(T))/\tan(T). \end{aligned}$$

Then enter the equations of the lines BC determining the angle, CQ determining Q , and that of the angle trisector:

$$\begin{aligned} X_{2T} &= T \\ Y_{2T} &= \tan(\theta)T \\ X_{3T} &= T \\ Y_{3T} &= \sin(\theta) \\ X_{4T} &= T \\ Y_{4T} &= \tan(\theta/3)T. \end{aligned}$$

In the **WINDOW** menu set

$$\begin{aligned} T_{\min} &= 0 \\ T_{\max} &= \pi \end{aligned}$$

to avoid drawing any of the other branch of the conchoid and the intrusive near vertical lines connecting the two branches. Then use **ZDecimal** to draw the graphs for a variety of angles including $\pi/2$, $\pi/3$, $\pi/4$, 0. For the angles $\pi/2$ and 0 you might want to enter the **FORMAT** menu and choose **AxesOff**. Note that the choice $\pi/3$ is the one of Fig. 2.12.

As I say, the graph is very nice, but one might like to zoom in on it a bit. The default zoom factor is 4, which is too large a zoom. One can use the **MEMORY** submenu accessed by the **ZOOM** button to access **SetFactors...** and set **XFact** and **YFact** equal to 2. Or one can enter parameters in the **WINDOW** menu. I found the following values worked well:

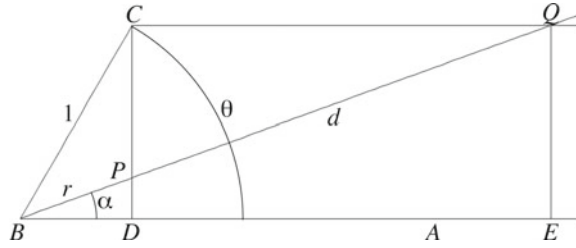
$$\begin{aligned} X_{\min} &= -1.3 \\ X_{\max} &= 3 \\ Y_{\min} &= -1.55 \\ Y_{\max} &= 1.55. \end{aligned}$$

The reader who has faithfully carried out this latest exercise has seen that the construction works, at least for acute angles.¹⁶ But he hasn't seen why the construction works or how the value 2 was chosen. We have seen that this choice works, but not how we knew to use 2 rather than, say, 3.

My favourite explanation is given by elaborating on the trigonometric proof cited by Coolidge,¹⁷ which yields the sought-after rationale. See Fig. 2.13.

¹⁶To the Greeks, angles were between 0° and 180° . As every obtuse angle is the sum of a right angle and an acute angle, and as the right angle is easily trisected, we need only concern ourselves here with acute angles.

¹⁷Julian Lowell Coolidge, *A History of Geometrical Methods*, Dover Publications, Inc., New York, 1963, pp. 46–47. This is a reprint of a volume originally published by Oxford University Press in 1940.

Fig. 2.13 Showing $d = 2$ 

First, note that if we are given $\theta = \angle ABC$ and another angle α we hope to be $\theta/3$, we can find the points P and Q where the line with angle α intersects CD and the horizontal line passing through C , respectively. If we let BC be the unit, we can ask for the value of $PQ = d$. To this end, let $r = BP$ and note that $BD = \cos \theta$, whence

$$\cos \alpha = \frac{BD}{BP} = \frac{\cos \theta}{r},$$

i.e.,

$$r = \frac{\cos \theta}{\cos \alpha}. \quad (2.17)$$

But, looking at triangle QBE ,

$$\sin \alpha = \frac{QE}{BQ} = \frac{CD}{d+r} = \frac{\sin \theta}{d+r},$$

whence

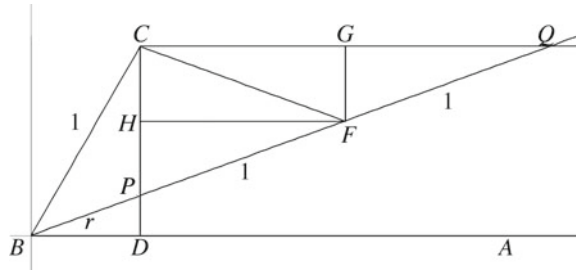
$$d+r = \frac{\sin \theta}{\sin \alpha} \quad (2.18)$$

Combining (2.17) and (2.18) we have

$$\begin{aligned} d &= d+r-r = \frac{\sin \theta}{\sin \alpha} - \frac{\cos \theta}{\cos \alpha} \\ d \sin \alpha \cos \alpha &= \sin \theta \cos \alpha - \sin \alpha \cos \theta \\ d \sin \alpha \cos \alpha &= \sin(\theta - \alpha). \end{aligned} \quad (2.19)$$

But $\theta = 3\alpha$ iff $\theta - \alpha = 2\alpha$ and, by (2.19),

Fig. 2.14 Geometric proof of trisection



$$\sin 2\alpha = \sin(\theta - \alpha) = d \sin \alpha \cos \alpha \text{ iff } d = 2.$$

Thus, to trisect the angle we must have $d = 2$. And, conversely, if we choose $d = 2$, the construction trisects the angle.

The trigonometric solution would not have been directly available to Nicomedes. The Greeks used chords rather than sines and cosines, which were later introduced by Indian mathematicians, and the trigonometric addition formulæ were, I believe, first proven some centuries after Nicomedes by Claudius Ptolemy (c. 85–c. 165). A more traditionally geometric proof that $\angle ABQ$ trisects $\angle ABC$, given the assumption that $PQ = 2$, proceeds as follows: Bisect PQ at F . From F drop perpendiculars to CD and CQ as in Fig. 2.14. Triangles PHF and FGQ are similar and $PF = FQ = 1$, whence they are congruent. Thus, $CG = HF = GQ$ and triangles CGF and QGF are congruent, sharing as they do two pairs of equal sides and equal right angles between them. Thus $\angle GCF = \angle GQF$, and $\angle GQF$ equals $\alpha = \angle ABQ$. We have

$$\angle CFP = \pi - \angle CFQ = \pi - (\pi - 2\alpha) = 2\alpha.$$

But $BC = 1 = CF$, whence BCF is an isosceles triangle and we have $\theta - \alpha = \angle CBF = \angle CFP = 2\alpha$, i.e., $\theta = 3\alpha$. Thus, $\angle ABQ$ is indeed the trisector of $\angle ABC$.

2.1.9 Remark I find adding the lines HF and GF to clutter up the diagram unnecessarily. To conclude that $CF = QF$, note that

$$\cos \alpha = \frac{CQ}{PQ} = \frac{CQ}{d},$$

i.e., $CQ = d \cos \alpha$. If we now apply the *Law of Cosines* to the triangle CQF and $\angle CQF = \alpha$, we get

$$\begin{aligned}
CF^2 &= CQ^2 + QF^2 - 2 \cdot CQ \cdot QF \cdot \cos \alpha \\
&= d^2 \cos^2 \alpha + \frac{d^2}{4} - 2 \cdot d \cos \alpha \cdot \frac{d}{2} \cdot \cos \alpha \\
&= d^2 \cos^2 \alpha + \frac{d^2}{4} - d^2 \cos^2 \alpha = \frac{d^2}{4},
\end{aligned}$$

whence $CF = \frac{d}{2} = QF$.

The next step of showing $\angle CFB = 2\alpha$ can be accomplished as before, assuming $d/2 = 1 = BC$, i.e., $d = 2$, or we can appeal to another result in Euclid¹⁸: If the same chord in a circle is subtended both by an angle with vertex at the centre and a vertex on the same side of the chord as the centre and lying on the circumference, then the former angle is twice the latter. In this case, one takes the circle centred at F of radius $d/2$, lets PC be the chord and Q the second vertex. One automatically has

$$\angle CFP = 2\angle CQF = 2\alpha.$$

One completes the proof by noting

$$\angle CBF = \angle CFB = \angle CFP \text{ iff } BC = CF \text{ iff } 1 = \frac{d}{2},$$

i.e., iff $d = 2$. Thus again we see that the choice of $d = 2$ does lead to the conclusion that $\angle ABQ = \alpha$ trisects $\angle ABC = \theta$. And we see again where the choice of $d = 2$ came from.

2.1.10 Remark The Law of Cosines is an important identity and pops up in Vector Analysis. Somewhat less generally important, but useful here, is the *Law of Sines*. Applied to triangle BCQ , not assuming a specific value for $d = PC$, it yields

$$\frac{\sin \alpha}{BC} = \frac{\sin(\theta - \alpha)}{CQ},$$

i.e.,

$$\frac{\sin \alpha}{1} = \frac{\sin(\theta - \alpha)}{d \cos \alpha}.$$

Thus $\sin(\theta - \alpha) = d \sin \alpha \cos \alpha$ and we conclude $\theta - \alpha = 2\alpha$ iff $d = 2$, i.e., the construction trisects $\angle ABC$ iff $d = 2$.

I've overindulged myself in presenting all these alternatives. But I think it is important here to stress that the choice of $d = 2$ is forced upon us by a simple

¹⁸The Law of Cosines, in what we might call a disguised form, appears as Propositions 12 and 13 in Book II of the *Elements*. To make this proof non-trigonometric and purely geometric requires merely a change in terminology.

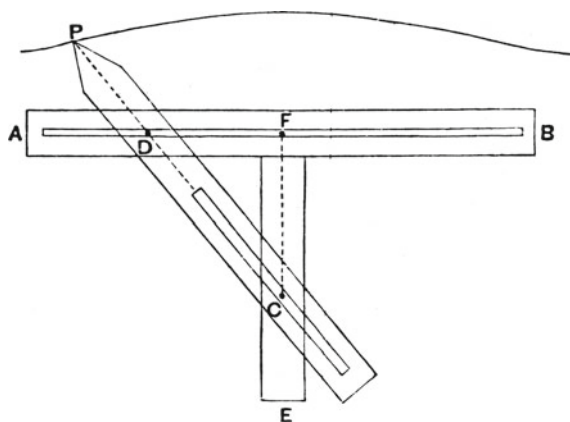
consideration and is not an act of inspiration or omniscience on the part of the presenter. The construction, which appears to come from nowhere, is actually quite natural. Some things like the drawing of Fig. 2.11 are fairly automatic. One would begin an analysis of the problem by drawing an angle $\angle ABC$ and its trisector $\angle ABQ$ as presented there. Dropping the perpendicular CD is a fairly natural thing to do in one's exploration, as is choosing $BC = 1$: we are dealing with angles and it is natural to place the vertex B in the centre of a circle of radius 1. Once this is done, the problem is to find P . P is determined by any of the distances BP , CP , and DP . At some stage one may draw the parallel CQ , perhaps to have a right triangle of altitude CD opposite $\angle ABP$. One then realises that PQ can also be used to determine P . If one knows about the conchoid, one now merely has to choose the right d and verify that it works. The most mysterious part of the presentation is the often unexplained choice of $d = 2$.

An analogous situation arises in the classroom proof of the Mean Value Theorem in which an auxiliary function is used. One of the criticisms levelled against this proof, which we will encounter in the next chapter, is the lack of motivating explanation behind the choice of this function. As with the choice here of $d = 2$, the choice there can be explained. The lack of explanation in a textbook or in a lecture speaks only of the laziness of the expositor, and not of the mysteriousness of the proof. This is not to say that such laziness is always bad: the expositor may choose not to explain such things if his intended readers or classroom students have sufficient background and ability to work out the details for themselves, or if the point is too minor to justify the necessary page count or classroom time. Is this the case here regarding the choice of d ? It seems not to be the case with the auxiliary function used in the proof of the Mean Value Theorem in the standard Calculus course.

But this is a matter for consideration later. For now we have to finish up with the conchoid, discuss two additional curves, and then give a tentative formal definition of a curve.

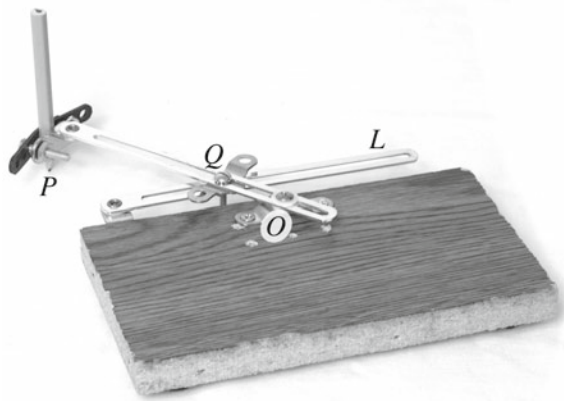
I don't have too much more to say about the conchoid. In his introductory essay on the work of Archimedes and his predecessors, Heath¹⁹ informs us that, according to Pappus, Nicomedes introduced the conchoid for the purpose of duplicating the cube. He also says that Pappus and another commentator Eutocius tell us that Nicomedes also constructed a mechanical instrument for use in drawing the conchoid. Such a device is illustrated in figure above, below, on the next page.

¹⁹T.L. Heath, *The Works of Archimedes Edited in Modern Notation with Introductory Chapters by T.L. Heath with a Supplement The Method of Archimedes Recently Discovered by Heiberg*, Dover Publications, Inc., New York, no date given. Heath's original edition was published in 1897 by Cambridge University Press, the supplement appearing subsequently in 1912. Cf. pp. cvi–cvii for his remarks on the conchoid.



Conchoidograph

Pictured above is Heath's drawing of a conchoidograph and below is a working model based on it that I made from scrap material found around the house. The point O is fixed and represents the pivot point of the conchoid (i.e., the point O of Fig. 2.10). The slider labelled L represents the line L of that figure. The pencil pointing to P represents the point P on the conchoid and Q the point on the line L . The screw at Q can be temporarily loosened to allow adjustment of the distance from P to Q ; when tightened PQ is fixed and Q is only allowed to move along L . With respect to my skill in constructing such, I must admit that my pencil holder is a bit wobbly and that the line L being fastened at only one end has a tendency to change its orientation while one is drawing the curve. But, when operated with three hands, it works quite well . . . Presumably a trip to the local hardware store will remedy these defects.



The conic sections can also be drawn by means of simple mechanical devices called linkages. Such devices show that these solutions to the duplication and trisection problems were genuine solutions, albeit solutions involving more than ruler and compass. The same cannot be said of the purely theoretical solutions to the trisection and quadrature problems afforded by the quadratrix. This solution shows that these problems can be solved *if* one can draw the quadratrix. It provides a reformulation of the problem rather than a solution. And, indeed, it can be shown that no similar linkage exists for drawing the quadratrix.

The same is true of another famous curve that can be used to trisect the angle and square the circle. This is the *spiral of Archimedes*. Exactly why Archimedes was drawn to the spiral is unclear. His work, *On Spirals*, is extant²⁰ and no explanation of his interest in spirals is given. The work is prefaced by a letter to a colleague named Dositheus, more than half of which summarises work not discussed in the book he is sending:

After these came the following propositions about the *spiral*, which are as it were another sort of problem having nothing in common with the foregoing; and I have written out the proofs of them for you in this book. They are as follows. If a straight line of which one extremity remains fixed be made to revolve at a uniform rate in a plane until it returns to the position from which it started, and if, at the same time as the straight line revolves, a point move at a uniform rate along the straight line, starting from the fixed extremity, the point will describe a spiral in the plane. I say then that the area bounded by the spiral and the straight line which has returned to the position from which it started is a third part of the circle described with the fixed point as centre and with radius the length traversed by the point along the straight line during the one revolution. And, if a straight line touch the spiral at the extreme end of the spiral, and another straight line be drawn at right angles to the line which has revolved and resumed its position from the fixed extremity of it, so as to meet the tangent, I say that the straight line so drawn to meet it is equal to the circumference of the circle.²¹

He cites a few more results before beginning the actual work of the book, but the two just cited are impressive enough.

Today, with Analytic Geometry and Calculus, these results are actually quite easy. First, referring to Fig. 2.15, one expresses the curve parametrically in terms of time t . Let P be the moving point, and assume the line segment L to be rotated is on the x -axis, with the fixed extremity at O and P initially coinciding with O . As P moves away from O along L , L is rotating around O , crossing the x -axis at D . Thus the motion of P is a composite of two motions, both assumed to be uniform. Thus there are constants a, b representing these uniform rates so that, at time t , the position of P in polar coordinates is given by

$$\theta = at, \quad r = bt.$$

Solving for t in terms of θ , we have $r = b\theta/a$. Thus, the spiral of Archimedes can be expressed in polar coordinates as

²⁰An English translation can be found in Heath's book cited in the preceding footnote. The work *On Spirals* occupies pp. 151–188.

²¹*Ibid.*, pp. 153–154.

Fig. 2.15 Spiral of Archimedes

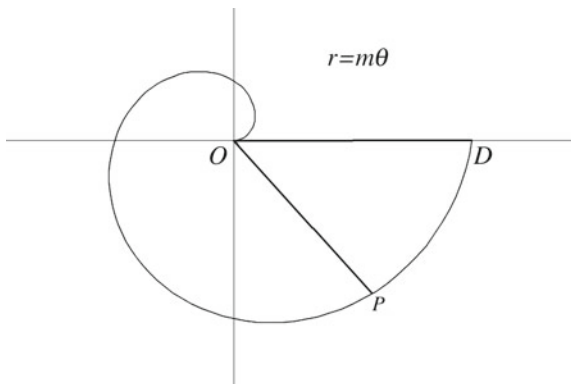
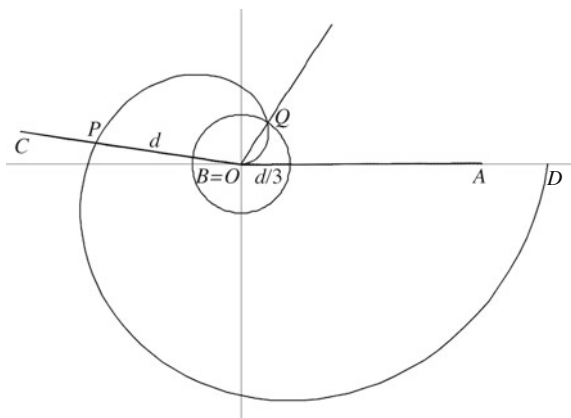


Fig. 2.16 Trisection via the spiral



$$r = \rho(\theta) = m\theta, \text{ for some } m.$$

Figure 2.15 shows the graph of this spiral for $0 \leq \theta \leq 2\pi$.

One can also express x and y parametrically in terms of θ :

$$\begin{aligned} x &= r \cos \theta = m\theta \cos \theta \\ y &= r \sin \theta = m\theta \sin \theta \end{aligned}, \quad 0 \leq \theta \leq 2\pi. \quad (2.20)$$

From the polar equation, for example, we can see immediately how to trisect an angle once the spiral is known. Let $\alpha = \angle ABC$ be given. Lay AB along OD , with B at O and find the point P where BC intersects the spiral. This has a certain length d , which is easily trisected using ruler and compass. Draw the circle of radius $d/3$ centred at $O = B$ and find the point Q where this circle intersects the spiral. Since $OQ = \frac{1}{3}OP$, we also have $\angle AOQ = \frac{1}{3}\angle AOP$. See Fig. 2.16.

By the same method any angle can be divided into any number of equal parts: the spiral solves the general multisection problem.

The first result cited by Archimedes in the preface to his book concerns the area of the region with boundary given by the spiral as θ ranges from 0 to 2π and the line OD as in Fig. 2.15. His claim is that this area is one third the area of the circle of radius OD . Today this is an easy calculation accessible to any student of the Calculus who has got as far as finding areas of polar curves by integration: For $r = \rho(\theta) = m\theta$, the area of the given region is

$$\begin{aligned} \text{Area} &= \int_0^{2\pi} \pi \cdot \rho(\theta)^2 \frac{d\theta}{2\pi} = \int_0^{2\pi} \frac{(m\theta)^2}{2} d\theta \\ &= \frac{m^2}{2} \cdot \frac{\theta^3}{3} \Big|_0^{2\pi} = \frac{m^2(2\pi)^3}{2 \cdot 3} = \frac{4m^2\pi^3}{3}, \end{aligned}$$

while the area of the circle with radius $OD = \rho(2\pi) = 2m\pi$ is $\pi \cdot (2m\pi)^2 = 4m^2\pi^3$.

Of greater interest here is the second property of the spiral cited by Archimedes in his letter to Dositheus. According to it, if one draws the tangent to the spiral at D in Fig. 2.15, it will meet the y -axis at a point E of distance $2\pi \cdot OD = 2\pi \cdot m2\pi = 4m\pi^2$ from the origin. Again, this is easy with modern Calculus. The slope of the tangent is

$$\frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta} = \frac{m \sin \theta + m\theta \cos \theta}{m \cos \theta - m\theta \sin \theta},$$

where we use the parametric representation (2.20). At $\theta = 2\pi$ this equals

$$\frac{m \cdot 0 + m \cdot 2\pi \cdot 1}{m \cdot 1 - m \cdot 2\pi \cdot 0} = \frac{2m\pi}{m} = 2\pi.$$

The equation of the tangent is thus

$$\frac{y - 0}{x - 2m\pi} = 2\pi,$$

i.e., $y = 2\pi x - 2\pi \cdot 2m\pi$, and the y -intercept E is given by $y = -2\pi \cdot 2m\pi = -2\pi \cdot m \cdot 2\pi$, whence $OE = 2\pi \cdot m2\pi = 2\pi \cdot OD$, the circumference of the circle of radius OD . From this the quadrature of the circle is an easy exercise.

The spiral of Archimedes can be continued by allowing $\theta > 2\pi$. If one does this, the curve spirals outward ever more, but at a constant rate of movement away from the origin. The radial distance between successive passes of the curve remains a constant $2m\pi$. Allowing θ to be negative produces a mirror image of the original curve reflected across the y -axis. The full curve thus intersects itself infinitely often in a rather attractive pattern, as the reader can see in Fig. 2.17.

Another famous spiral discovered some centuries later, the *logarithmic spiral* has completely different properties. Its definition in polar coordinates is

$$r = \rho(\theta) = ae^{b\theta}, \quad \theta \in (-\infty, \infty),$$

where a, b are positive constants and e is the base of the natural logarithms. As θ assumes larger positive values, its radial growth is unbounded, the radial distance between passes increasing without bound. At $\theta = 0$, the curve does not begin at the origin; in fact, as θ assumes larger and larger negative values, the curve spirals in towards the origin at an ever decreasing rate. See Fig. 2.18.

Our final curve for consideration is the *cycloid*. Boyer introduces the cycloid as follows:

However, it is reported that the imaginative Nicholas of Cusa (1401–1464) had noted the curve traced out by a point on the rim of a cart wheel as the wheel rolled along the road. Although he seems to have been unable to determine its nature or properties, this observation constituted a significant step in the study of curves, for it seems to represent the first modern instance in which a new curve was suggested by natural phenomena. The ancients had invented new curves *ad hoc* to solve specific geometrical problems: they had not discovered

Fig. 2.17 Full archimedean spiral

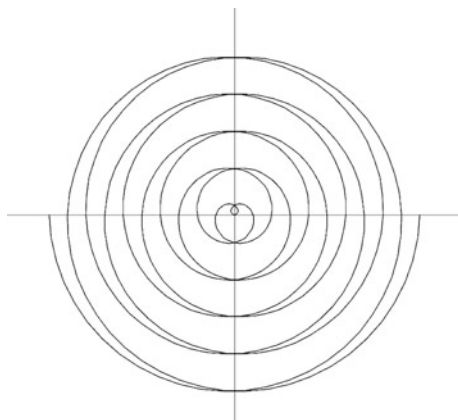
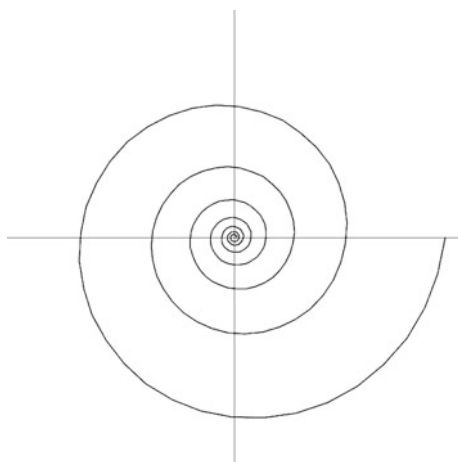


Fig. 2.18 Logarithmic spiral



these, except for the line and the circle, in the world of nature. The new curve of Cusanus²² was followed two centuries later by other curves which were disclosed by, and useful in the study of physical science.²³

The cycloid has its own mini-history, having been studied by a number of excellent mathematicians over the ensuing centuries. Apparently, Nikolaus von Kues played no rôle in it aside from a later misinterpretation on the part of John Wallis (1616–1703) of one of the diagrams in the Oxford manuscript of the work *De mathematicis complementis* of Kues.²⁴ A more accurate introduction to the cycloid and its history reads

The seventeenth century is one of the most exciting periods in the history of mathematics. The first half of the century saw the invention of analytic geometry and the discovery of new methods for finding tangents, areas, and volumes. These results set the stage for the development of the calculus during the second half. One curve played a central role in this drama and was used by nearly every mathematician of the time as an example for demonstrating new techniques. That curve was the cycloid.

The cycloid is the curve traced out by a point on the circumference of a circle, called the *generating circle*, which rolls along a straight line without slipping. It has been called the “Helen of Geometry,” not just because of its many beautiful properties but also for the conflicts it engendered...

The earliest mention of a curve generated by a point on a moving circle appears in 1501, when Charles de Bouvelles used such a curve in his mechanical solution to the problem of squaring the circle.²⁵

The history of the cycloid is a matter of some interest, but this interest is tangential to our purpose here. The reader curious about its central rôle in the development of the Calculus, the controversies it engendered (criticisms of proofs, and priority conflicts), and the fascinating physical properties of the curve is referred to the paper of John Martin from which the above quote has been taken.²⁶

What is relevant here, aside from its introduction of the physical world as a source of curves, is that it is, with our modern algebraic notation, easy to obtain a parametric representation of the curve and therewith a means to divine its properties.

The cycloid is depicted in Fig. 2.19. Here, the point P is assumed to coincide with the origin at the time $t = 0$, but remains fixed relative to the circle as it rolls along at

²²Nikolaus von Kues is often cited under variants of his name. The Latin form is Nicolaus Cusanus, though Cusanus often suffices. Other variants are Nikolaus von Cusa, Nicholas of Cusa, or simply Nicholas Cusa.

²³Boyer, *op. cit.*, p. 72.

²⁴Nikolaus von Kues, *Die mathematischen Schriften*, 2nd. edition, Verlag von Felix Meiner, Hamburg, 1979, p. 220. The volume contains translations of Kues's manuscripts from the Latin by Josepha Hofmann and an introduction and notes by Joseph Ehrenfried Hofmann. Footnote 37 on page 217 includes the remark, “The figure contained in the *Oxford* manuscript has led WALLIS to the rash claim that CUSANUS had already arrived at the construction of the cycloid”.

²⁵John Martin, “The Helen of Geometry”, *The College Mathematics Journal* 41, no. 1 (2010), pp. 17–28; here: p. 17.

²⁶I also suggest V. Frederick Rickey, “Build a brachistochrone and captivate your class” in: Amy Shell Gellasch (ed.), *Hands on History. A Resource for Teaching Mathematics*, Mathematical Association of America, 2007.

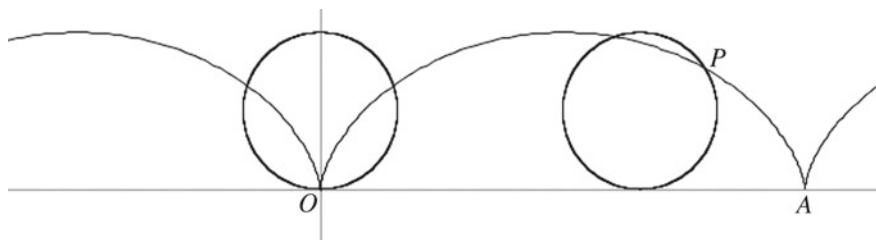


Fig. 2.19 Simple cycloid

a constant rate. If we affixed a movie camera to a dolly and filmed the wheel as the camera was moved along with the wheel, the film would show the point P moving around the circumference of a circle in a clockwise direction. Thus, if we imagine the axes moving too, the position of P would be described by

$$x = r \cos(\theta), \quad y = r + r \sin \theta,$$

where r is the radius of the circle and θ is the angle at which P sits on the circle as measured from its centre. For simplicity's sake we can take $r = 1$ and assume the rate of rotation is 1 radian/second. Thus in terms of time t , we have²⁷ $\theta = -\frac{\pi}{2} - t$ and

$$x = \cos\left(-\frac{\pi}{2} - t\right) = -\sin t, \quad y = 1 + \sin\left(-\frac{\pi}{2} - t\right) = 1 - \cos t.$$

Now, viewed from a stationary position, the vertical position is unchanged; thus we still have $y = 1 - \cos t$. But, horizontally, the circle itself has moved t radians: $x = t - \sin t$. The parametric equations of the cycloid are thus

$$\begin{aligned} x &= t - \sin t \\ y &= 1 - \cos t, \quad t \in (-\infty, \infty). \end{aligned}$$

It is not relevant to our purposes here, but the cycloid, like the conchoid, has a couple of variants if P is not on the rim but is elsewhere on the radius of the circle. If P lies inside the circle there is a dip in place of the cusp, and if P is outside the circle there is a loop. There are also *epicycloids* where we imagine a wheel rolling not in a straight line but around the exterior of the circle, and *hypocycloids* obtained when the wheel rolls around inside a circle. I suppose one could roll it along spirals as well and see what develops. It is a good subject for experimentation with one's graphing software or graphing calculator.

2.1.11 Exercise Had the Greeks been aware of the cycloid, would they have accepted it as a mechanical curve that effectively squares the circle? Presumably the distance OA between the places where P touches the x -axis equals the circumference of

²⁷ t has a minus sign because the clockwise rotation is the reverse of the usual rotation.

the circle. However, they might have had some difficulty with this because of *Aristotle's Wheel*, a geometric paradox dubiously ascribed to Aristotle. One imagines two wheels, a larger and a smaller, rigidly fixed to each other at the hub. After a complete revolution, points on their rims have traced out lines of equal distance. (See Fig. 2.20.) If the radii of the wheels are r and R , respectively, have the wheels travelled a distance of $2\pi r$ or $2\pi R$? The ratio $P'P'$ to PP should be R/r , but the distances are clearly equal. How is one to explain this?

This paradox puzzled scholars for centuries before Galileo (1564–1642) accounted for the discrepancy. Not everyone would accept Galileo's explanation today²⁸ and one refers blithely to “slippage”. What is happening is clearest if one imagines the small wheel rolling on a rail and carrying the large wheel with it as it turns. Assume for convenience that $r = 1$, $R = 2$ as in Fig. 2.20 and graph the paths of P and P' on your calculator as t goes from $-\pi$ to π . In the equation editor enter

$$\begin{aligned} X_{1T} &= T - \sin(T) \\ Y_{1T} &= 1 - \cos(T) \\ X_{2T} &= T - 2\sin(T) \\ Y_{2T} &= 1 - 2\cos(T) . \end{aligned}$$

In the WINDOW screen enter

$$\begin{aligned} T_{\min} &= -\pi \\ T_{\max} &= \pi \end{aligned}$$

and graph the functions using ZDecimal. You will see the paths of P and P' between the times when they are at the high points of the wheel. Notice that P travels only from left-to-right in the x -direction, while P' does some backtracking. If one does the calculation of the total horizontal movement of P' without regard for direction, one will find P' has actually travelled $10\pi/3$, still $2\pi/3$ short of the expected 4π , but at least one sees some of the discrepancy simply explained.

It is time we reconsider the problem of defining what a curve is. With our modern knowledge of algebraic notation, we spot immediately that all the curves cited have one thing in common: they all have parametric definitions with the parameter ranging over some interval or intervals. The Greeks had no such symbolism and resorted either to vague descriptions of what their curves had in common, or classified them according to their obvious differences.

The Euclidean definition of a curve as “breadthless length” and the Aristotelian definition as “magnitude extended in one direction”, both nods to the one-dimensionality expected of a curve, are too vague to serve as actual definitions. To do anything with these, one needs to define the term “dimension”, which would only be done adequately in the 20th century by L.E.J. Brouwer, P.S. Urysohn (1898–1924), and eventually Karl Menger (1902–1985). Aristotle's definition of a curve as the “flowing of a point” is more promising. It suggests to the modern mind

²⁸Cf., e.g., my exposition: Craig Smoryński, *Adventures in Formalism*, College Publications, London, 2012, pp. 99–104.

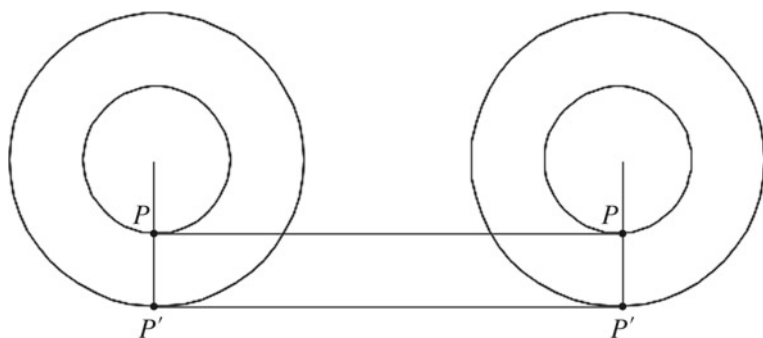


Fig. 2.20 Aristotle's wheel

a parametrisation of the position of the point P as a function of the time t from some interval during which the point is doing its flowing.

Heath's discussion of these definitions in his annotated edition of the *Elements* was followed by a discussion of two classifications of curves by Geminus (c. 10 B.C. – c. 60 A.D.). The first of these is very crude: Lines (i.e., curves) can be composite (broken lines forming angles) or incomposite; the latter can then form a figure (circle, ellipse) or not form a figure (straight line, parabola, hyperbola, conchoid). His second classification was more elaborate, but still crude.²⁹

Boyer cites a better classification:

The classical Greek geometers divided curves into three ranks or orders: the highest place was reserved only for the perfect curves, the line and the circle. These were called plane loci. Second place was granted to the Menaechmian conics which, probably on account of their original mode of definition, were known as solid loci. All other curves, whether algebraic or transcendental, were grouped together under the heading linear loci. Pappus described this last category as made up of those curves “the origin of which is more complicated and less natural [than that of the plane and solid loci], as they are generated from more irregular surfaces and intricate movements.” In this description we see the two types of curve definition which the Greeks recognized — the kinematic and the stereometric.³⁰

Since the Greeks individual curves could be defined in any of three ways:

- (1) kinematically, as the “flowing of a point” — as we've defined the conchoid, the spiral of Archimedes, and the cycloid;
- (2) as loci — as we've defined the individual conic sections;
- (3) stereometrically, i.e., in terms of solids — as the conic sections were originally defined.

It is not clear what the curves given rise to by these three modes of definition have in common. The kinematic approach is nowadays naturally formulated in terms of parametrisation. The various loci we've encountered were naturally supplied with

²⁹Heath, *Elements*, *op. cit.*, pp. 160–165.

³⁰Boyer, *op. cit.*, p. 32. The bracketed insertion is Boyer's.

parametrisations in part because they were described as loci of points satisfying certain distance requirements easily expressed equationally. But can one easily isolate those modes of definition of loci the solutions to which yield curves? And, although once given a curve in the plane one can readily concoct a solid, one of the edges of which happens to be that curve, using stereometry to define the notion of a curve presupposes a definition of a solid and its surface or of a solid and the boundary of a region obtained by intersecting the solid with a plane. And defining these concepts entails the same difficulties as defining a curve, but in higher dimensions.

So one's best bet for a usable formal definition of a curve to replace the vague informal conception seems to be the kinematic one encapsulated by a parametrisation of the position of a point P in terms of time t taken over an interval or intervals. Before making this our official choice, however, note that we have seen another way of defining a curve, namely another algebraic method that became available in the 17th century: With Eq. (2.9) we noted that every conic section was the locus of points satisfying a quadratic equation

$$f(x, y) = 0.$$

The ability to define curves as solution sets to equations was ushered in by the near simultaneous invention of Analytic Geometry by Fermat and Descartes. Fermat applied the algebraic symbolism of François Viète (1540–1603) to classical locus problems, deriving equations and declaring that such equations described curves. Boyer waxes eloquently on Fermat's statement that the equations described curves:

This brief sentence represents one of the most significant statements in the history of mathematics. It introduces not only analytic geometry, but also the immensely useful idea of an algebraic variable. The vowels in Viète's terminology previously had represented unknown, but nevertheless fixed or determinate, magnitudes. Fermat's point of view gave meaning to indeterminate equations in two unknowns — which previously had been rejected in geometry — by permitting one of the vowels to take on successive line-values, measured along a given axis from an initial point, the corresponding lines representing the other vowel, as determined by the given equation, being erected as ordinates at a given angle to the axis.³¹ In ancient Greek works, certain lines associated with a given curve had played a role equivalent to that of a coordinate system, and the properties of the curve had been expressed in terms of these lines by means of rhetorical algebra.³² The curve came first, the lines were then superimposed upon it, and finally the verbal description (or algebraic equation) was derived from the geometrical properties of the curve. Fermat's genius made it possible to reverse this situation. *Beginning* with an algebraic equation, he showed how this equation could be regarded as defining a locus of points — a curve — with respect to a given coordinate

³¹ A brief word of explanation: For some time mathematicians viewed curves as the paths traced out by the intersection of two lines, eventually a vertical line moving along the x -axis and a horizontal one moving up and down the y -axis. With Fermat, however, the axes were not necessarily perpendicular but met at a given angle. The variables thus stood for the positions of the lines parallel to these axes. Viète had begun a short-lived practice of using vowels to denote variables and consonants to denote unspecified constants and Fermat adhered to this tradition.

³² Mathematical historians distinguish 3 phases in the development of algebraic symbolism: *rhetorical*, in which everything is expressed in words; *syncopated*, in which some abbreviations are introduced; and *symbolic*, in which everything is expressed in abstract symbols and calculations follow strict term rewriting rules.

system. Fermat did not invent coordinates and he was not the first one to use graphical representation. Analytic reasoning had long been used in mathematics, and the application of algebra to geometry had become a commonplace. However, there appears to have been no appreciation before the times of Fermat and Descartes of the fact that, in general, a given algebraic equation in two unknown quantities determines, *per se*, a unique geometric curve. The recognition of this principle, together with its use as a formalized algorithmic procedure, constituted the decisive contribution of these two men.³³

Fermat went so far as to demonstrate that quadratic equations yielded exactly the plane and solid loci, i.e., the conic sections, but did not consider general algebraic equations that arise when f is a polynomial in two variables of arbitrary degree. Descartes did. Boyer summarises Descartes's approach as follows:

Whereas Viète had been interested in the constructibility of determinate problems, Descartes went further and applied the criteria to loci as well. It was here that he found it necessary to use a coordinate system. One may say that, in a general sense, the invention of analytic geometry by Descartes consisted in the extension of the analytic art of Viète³⁴ to the construction of *indeterminate* equations, just as in the case of Fermat it was the study of loci, by the analytic art, which led to the same result. But Descartes continued to regard the construction of *determinate* equations as his ultimate purpose.

The plotting of curves in the now customary manner was not a part of Cartesian analytic geometry. Even the Pappus loci are not sketched. Descartes knew that an equation in two unknowns determines a curve, but oddly enough, he seems not to have regarded such an equation as an adequate definition of the curve, and felt constrained to exhibit an actual mechanical construction in each case. It has been conjectured that the ancient Greeks stressed constructions because these served as existence theorems. One is tempted to apply this idea to Descartes and say that he doubted the existence of a curve corresponding to an equation unless he could supply a kinematic construction for it. Like the ancient Greeks, he felt that a locus had to be legitimized by associating it geometrically or kinematically with another known curve. Perhaps it was the traditionally axiomatic form of geometry that led him in this direction... This represents, of course, a clear-cut break with the Platonic limitation of instruments to compasses and straight-edge, and Descartes makes free use of various linkages and mechanical devices. The concept of movement plays a far more prominent role in his work than in that of Fermat.³⁵

The success of analytic (i.e., algebraic) techniques in solving locus problems dictates that we want an algebraic definition of what a curve is. We have two candidates at our disposal — kinematic, defining them via parametric equations, and algebraic, defining them as the solution set of an equation. Neither definition is perfect, but the latter appears more immediately recognisably imperfect than the former.³⁶ Descartes's elevation of the kinematic over the equational as the definitive hallmark

³³Boyer, *op. cit.*, pp. 75–76.

³⁴This “analytic art” was the beginning of symbolical algebra. The adjective “analytic” here referred to the algebraic analysis of a problem — its expression in algebraic terms and the solution of the resulting equations. Except for “Analytic Geometry”, the adjective “analytic” today refers more generally to those areas of mathematics that the Calculus evolved into, Calculus itself having evolved from Analytic Geometry.

³⁵Boyer, *op. cit.*, pp. 88–89.

³⁶Consider, e.g., the “curve” defined by the constant function $f(x, y) = 0$.

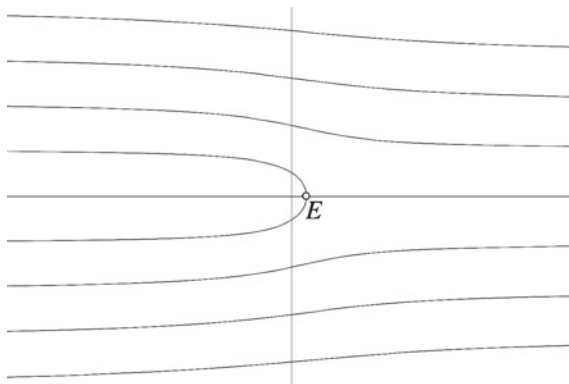


Fig. 2.21 Full quadratrix

of *curveness* may have indicated lingering doubts about an equation's always defining a curve. Moreover, singling out those functions f which define curves is merely an algebraic reformulation of the basic problem of defining what a curve is in the first place. So we will tentatively define a curve as one that is parametrically definable.

2.1.12 Definition Let I be an interval and $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ a function from I to the real plane. The *curve* defined by γ is the range $\gamma(I) = \{\gamma(t) \mid t \in I\}$ of γ ; γ itself is called a *parametrisation* of the curve $\gamma(I)$.

The interval I can be open, closed, half-open, bounded, or unbounded. Moreover, we can relax the requirement that I be an interval to I being a union of disjoint intervals so as to accommodate curves with multiple branches.

Note that the *graph* of a function $y = f(x)$ or $x = f(y)$ falls under the scope of this definition: one simply defines $\gamma(t) = \langle t, f(t) \rangle$ or $\gamma(t) = \langle f(t), t \rangle$, respectively.

Definition 2.1.12 is tentative as it is still a bit too inclusive. Not every parametrically defined function $\gamma(t) = \langle x(t), y(t) \rangle$ has as its range something we would call a curve. We have already noted the two branches of the hyperbola. And if one has used (2.1) to graph the quadratrix on one's graphing calculator, one will have noticed that it consists of lots (in fact, infinitely many) branches. (Cf. Fig. 2.21. Note that the point E is not on the curve.) And one will have seen the graphs of the trigonometric functions $\sec x$, $\csc x$, $\tan x$, and $\cot x$ with their infinite collections of branches.

Even if the range of γ has only a single branch, it might not be something we want to call a curve. Consider the graph of the function

$$y = \begin{cases} \sin 1/x, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

given in Fig. 2.22. In topological terms, it is *connected*, but it is not nicely connected in that in going from the left of the y -axis to the right one does not pass smoothly through the origin at $x = 0$. As x moves closer and closer to 0, y infinitely often

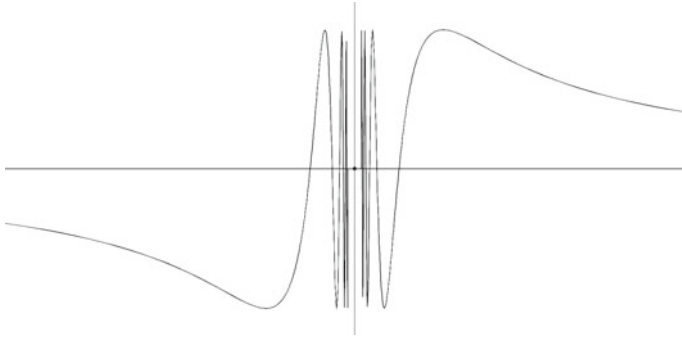


Fig. 2.22 Graph of $\sin(1/x)$

assumes the value 0 only to veer away from it. It does not have 0 as its unique limiting value as x goes to 0. Indeed, we could have defined $y(0)$ to be any value in the interval $[-1, 1]$ with the same result.

More seriously, it can happen that the range of γ can be two dimensional. Every real number in the interval $[0, 1]$ has a decimal expansion $.r_0r_1r_2\dots$ which does not end in an infinite repeating sequence of 9's (with the exception of the value $1 = 1.\overline{0}$). If we define γ by

$$\begin{aligned}\gamma(1) &= \langle 1, 1 \rangle \\ \gamma(.r_0r_1r_2\dots) &= \langle .r_0r_2r_4\dots, .r_1r_3r_5\dots \rangle,\end{aligned}$$

the range of γ is the entire unit square $[0, 1] \times [0, 1]$.³⁷

In the next section we will narrow the definition further and define the class of continuous curves, which more closely fit our intuitive geometric conceptions of curve and kinematic motion, the latter at least as conceived before the advent of the quantum leap.

³⁷The only tricky part is recognising that

$$\begin{aligned}\gamma(.r_09r_19r_29\dots) &= \langle .r_0r_1r_2\dots, .999\dots \rangle = \langle .r_0r_1r_2\dots, 1 \rangle \\ \gamma(.9r_09r_19r_2\dots) &= \langle .999\dots, .r_0r_1r_2\dots \rangle = \langle 1, .r_0r_1r_2\dots \rangle.\end{aligned}$$

2.2 Continuous Curves

2.2.1 Defining Continuity

The notion of a curve as the “flowing of a point” would seem to entail a bit more than a succession of positions through time described parametrically by some listing function γ defined on a time interval as tentatively specified in Definition 2.1.12. The word “flowing” also promises some smoothness to the motion, with no gaps or sudden jumps. It may also suggest no sudden changes in direction, as given by corners and cusps. To accommodate these additional expectations of curveness, we have two refinements of our definition of a curve — definitions of *continuous curves* and *smooth curves*. Continuous curves have no gaps or strange jumps, but may have corners and cusps; smooth curves are allowed none of this sort of bad behaviour. As our central interest in this book is the Mean Value Theorem, we will eventually want to consider smooth curves. But in the Calculus in general, one wants to consider the broader class of continuous curves. Not all motions, after all, are simple smooth “flows”. We have seen cusps, for example, in the motion-defined cycloid; and corners will appear when moving objects are reflected (i.e., bounced) or refracted. In the present section we will consider continuous curves and in the next section we will consider smooth curves.

Modern definitions of “curve” were given in the early decades of the 20th century. Continuity was adequately defined in the 19th, yet it too required some preparation. Where today we would first define a continuous function and then declare a curve to be continuous if it possessed a continuous parametrisation, continuity was a property of curves long before one spoke of functions. A slight familiarity with the history of continuity is not necessary here for our understanding of the Mean Value Theorem, but it will bear on the later history of this Theorem.

Philosophical discussions of continuity generally begin with Aristotle and his belief in the *potentially* infinite divisibility of the line. Any line segment can be divided into two properly smaller segments, each of which can again be divided, and so on — where “so on” means the process can be repeated any finite number of times without bound, not that one can actually do it infinitely often and still have a line rather than a single point at the end. A relatively modern version of this is discussed by Bertrand Russell (1872–1970):

It is generally held by philosophers that numbers are essentially discrete, while magnitudes are essentially continuous. This we shall find to be not the case. Real numbers possess the most complete continuity known, while many kinds of magnitude possess no continuity at all.³⁸

³⁸Bertrand Russell, *Principles of Mathematics*, 2nd ed., W.W. Norton & Company, Inc., New York, no date given, p. 193. The first edition was published in 1903, the second originally in 1938. The printing I quote from is a paperback that I acquired new in the late 1960s or early 1970s and is thus a reprint of the second edition.

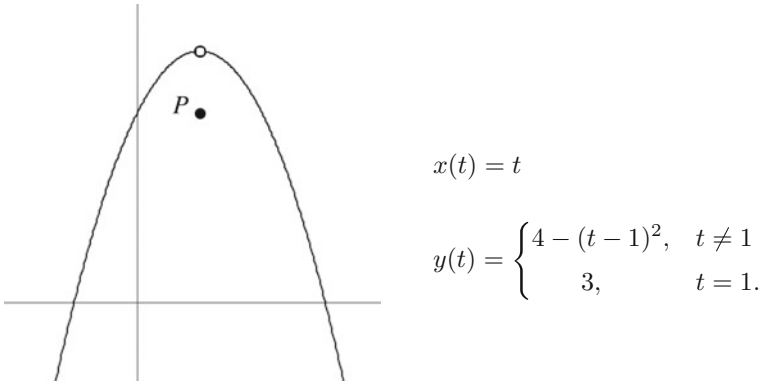


Fig. 2.23 An isolated point

Russell proceeds to define an ordered set to be continuous if, between any two elements of the set, an intermediate one exists. This makes the rational numbers, as well as the reals, continuous. Later in his book he admits that this is insufficient and calls such sets *compact*.³⁹ His new definition of continuity is due to Georg Cantor and Russell spreads the definition over two chapters. This requires the introduction of two additional notions of *perfect* and *cohesive* sets.

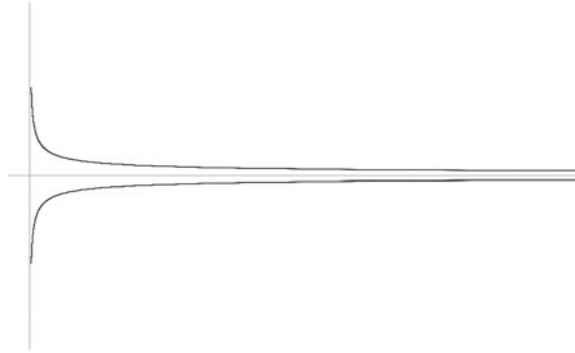
Russell reverses the order of the terms and defines the second one first. A set T is *cohesive* iff for any $t, t' \in T$ and any $\epsilon > 0$, a chain $t_0 = t, t_1, t_2, \dots, t_n = t'$ can be found such that each of the distances $d(t_i, t_{i+1})$ is less than ϵ .

As Russell states, cohesiveness is a sort of connectedness condition, by which the rational numbers would be considered connected. There are gaps, but they all have length 0. If we consider Fig. 2.21 we can see that that part of the full quadratrix consisting of the two branches that approach each other at E is cohesive, but the whole graph is not because the distances between points on any other pair of branches are all at least $\pi > \epsilon$ for small ϵ .

2.2.1 Exercise Is the graph of Fig. 2.22 cohesive? Does your answer depend on whether or not you measure the distances taken along the curve or “as the crow flies”, i.e., the distances between these successive points in the plane?

The second condition, that the set T be *perfect* refers to limits and has two subconditions, namely, i. every point on T is the limit of a sequence of elements of T , and ii. T contains all of its limits. The first subcondition rules out *isolated* points, examples of which have not yet appeared in our illustrations, but are readily given as in Fig. 2.23. The second subcondition requiring that the set contain all of its limits is, even without a formal definition of limit, clearly not satisfied by the curves of Figs. 2.1, 2.21, 2.22, or 2.23.

³⁹The modern term for this is “dense”; “compact” has an altogether different meaning in mathematics.

Fig. 2.24 Second order hyperbola

Cantor's definition is not the best possible. According to it, an open interval, say $(0, 1)$, is not continuous because it fails to include its endpoints. And the simple quadratrix of Fig. 2.1 is not continuous because it lacks the limit point E .

On the other hand, some discontinuous curves are cohesive in Cantor's sense. A particularly simple example would be given by simply removing the point P from Fig. 2.23. A slightly more subtle example is given by the second order hyperbola

$$\begin{aligned} x(t) &= \frac{1}{t^2}, \quad t \in [-4, 0) \cup (0, 4]. \\ y(t) &= t \end{aligned}$$

(See Fig. 2.24.)

Inadequacy aside, Cantor's definition of continuity won't do here because it is too technically advanced for the first year Calculus course. With the post-Cantor advent of Topology, more adequate definitions of the continuity of a set emerged. Roughly, one defines a set T to be *connected* if it contains no gaps, where one recognises a gap by its ability to separate two portions of T . One does not necessarily determine the gap or separation by distance. In Fig. 2.23, the distance from P to the rest of the curve is $3/4$ (*Exercise.*); the distance between the two branches of the quadratrix converging to each other at E in Fig. 2.21 is 0. A disconnecting gap can have measure 0.

Topology offers a formal definition of connectedness in terms of open sets:

2.2.2 Definitions A set U is said to be *open* if, for every point $u \in U$, all points of the space sufficiently close to u also lie in U :

$$\forall u \in U \exists \epsilon > 0 \forall x (\text{dist}(x, u) < \epsilon \Rightarrow x \in U).$$

If U is a subset of the plane, this means that if $P \in U$, then some *open disc* centred at P (i.e., the interior of a circle with centre P) is a subset of U . A set T is *disconnected* by a pair of disjoint open sets U, V if there are nonempty sets $X, Y \subset T$ such that

$$X \subseteq U, Y \subseteq V, T = X \cup Y, U \cap V = \emptyset.$$

A set T that is not disconnected is called *connected*.

2.2.3 Examples i. The graph of the hyperbola $x^2 - y^2 = 1$ of Fig. 2.7 is disconnected. Here we can take

$$U = \{(x, y) \mid x < 0\}, \quad V = \{(x, y) \mid x > 0\};$$

ii. The two branches of the conchoids of Fig. 2.9 are disconnected by the sets

$$U = \{(x, y) \mid x < a\}, \quad V = \{(x, y) \mid x > a\};$$

iii. The full quadratrix of Fig. 2.21 is disconnected by, among others,

$$U = \{(x, y) \mid y > 0\}, \quad V = \{(x, y) \mid y < 0\};$$

iv. The graph of $\sin(1/x)$ as given by Fig. 2.22 is connected, but if one drops the point $(0, 0)$ what remains is disconnected by

$$U = \{(x, y) \mid x < 0\}, \quad V = \{(x, y) \mid x > 0\}.$$

In each of these examples, the openness of the sets U, V is easy to establish and we see that it can be quite easy to show that a curve is disconnected when it is. It can be a lot harder to prove connectivity, as Example 2.2.3.iv illustrates.

Russell and Georg Cantor (1845–1918) were late arrivals on the continuous scene and I mention them first because their divisibility criterion for continuity traces back to Aristotle. Additionally there was always the unconscious assumption that curves which crossed each other actually met in a point. However, until Fermat and Descartes opened up the field by introducing numerous new curves algebraically, curves could be discussed on a case-by-case basis and the need for general definitions never arose. In his classic *La Géométrie*, Descartes used the word continuous, but offered no attempt to analyse the notion or to explain what he meant. It was the explanation, the informal intuition behind the scenes:

...if we think of geometry as the science which furnishes a general knowledge of the measurement of all bodies, then we have no more right to exclude the more complex curves than the simpler ones, provided they can be conceived of as described by a continuous motion or by several successive motions, each motion being completely determined by those which precede; for in this way an exact knowledge of the magnitude of each is always obtainable.⁴⁰

The importance of continuity is again noted later:

⁴⁰René Descartes (David Eugene Smith and Marcia L. Latham, trans.), *The Geometry of René Descartes*, Dover Publications, Inc., New York, 1954, pp. 42 (original French version) and 43 (English translation). The French original was published in 1637 as an appendix to Descartes's philosophical work *Discours de la Methode*. The English translation was first published in 1925 by the Open Court Publishing Company.

But the fact that this method of tracing a curve by determining a number of its points taken at random applies only to curves that can be generated by a regular and continuous motion does not justify its exclusion from geometry.⁴¹

La Géométrie divides into three “books”. The first introduces the work, describing a locus problem of Pappus that he solved using his methods. In the second book he solves various locus problems by deriving equations describing the curves defined, and uses this algebraic formulation to solve various problems involving the curves. This included a computationally intense method of finding tangents and normals to such curves by determining circles that “touch” the curves and then finding the tangents and normals to these circles. The third book deals mainly with the Theory of Equations and the problem of finding roots of polynomials. Descartes did not introduce functions in this work.

Fermat considered functions given by expressions and even came close to the notion of the derivative with a technique for finding the maximum of a curve $y = f(x)$ by manipulating the difference quotient⁴²

$$\frac{f(A + E) - f(A)}{E}.$$

This did not yet bring the notion of function to centre stage. With Descartes and Fermat algebraic expressions entered the stage but there were still other perspectives. Isaac Newton (1642–1727) envisioned a curve as the path traced out by the points of intersection of two lines, a vertical one moving horizontally and a horizontal one moving vertically. Their respective positions x and y were dependent on time, thus, in our modern terminology parametric functions $x(t)$ and $y(t)$ of time. His younger contemporary Gottfried Wilhelm Leibniz (1646–1716) introduced the term “function”, but two mathematical generations later, Leonhard Euler (1707–1783) considered a function to be continuous if the same expression was used throughout an interval — “continuity” meant “continuity of definition”. And curves were still on one’s mind when one began to speak of continuous functions as opposed to continuous curves.

As seen from Euler’s standard textbook *Introductio in Analysin infinitorum* of 1748, continuity was at first understood as a geometrical quality: as a quality of curves. Continuous curves were characterized by the fact that they could be expressed by an analytic expression. In contrast, discontinuous curves consisted of several segments that belonged to different functions and hence did not correspond to just one analytic expression, but to several. This explains why Euler called the non-continuous curves “discontinuous” or “mixed” curves...

In his later treatise of 1763 *De usu functionum discontinuarum in analysis*..., Euler, specifying the concept of continuity, stressed that it is necessary for continuous curves to obey a single analytic law. A hyperbola’s two branches thus form a continuous curve.

The historical literature always refers to Arbogast’s treatise of 1791 as to that which offered new conceptual proposals. This is said firstly because he explicitly formulated the interme-

⁴¹ *Ibid.*, pp. 90 (French) and 91 (English).

⁴² Fermat followed Viète in using vowels A, E, I, O, U for variables x, y, \dots Cf. p. 83 *f.*, below, for a more precise description of Fermat’s technique.

diate value property for continuous functions and secondly because he introduced a new term: “discontigue.” While curves, according to Euler’s specification, had been considered to be discontinuous as well if their various parts were attached to one another, provided that these were defined by different “laws,” Arbogast now called curves *discontigue* if their various parts were unconnected. In all these works, this continual conceptual differentiation is emphasised as an important achievement, because with it, and with the novel term, the discipline had come closer to the meaning of discontinuity as it is understood today.

It must be pointed out, however, that Arbogast’s reflections on the meaning of *continue*, *discontinue* and *discontigue* still refer to curves, and that functions, for him, were only of secondary importance for representing particular parts of a curve. Arbogast assumed functions as basic concepts only when reflecting on intermediate values.⁴³

The distinction between continuity and contiguity, something of a non-issue today, actually lies at the heart of the matter. It was the climax of a four-and-a-half decade long controversy over the vibrating string problem. In 1746 Jean le Rond d’Alembert (1717–1783) wrote his first paper on the vibrating string. This was a geometric problem, but the method of attack was analytic. D’Alembert insisted on functions continuous in the Eulerian, algebraic sense of continuity of expression. Euler, who also considered the problem, allowed mixed functions, both in determining the initial shape of the string and in specifying the solution. These solutions involved the solution of what are called partial differential equations and in 1787 the St. Petersburg Academy proposed a prize competition concerning these solutions:

“Do they belong to any curves or surfaces either algebraic, transcendental,⁴⁴ or mechanical, either discontinuous or produced by a simple movement of the hand? Or shouldn’t they legitimately be applied only to continuous curves susceptible of being expressed by algebraic or transcendental equations?”⁴⁵

The prize was awarded to Louis Arbogast (1759–1803), whose treatise on the matter was published in 1791.

Historian Judith Grabiner sums up Arbogast’s contribution succinctly:

To get a feeling for the climate of mathematical opinion about continuous functions in which Bolzano and Cauchy worked, it will suffice to quote from the work of L.F.A. Arbogast, who won the St. Petersburg Academy’s prize competition in 1787 by giving the best characterization of those functions that would be allowable solutions to the vibrating-string equation. Arbogast described these functions in several ways: The functions make “no jumps”; they have the intermediate-value property; and they increase in increments whose sizes correspond to the sizes of the increments in the variable. For instance, if the function is represented by

⁴³Gert Schubring, *Conflicts between Generalization, Rigor, and Intuition: Number Concepts Underlying the Development of Analysis in 17–19th Century France and Germany*, Springer Science+Business Media, Inc., New York, 2005, pp. 26–27. In quoting this I have omitted Schubring’s citations to the literature.

⁴⁴Descartes and Fermat had introduced algebraic descriptions $f(x, y) = 0$ for curves, where f was a polynomial; very quickly transcendental functions like sines, cosines, logarithms, etc., were introduced into the composition of f .

⁴⁵Citation from Jean Itard, “Arbogast, Louis François Antoine”, in: Charles Coulston Gillispie (ed.), *Dictionary of Scientific Biography*, vol. 1, Charles Scribner’s Sons, New York, 1970, p. 207. Itard adds, “The Academy was thus requesting a drastic settlement of the dispute between Jean d’Alembert, who adopted the second point of view, and Leonhard Euler, partisan of the first”.

two different formulas on adjacent intervals, “the last ordinate of the old form, and the first of the new, are equal to each other, or differ only by an infinitely small quantity.” Again, “The ordinate y cannot pass brusquely from one value to another; there cannot be a jump from one ordinate to another which differs from it by an assignable quantity.” This “no-jumps” characterization, though it helps call attention to the crucial property of continuous function as defined by Cauchy and Bolzano, is not in itself an anticipation of that definition; it deals with functions that are piecewise continuous, and discusses the behavior of the function in detail only at the break. Thus a definition of the continuity of the “piece” is still lacking. But Arbogast was concerned about this question. He linked his no-jumps property to the intermediate-value property, saying that the functions had to obey what he called the “law of continuity”: “A quantity cannot pass from one state to another without passing through all the intermediate states subject to the same law.” The closest Arbogast came to the Cauchy-Bolzano definition was to say “assuming that the variable increases continually, the function receives corresponding variations,” though the language is not sufficiently precise to be a real anticipation of that definition.⁴⁶

The property cited is sufficiently important to be singled out and given a formal definition.

2.2.4 Definition A real-valued function defined on an interval I has the *intermediate value property* if, whenever $a, b \in I$ and d are such that $f(a) < d < f(b)$, there is some c between a and b for which $f(c) = d$.

In the standard course in the Calculus one learns that every continuous function $f : I \rightarrow \mathbb{R}$ has the intermediate value property. The intermediate value property is clearly a version of the no gaps requirement for the continuity of the curve $y = f(x)$ and is the requirement most explicitly stated by Arbogast for the continuity of such a function. The question arose: does the intermediate value property characterise continuity of real-valued functions of reals?

2.2.5 Example Consider the function graphed in Fig. 2.22,

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right), & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Intuitively it is clear that f has the intermediate value property. But it fails to be continuous in two senses. First, the graph, considered as a set, is not continuous in Cantor’s sense: The upper points of the oscillations of the curve, i.e., the points

⁴⁶Judith Grabiner, “Cauchy and Bolzano: tradition and transformation in the history of mathematics”, in: Everett Mendelsohn (ed.), *Transformation and Tradition in the Sciences: Essays in Honor of I. Bernard Cohen*, Cambridge University Press, Cambridge, 1984, p. 112. A similar, earlier, discussion of the matter was given by Grabiner in: Judith V. Grabiner, *The Origins of Cauchy’s Rigorous Calculus*, The MIT Press, Cambridge (Mass.), 1981, pp. 91–92. This book was reprinted by Dover Publications, Inc., in 2005. Accessible fuller quotations from Arbogast can be found in: C.H. Edwards, Jr., *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979, pp. 303–304; Umberto Bottazzini (Warren van Egmond (trans.)), *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass*, Springer-Verlag, New York, 1986, pp. 34–35; and (in German) Klaus Volkert, *Geschichte der Analysis*, Bibliographisches Institut & F.A. Brockhaus AG, Zürich, 1988, pp. 170–171.

$$\left\langle \frac{2}{(4n+1)\pi}, \sin \frac{(4n+1)\pi}{2} \right\rangle = \left\langle \frac{2}{(4n+1)\pi}, 1 \right\rangle$$

for $n = 0, 1, 2, \dots$ converge to the point $\langle 0, 1 \rangle$, which is not on the curve. Viewed as a set, the curve is not closed under taking limits and is thus not perfect in Cantor's sense, i.e., he wouldn't accept it as continuous.

And it does not satisfy Arbogast's less emphasised condition mentioned by Grabiner that "functions...increase in increments whose sizes correspond to the sizes of the increments in the variable". This definitely fails at $x = 0$ where the tiniest increment Δx can take one from $f(0) = 0$ to $f(\Delta x) = \pm 1$.

2.2.6 Remark I confess to ignorance of the origin of this Example. Augustin Louis Cauchy (1789–1857) cites⁴⁷ the function $\sin(\frac{1}{x})$ as one which, as x tends to 0, "admits an infinity of limits between the limits -1 and $+1$ "⁴⁸ but doesn't use the function as an explicit counterexample to anything. In an unfinished manuscript written in the 1830s or so, Bernard Bolzano (1781–1848) addresses the problem, but his explanation is a bit vague and only seems to yield a weak counterexample.⁴⁹ However, a bit later in the same work,⁵⁰ he cites the function

$$f(x) = \sin \ln(1 - x)$$

as one which oscillates infinitely often on the interval $[0, 1)$. This function readily yields an example analogous to that of Example 2.2.5, namely

$$f(x) = \begin{cases} \sin \ln |1 - x|, & x \neq 1 \\ 0, & x = 1. \end{cases}$$

Klaus Volkert credits Gaston Darboux (1842–1917) with being in 1875 the first to definitively answer in the negative the question of whether or not the intermediate value property implies continuity.⁵¹ The oscillating sine function has become a popular example. Multiplication by x ,

$$g(x) = \begin{cases} x \sin \left(\frac{1}{x} \right), & x \neq 0 \\ 0, & x = 0, \end{cases}$$

⁴⁷ Augustin Louis Cauchy, *Cours d'analyse de l'École Royale Polytechnique; I.^{re} Partie. Analyse algébrique* [Course in Analysis of the Royal Polytechnical School; Part I. Algebraic Analysis], de Bure, Paris, 1821. English translation: Robert E. Bradley and C. Edward Sandifer (eds. and trans.), *Cauchy's Cours d'analyse; An Annotated Translation*, Springer Science+Business Media, LLC, New York, 2009. The function $\sin(\frac{1}{x})$ is cited on p. 12 of the Bradley/Sandifer edition.

⁴⁸ The word "limit" is used in two senses here. The first occurrence refers to what we now call *limit points*; the second refers to the endpoints of the interval $[-1, 1]$ on the y -axis.

⁴⁹ The manuscript is called "Functionenlehre" ["Theory of functions"] and can be found in: Steve Russ (ed.), *The Mathematical Works of Bernard Bolzano*, Oxford University Press, Oxford, 2004. For Bolzano's counterexample, cf. pp. 471–472 (§§83–84, but see also §46, pp. 453–454).

⁵⁰ *Ibid.*, p. 481, §102.

⁵¹ Volkert, *op. cit.*, p. 187. Cf. Lemma 3.1.5 on page 187, below.

Fig. 2.25 Graph of $x \sin(1/x)$

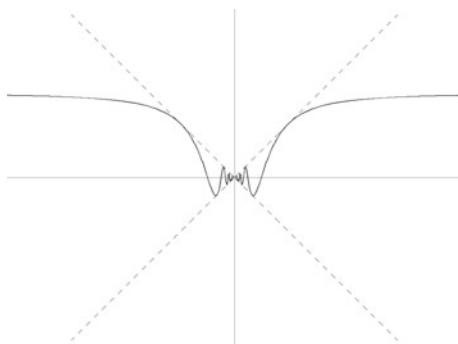
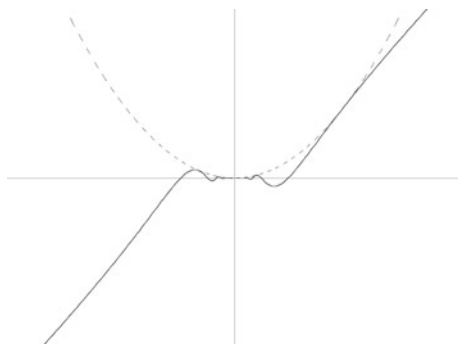


Fig. 2.26 Graph of $x^2 \sin(1/x)$



results in a continuous function (see Fig. 2.25.) with infinite arc length over any interval containing 0. And Darboux uses the variant,

$$h(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right), & x \neq 0 \\ 0, & x = 0, \end{cases}$$

(See Fig. 2.26.), which we will refer to in the sequel as *Darboux's function*, as an example of an infinitely oscillating differentiable function with a discontinuous derivative at $x = 0$.⁵² Functions of the form $f(x) = x^\alpha \sin(\frac{1}{x})$ for $\alpha > 0$ form a rich class of counterexamples in the Calculus.⁵³

Returning to the intermediate value property, we see in it a manifestation of continuity — the “flowing of a point” —, but not a sufficient condition for an adequate definition thereof. Moreover, if we reflect on general curves $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$, we must ask: if we were to try to define continuity for $f : I \rightarrow \mathbb{R}$ by means of the intermediate value property, how could we extend this definition to γ , i.e., what is the proper ana-

⁵²Gaston Darboux, “Mémoire sur les fonctions discontinues”, *Annales scientifiques de l'École Normale Supérieure*, 2nd series, vol. 4 (1875), pp. 57–112; here: p. 109.

⁵³H. Turgay Kaptanoğlu, “In praise of $y = x^\alpha \sin(\frac{1}{x})$ ”, *American Mathematical Monthly* 108 (2001), pp. 144–150.

logue of the intermediate value property in two-dimensional space? The intermediate value property emerges as more of a deflection from our goal than a path towards it. Nevertheless, it is a useful property for a function to have and numerous mathematicians attempted to prove that those functions defined by “analytic expressions” possessed the property. Bolzano, who gave the first correct proof for continuous functions, cites attempts by Abraham Gotthelf Kästner, Alexis Claude Clairaut, Sylvestre François Lacroix, Mathias Metternich, Georg Simon Klügel, Joseph Louis Lagrange, and Christian Lebrecht Rösling, and others.⁵⁴ To carry out his proof, Bolzano gave the first “correct” definition of continuity.

I have put cautionary quotation marks around the word “correct” to emphasise that correctness here is not absolute. The definition is correct in that it more-or-less agrees with our modern definition. And it is correct in that it works — those functions we think of as being continuous are continuous under his definition, and theorems we would expect to hold for continuous functions are indeed provable. Without further ado, I quote Bolzano’s definition.

Following the *correct definition* one understands by the expression that *a function $f(x)$ varies according to the law of continuity for all values of x , which lie inside or outside certain bounds, only so far as, if x is any such value, the difference $f(x + \omega) - f(x)$ can be made to be smaller than any given value, if x is so small as one wishes to make it.*⁵⁵

In a footnote, Bolzano explains the conditions on the domain:

There are functions which vary continuously for all values of their arguments, e.g. $\alpha x + \beta x$. But there are others which vary according to the law of continuity only within or without certain limits of their roots. So $x + \sqrt{(1-x)(2-x)}$ varies continuously only for all values of x which are $< +1$ or $> +2$; but not for the values which lie between $+1$ and $+2$.⁵⁶

And in the text itself he remarks on the intermediate value property as not defining continuity:

But...the continuous function never reaches a higher value without first going through all lower values, i.e., that $f(x + n\Delta x)$ can take on every value lying between $f(x)$ and $f(x + \Delta x)$, if one takes n arbitrarily between 0 and $+1$: this is indeed a very true conjecture, but it cannot be seen as the *definition* of the concept of continuity but rather is a *theorem* about the same.⁵⁷

⁵⁴Bernard Bolzano, *Rein analytischer Beweis des Lehrsatzes, daß zwischen je zwey Werthen, die ein entgegengesetztes Resultat gewähren, wenigstens eine reelle Wurzel der Gleichung liege*, Gottlieb Haase, Prague, 1817. The work also appeared the following year in volume 5 of the *Abhandlungen der königlichen böhmischen Gesellschaft der Wissenschaften*, and was edited and reprinted by Philip E.B. Jourdain in 1905 as half of number 153 of *Ostwalds Klassiker der exakten Wissenschaften*. English translations by Steve Russ and William Ewald appeared first in 1980 and 1996, respectively. The most recent English version appears in Russ’s edition of *The Mathematical Works*, *op. cit.* Below, I shall refer to the *Ostwald Klassiker* reprint as “Bolzano, *Klassiker*” in what follows, but will also give references to *The Mathematical Works* for English translations. Thus, the above list of names can be found in: Bolzano, *Klassiker*, p. 6; Russ, *op. cit.*, p. 253.

⁵⁵Bolzano, *Klassiker*, pp. 3–4, Russ *op. cit.*, p. 256.

⁵⁶*Ibid.* Presumably Bolzano intends “ $<$ ” here to read “less in absolute value than”.

⁵⁷Bolzano, *Klassiker*, p. 6; Russ, *op. cit.*, pp. 256–257.

It was in 1817 that Bolzano published his definition of continuity in a short pamphlet, the long title of which translates to *Purely Analytic Proof of the Theorem that between any two Values, which give Results of Opposite Sign, there lies at least one real Root of the Equation*. This is one of the major works in the history of the foundations of the Calculus, including the definition of continuity, the introduction of Cauchy sequences and a proof of their convergence,⁵⁸ a proof of the least upper bound property as a corollary, yielding the Bolzano-Weierstrass Theorem⁵⁹ in the process, and, finally, the statement and proof of the Intermediate Value Theorem:

Theorem. If two functions of x , $f(x)$ and $\varphi(x)$, vary according to the law of continuity either for all values of x or for all which lie between α and β , if further $f(\alpha) < \varphi(\alpha)$ and $f(\beta) > \varphi(\beta)$, then between α and β there is always a value of x for which $f(x) = \varphi(x)$.^{60, 61}

Four years after Bolzano's paper was published, Cauchy's lectures, the famous *Cours d'analyse*, offering an independent treatment, was published. This work is the first part of what was intended to be a two-part textbook on the Calculus. The second part, which would have included the Differential and Integral Calculus was delayed a couple of years and published under a different title. The first part, which runs several hundred pages, does not get as far as differentiation or integration, but lays the foundations of the Calculus, treating the real numbers, continuity, infinite series, complex numbers, and related topics. In the midst of all of this is Cauchy's definition of continuity:

⁵⁸A *Cauchy sequence* is a sequence a_0, a_1, a_2, \dots of numbers satisfying: for any $\epsilon > 0$ a number n_0 can be found such that for all $m, n > n_0$ one has $|a_m - a_n| < \epsilon$. The convergence of such sequences had been used without note by Euler. Bolzano drew attention to them and proved their convergence relative to his notion of real number as incompletely treated in a later work not published in his lifetime. Jacqueline Stedall finds the proof "incorrectly argued" (Stedall, *op. cit.*, p. 496):

It turned out to be more difficult than it might seem, and Bolzano was forced to introduce [as] a fresh assumption the existence of a quantity X to which the terms of the series approach as closely as we please. Such a hypothesis, Bolzano claimed 'contains nothing impossible' ..., but it was precisely what he was trying to prove in the first place. The problem was deeper than Bolzano realized. Convergence of Cauchy sequences requires *completeness* of the real numbers or, simply speaking, that the number line is an unbroken continuum with no gaps. Convergence of Cauchy sequences is in fact mathematically *equivalent* to completeness: either must be assumed in order to prove the other. Without some such assumption, Bolzano was forced to introduce his hypothetical quantity X .

This is a fair criticism, but I give Bolzano full credit nonetheless as he later offered some justification for his variant of completeness on which his proof of the convergence of Cauchy sequences was based. I discuss this sort of thing in some detail in Smoryński, *Formalism, op. cit.*, pp. 232–265.

⁵⁹The Bolzano-Weierstrass Theorem asserts that any bounded sequence a_0, a_1, a_2, \dots of numbers, i.e., any such sequence for which there is a bound $B > |a_n|$ for all n , has a convergent subsequence. It is a fundamental result of Analysis.

⁶⁰Bolzano, *Klassiker*, p. 31; Russ, *op. cit.*, p. 273.

⁶¹Bolzano is a little sloppy here: In his example cited above, the law of continuity is two-sided and does not apply to the endpoints α, β of an interval, but his proof of the Theorem assumes the one-sided continuity of f and ϕ at the endpoints of the interval.

Let $f(x)$ be a function of the variable x , and suppose that for each value of x between two given limits, the function always takes a unique finite value. If, beginning with a value of x contained between these limits, we add to the variable x an infinitely small increment α , the function itself is incremented by the difference

$$f(x + \alpha) - f(x),$$

which depends both on the new variable α and on the value of x . Given this, the function $f(x)$ is a *continuous* function of x between the assigned limits if, for each value of x between these limits, the numerical value of the difference

$$f(x + \alpha) - f(x)$$

decreases indefinitely with the numerical value of α . In other words, *the function $f(x)$ is continuous with respect to x between the given limits if, between these limits, an infinitely small increment in the variable always produces an infinitely small increment in the function itself.*

We also say that the function $f(x)$ is a continuous function of the variable x in a neighborhood of a particular value of the variable x whenever it is continuous between two limits of x that enclose that particular value, even if they are very close together.

Finally, whenever the function $f(x)$ ceases to be continuous in the neighborhood of a particular value of x , we say that it becomes discontinuous, and that there is a *solution of continuity* for this particular value.⁶²

In 1817 Bolzano had been careful to avoid using the words “infinitely small”. And half a century later, in completing the work of Bolzano and Cauchy on the “arithmetisation of analysis” as their programmes of bringing rigour to the Calculus came to be called, Karl Weierstrass (1815–1897) treated these words as a mere figure of speech. Cauchy, however, used infinitesimals in an essential way. To him, a function continuous in an interval was continuous at all numbers in the interval,⁶³ not just at the real numbers in the interval. This has powerful consequences and Cauchy’s notion of continuity is strictly stronger when the domain is an open interval than is ordinary continuity.

Speaking of ordinary continuity, Weierstrass gives an equally prosaic definition in his lectures of 1861:

If $f(x)$ is a function of x and x is a definite value, then the function will change into $F(x + h)$ ⁶⁴ if x passes from x to $x + h$; the difference $f(x + h) - f(x)$ is called the change which the function experiences through the passage of the argument from x to $x + h$. If it is now possible to determine a bound δ for h so that for *all* values of h of absolute value smaller than δ , $f(x + h) - f(x)$ will be smaller than any ε however small, one says to infinitely small changes in the argument correspond infinitely small changes of the function. Because one says, if the absolute value of a quantity can be made smaller than any arbitrarily chosen value, however small, it can be chosen infinitely small. If now a function is so obtained that [to] infinitely small changes in the argument correspond infinitely small changes in

⁶²Bradley and Sandifer, *op. cit.*, p. 26. The editors explain that “solution of continuity” is to be read as “dissolving of continuity”, i.e., the breakdown of continuity is meant. Note again, as in footnote 48, the use of the word “limits” to mean “endpoints”.

⁶³I.e., at every number $r + \eta$ in the interval, where r is real and η is infinitesimal.

⁶⁴*Sic.* This should read $f(x + h)$.

the function, then one says this function is a *continuous function* of its argument, or that it changes continuously with this argument. — For individual values of the arguments of functions, which are in general continuous, the continuity can become interrupted. For such values the function will be discontinuous.⁶⁵

The truth be told, this is not at first sight any clearer than Bolzano's or Cauchy's definitions. All three are at least partly ambiguous. Bolzano and Cauchy depart from our modern practice of defining what it means for a function to be continuous at a point, and define what it means for a function to be continuous on an open interval or intervals. Both authors can produce δ for given ϵ when necessary, but it is not clear from the given definitions whether δ depends only on ϵ and the interval in question, or if it is allowed to depend on x as well. Reading their proofs suggests Bolzano allows δ to depend on ϵ and x (ordinary continuity) while Cauchy insists δ depends only on ϵ (uniform continuity). Weierstrass starts out defining continuity at a point, but his continued clarification makes this less clear.

Bolzano was writing a major work on the foundations of the Calculus when he died and he never completed the task, his impressive partial work only first published in the 20th century. Cauchy prepared lectures on the subject and published them. They were widely read in France and Germany. Weierstrass lectured regularly on *Functionenlehre*, the Theory of Functions, covering real and complex number systems, and the foundations of the Calculus and that of the theory of functions of a complex variable. He generally did not publish the results of these lectures, but copies were deposited in Mathematical Institute libraries around Germany and, additionally, his students were not shy about publishing expositions of the work of their master. One of these was Eduard Heine (1821–1881), whose “Die Elemente der Functionenlehre”⁶⁶ [“The elements of the theory of functions”] is occasionally cited as the first published modern definition of continuity.

In this paper, Heine begins by singing the praises of Weierstrass:

The advance of the Theory of Functions is actually limited by the circumstance, that certain of its elementary propositions, although proven by a penetrating researcher, will still be doubted, so that the results of an investigation are not held universally as correct, if they rest on these indispensable fundamental assertions. The explanation I find therein, that indeed the principles of Mr. Weierstrass, directly through his lectures and other oral communications, indirectly through copies of exercise books, which would have been worked out following these lectures, themselves having been disseminated in wider circles, that however have not been published by him himself in authentic versions, so that there is no place at which one can find these propositions *developed in context*.⁶⁷

Heine's paper divides into a Part A on numbers and a Part B on functions. Part A gives an infinitistic construction of the real numbers from the rational numbers

⁶⁵Karl Weierstrass and Hermann Amandus Schwarz, *Differential Rechnung, nach einer Vorlesung des Herrn Weierstrass im Sommersemester 1861*, Hdschr. Koll. N 37 (Humboldt-Universität zu Berlin), pp. 2–3.

⁶⁶Eduard Heine, “Die Elemente der Functionenlehre”, *Journal für die reine und angewandte Mathematik* 74 (1872), pp. 172–188.

⁶⁷*Ibid.*, p. 172.

using Cauchy sequences and depends more on Cantor than on Weierstrass.⁶⁸ Part B on functions followed Weierstrass more closely.

In Part B, §1 Heine gave a modern definition of function and in §2 he led off with a definition of what is now called *pointwise continuity*:

1. *Definition.* A function $f(x)$ is called *continuous at a given individual value* $x = X$ if, for every given magnitude ε , however small, there exists another positive number η_0 of such a nature, that for no positive magnitude η , which is smaller than η_0 , does the absolute value of $f(X \pm \eta) - f(X)$ exceed ε .⁶⁹

There may be some awkwardness in the phrasing, but there is no ambiguity. It agrees substantially with the modern definition. The only major difference is that we are more explicit in assuming f to be defined on an interval around the given point at which f is to be continuous:

2.2.7 Definition Let I be an interval, $f : I \rightarrow \mathbb{R}$ a function defined for all elements of I , and $x \in I$ a point in the interval. f is *continuous at x* if for every $\epsilon > 0$ there is a $\delta > 0$ such that for all $y \in I$ whenever $|y - x| < \delta$ we have $|f(y) - f(x)| < \epsilon$:

$$\forall \epsilon > 0 \exists \delta > 0 \forall y \in I (|y - x| < \delta \Rightarrow |f(y) - f(x)| < \epsilon).$$

Defining $|\langle x, y \rangle| = \sqrt{x^2 + y^2} = \text{dist}(\langle 0, 0 \rangle, \langle x, y \rangle)$ for $\langle x, y \rangle \in \mathbb{R} \times \mathbb{R}$, we can readily adapt this definition to functions $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$:

2.2.8 Definition Let I be an interval, $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ a function defined for all elements of I , and $t \in I$ a point in the interval. γ is *continuous at t* if

$$\forall \epsilon > 0 \exists \delta > 0 \forall t' \in I (|t' - t| < \delta \Rightarrow |\gamma(t') - \gamma(t)| < \epsilon).$$

We need one more definition before we can define what it is for a curve to be continuous.

2.2.9 Definition Let I be an interval and f a function mapping I either to \mathbb{R} or $\mathbb{R} \times \mathbb{R}$. f is said to be *continuous on I* if f is continuous at all points in I .

With this, we have the following.

2.2.10 Definition Let $C \subseteq \mathbb{R} \times \mathbb{R}$ be a curve. C is a *continuous curve* if there is an interval I and a continuous function $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ such that $C = \gamma(I)$, i.e., C is a continuous curve if it has a continuous parametrisation.

⁶⁸In the 1830s Bolzano offered a description of real numbers that nowadays one would treat as such a construction, but this went unpublished until the 20th century. At some later, undetermined, date (cf. pp. 334–335, below), Weierstrass offered such a construction treating real numbers as abstract sums of rationals. And in 1858 Richard Dedekind independently constructed the reals using sets of rationals. None of this was published until 1872 when several such constructions, new and old, simultaneously made it into print. Cantor's, Charles Méray's, and Heine's constructions used Cauchy sequences.

⁶⁹Heine, *op. cit.*, p. 182.

2.2.2 Properties of Continuity

We have finally fulfilled our promise that we would define what a continuous curve is in this section, and presumably we could now turn to our next task, namely, that of defining what we mean by a smooth curve. The reader who has had his fill of the discussion of continuity will doubtless want to skip ahead to the next section for this topic. I beg the reader's indulgence, however, as there is a bit more to be said about continuity, both generalities and results that will be needed later.

The first thing to note is that the definition of continuity at a point is an unnatural and non-intuitive notion. For centuries continuity referred to a flow, a smoothness of motion, a blending (as of shades of colour), or the non-existence of gaps — it had nothing whatsoever to do with a single stationary point. The natural notion refers to continuity on an interval: the point flows over an interval of time, the motion is smooth for a while, etc. What we have with Definition 2.2.7 is a technically useful generalisation of the expected formalisation of the concept of a function's being continuous on an interval. Bolzano's and Cauchy's definitions were of the continuity of a function on an interval. It is not clear which of the two notions Weierstrass is referring to in the translated quotation from his 1861 lecture; he begins by referring to a “definite value” x , but finishes with the sentence,

If now a function is such that to infinitely small changes of the argument...

Does “the argument” still refer to a fixed x or is he referring to arbitrary elements in the function's domain? In the intervening years, as he lectured repeatedly on the Theory of Functions, Weierstrass undoubtedly cleared up the ambiguity, but in print we have Heine to thank for this clarification.

Heine leads off §3 of his paper defining *two* notions of continuity on a closed interval:

1. *Definition.* A function $f(x)$ is called *continuous* from $x = a$ to $x = b$, if it is continuous (B, §2, Def. 1)⁷⁰ for each individual value $x = X$ between $x = a$ and $x = b$, with the inclusion of the values a and b ; it is called *uniformly continuous* from $x = a$ to $x = b$, if for every magnitude ε , however small, there is a positive magnitude η_0 such that for all positive values η , which are smaller than η_0 , $f(x \pm \eta) - f(x)$ remains below ε . Whatever value one may give to x , assuming only that x and $x \pm \eta$ belong to the region from a to b , the same η_0 must yield the demanded [inequality].^{71, 72}

Once again, the use of natural language makes for awkward phrasing and mathematicians might prefer using more precise formal language:
 f is continuous on I if

$$\forall x \in I \forall \epsilon > 0 \exists \delta > 0 \forall y \in I (|y - x| < \delta \Rightarrow |f(y) - f(x)| < \epsilon);$$

⁷⁰The reference is to his earlier definition cited on page 183, above.

⁷¹*Ibid.*, p. 184.

⁷²One way of visualising this is to imagine a rectangle $[a - \delta, a + \delta] \times [f(a) - \epsilon, f(a) + \epsilon]$ of fixed size $2\delta \times 2\epsilon$. As one moves $\langle a, f(a) \rangle$ along the curve, the graph over the interval $[a - \delta, a + \delta]$ always remains inside the rectangle.

and f is uniformly continuous on I if

$$\forall \epsilon > 0 \exists \delta > 0 \forall x \in I \forall y \in I (|y - x| < \delta \Rightarrow |f(y) - f(x)| < \epsilon).$$

One would then point out that in the former definition δ depends on x as well as on ϵ , while in the latter case, given ϵ , the choice of δ is uniform for all x .

But I betray here my background as a mathematical logician. Mathematicians are not generally so happy with alternations of quantifiers and prefer instead to introduce the *modulus of continuity*, by which is meant a function yielding δ — a function $\delta(x, \epsilon)$ of two arguments in the former case and a function $\delta(\epsilon)$ in the latter.

Without great care in its formulation, the definition of continuity on an interval can be read either as continuity on the interval or as uniform continuity on the interval. It is generally agreed that Bolzano had in mind the former concept and it has been put forward without yet achieving universal agreement that Cauchy meant the latter. In work that lay unpublished until the 20th century, in the 1830s Bolzano recognised the distinction and proved⁷³ that a function continuous on a closed bounded interval is uniformly continuous there. Cauchy apparently never noted the distinction, always dealing with the uniform notion.

In the United States uniform continuity is not mentioned in the introductory Calculus course, being deemed a topic for an advanced course in the subject. It is not deeper or more difficult a concept than ordinary continuity on an interval. Indeed, direct proofs that various functions are continuous usually yield uniform continuity and I imagine the failure to mention the notion is the fact that one would feel obliged to discuss the relation between the two notions of continuity on an interval, a relation easy enough to state but requiring a proof taking one to a higher level of abstraction. There are, however, several results that are asserted without proof in the first year Calculus course — the Intermediate Value Theorem, the Extreme Value Theorem, the existence of the integral of a continuous function on a closed bounded interval, etc.

In a footnote in his paper “Die Elemente der Functionenlehre”,⁷⁴ Heine remarks that the results of §3 of his paper generally follow the principles of Weierstrass, Heine himself contributing only to the details of execution. The important results in fact predate Weierstrass: the Intermediate Value Theorem (Bolzano 1817, Cauchy

⁷³His exposition is muddled and not everyone accepts it, but a correct proof was certainly within his grasp. Cf. pages 301–302, below, for details.

⁷⁴Heine, *op. cit.*, p. 182.

1821), the Extreme Value Theorem (Bolzano 1830s),⁷⁵ and the Uniform Continuity Theorem (Bolzano 1830s).

Both the Intermediate Value Theorem and the Extreme Value Theorem are intimately connected with the Mean Value Theorem and I ought to say something about their proofs. The Uniform Continuity Theorem is not obviously⁷⁶ as central to our present purpose, but its discussion is not a great digression, and, in any event, the result will be used repeatedly in the sequel.

2.2.11 Theorem (Intermediate Value Theorem) *Let $a < b$ and let $f : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$. Suppose $f(a) < 0 < f(b)$. Then there is some $c \in (a, b)$ such that $f(c) = 0$.*

Proof sketch. Probably the simplest proof uses infinite integers and infinitesimals, à la Euler and Cauchy⁷⁷ Let N be an infinite integer and consider the *finite⁷⁸ set of values

$$f(a), f\left(a + \frac{b-a}{N}\right), f\left(a + 2\frac{b-a}{N}\right), \dots, f\left(a + N\frac{b-a}{N}\right).$$

Let K be the smallest integer such that $f\left(a + K\frac{b-a}{N}\right) \geq 0$. Then $a + K\frac{b-a}{N}$ differs from a standard real c by an infinitesimal amount. $f(c)$ must of necessity be 0 as, by continuity it is infinitesimally close to $f\left(a + (K-1)\frac{b-a}{N}\right) < 0$ and to $f\left(a + K\frac{b-a}{N}\right) \geq 0$. \square

For one not familiar with Nonstandard Analysis, this may make no sense at all, but it is quite rigorous. The nice thing about it, in addition to its simplicity, is that it adapts quickly to a proof of the Extreme Value Theorem.

2.2.12 Theorem (Extreme Value Theorem) *Let $a < b$ and let $f : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$. There are $c, d \in [a, b]$ such that for all $x \in [a, b]$,*

⁷⁵In Craig Smoryński, *A Treatise on the Binomial Theorem*, College Publications, London, 2012, p. 138, I also credit Cauchy with a proof of this Theorem. In glancing over his two main textbooks I have not found the result proven although it is appealed to in the *Résumé des leçons données à l'École Royale Polytechnique sur le calcul infinitesimal*, de Bure, Paris, 1823. The nonstandard proofs of the Intermediate Value Theorem, which Theorem is proven in the *Cours d'analyse*, and the Extreme Value Theorem being virtually identical, I must have simply assumed Cauchy had proven the latter. It would naturally have fit into the projected second volume of the *Cours*. As this volume was intended as a textbook and policies at the École had changed, Cauchy did not include as much foundational material in the *Résumé* when he came to write this later. So he might have proven the result and simply neglected to include the proof in any of his textbooks.

⁷⁶But see Sect. 6, below.

⁷⁷Such a proof, long discredited, is nowadays acceptable thanks to the rigorous foundation and development of Nonstandard Analysis. The reader unfamiliar with these modern developments may consider the proof merely heuristic. The curious reader who would like to know more is referred to Chapter II, Sect. 6, of Smoryński, *Formalism*, for an introduction to and some references on the subject.

⁷⁸In Nonstandard Analysis, a set of nonstandard numbers is called *finite if it can be put into one-to-one correspondence with an integer, finite or infinite, by an “internal” function. In simple terms, a *finite set is a possibly infinite set that behaves like a finite set.

$$f(c) \leq f(x) \leq f(d),$$

i.e., f assumes minimum and maximum values on $[a, b]$.

Proof sketch. Again, let N be an infinite integer and again consider the values

$$f(a), f\left(a + \frac{b-a}{N}\right), f\left(a + 2\frac{b-a}{N}\right), \dots, f\left(a + N\frac{b-a}{N}\right).$$

This set has a least element $f\left(a + K_1\frac{b-a}{N}\right)$ and a greatest element $f\left(a + K_2\frac{b-a}{N}\right)$. Then c and d are the “standard parts” of $a + K_1\frac{b-a}{N}$ and $a + K_2\frac{b-a}{N}$, respectively, i.e., the real numbers infinitesimally close to these. \square

Those not schooled in Nonstandard Analysis might find these proofs too simple to be believable. The usual proofs are a bit deeper, requiring an iterated partitioning, a sequence of approximations, an appeal to the convergence of Cauchy sequences, and, in the latter proof, an invocation of the Bolzano-Weierstrass Theorem. Against this I point out that much of the difficulty in the standard proofs has been transferred in Nonstandard Analysis to the construction of the nonstandard reals. There is thus no reason to distrust the nonstandard proofs on the grounds of their simplicity.

Note that the Extreme Value Theorem of necessity is valid for closed intervals $[a, b]$ and not open ones (a, b) . The open interval (a, b) being without extreme values itself maps trivially to an interval without extreme values via any strictly increasing function, e.g., $f(x) = x$.

2.2.13 Exercise Graph the following functions and find open intervals over which they fail to satisfy the Extreme Value Theorem:

- i. $y = f(x) = \frac{x}{(x+1)(x-1)}.$
- ii. $y = \frac{x^2}{(x+1)(x-1)}.$

Bolzano originally proved Theorem 2.2.11 by appeal to the Least Upper Bound Principle, which he proved by appeal to the convergence of Cauchy sequences. In standard Analysis, some completeness axiom must be assumed. My own preference is to choose the Least Upper Bound Principle itself.

2.2.14 Definitions Let $X \subseteq \mathbb{R}$ be a nonempty set of real numbers. A number B is an *upper bound* on X if no element of X exceeds B :

$$\forall x \in X (x \leq B).$$

A number B_0 is the *least upper bound* of X if B_0 is an upper bound on X and if $B_0 \leq B$ for any other upper bound B on X .

Note that B_0 need not be an element of X itself. For example, 1 is the least upper bound of the interval $(0, 1)$ yet does not belong to the interval. It does, however,

belong to the set of upper bounds on the interval, being itself the least element of this set.

The Least Upper Bound Principle is simply the assertion that bounded nonempty sets possess least upper bounds and optionally follows from or is taken as the completeness axiom of the reals.

2.2.15 Axiom (*Completeness Axiom; Least Upper Bound Principle*) Let $X \subseteq \mathbb{R}$ be a bounded nonempty set of real numbers. There is a least upper bound of X .

2.2.16 Remark One can also define the Greatest Lower Bound of a set bounded below and postulate a Greatest Lower Bound Principle. But this new Principle is redundant: if B_0 is a least upper bound for $\{-x \mid x \in X\}$, then $-B_0$ is the greatest lower bound of X .

From this axiom Theorems 2.2.11 and 2.2.12 are easily derived.

Proof of Theorem 2.2.11. Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous with $f(a) < 0 < f(b)$. Define

$$X = \{x \in [a, b] \mid \forall y \in [a, b](y \leq x \Rightarrow f(y) < 0)\}.$$

X is nonempty since $a \in X$ and it has b as an upper bound as $X \subseteq [a, b]$. By the Least Upper Bound Principle, X has a least upper bound c . The claim is that $f(c) = 0$. To prove this we need a simple lemma.

2.2.17 Lemma Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous and let $x \in [a, b]$ be such that $f(x) \neq 0$. There is a $\delta > 0$ such that, for all $y \in [a, b]$,

$$|y - x| < \delta \Rightarrow f(y) \neq 0 \text{ and } f(y) \text{ has the same sign as } f(x).$$

Proof Assume for the sake of definiteness that $f(x) > 0$. Let $\epsilon = f(x)/2$ and choose $\delta > 0$ such that, for all $y \in [a, b]$,

$$|y - x| < \delta \Rightarrow |f(y) - f(x)| < \epsilon,$$

i.e.,

$$\begin{aligned} |y - x| < \delta &\Rightarrow |f(y) - f(x)| < \frac{f(x)}{2} \\ &\Rightarrow -\frac{f(x)}{2} < f(y) - f(x) < \frac{f(x)}{2} \\ &\Rightarrow f(x) - \frac{f(x)}{2} < f(y) - f(x) + f(x) \\ &\Rightarrow f(y) > \frac{f(x)}{2} > 0. \end{aligned}$$

□

Returning to the proof of Theorem 2.2.11, assume $f(c) \neq 0$. Choose δ according to the Lemma such that for $x \in [a, b]$ $f(x)$ has the same sign as $f(c)$ whenever $x \in (c - \delta, c + \delta)$. If $f(c) < 0$, then $c < b$, and for $\eta = \min\{\delta, b - c\}$,

$$\begin{aligned} c < x \leq c + \frac{\eta}{2} &\Rightarrow x \in (c - \delta, c + \delta) \ \& \ x \in [a, b] \\ &\Rightarrow f(x) < 0 \\ &\Rightarrow c + \frac{\eta}{2} \in X \\ &\Rightarrow c \text{ is not an upper bound of } X. \end{aligned}$$

Similarly, $f(c)$ cannot be positive.

It follows that $f(c) = 0$ and, since neither $f(a)$ nor $f(b)$ is 0, c lies in the interior of the interval. \square

Proof of Theorem 2.2.12. I outline the proof for the existence of a maximum.

This is slightly more complicated than the proof of Theorem 2.2.11. First we show that $\{f(x) \mid x \in [a, b]\}$ is bounded by considering the set

$$X = \{x \in [a, b] \mid \exists B \forall y \in [a, b] (y \leq x \Rightarrow f(y) \leq B)\}.$$

X is again nonempty because $a \in X$ and it is again bounded by b because $X \subseteq [a, b]$. Let c be its least upper bound. The claim is that $c = b$. To prove this we need another simple lemma.

2.2.18 Lemma *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous and let $x \in [a, b]$. There is a $\delta > 0$ such that f is bounded on $(x - \delta, x + \delta)$.*

Proof Let $\epsilon > 0$ and choose δ such that, for all $y \in [a, b]$,

$$|y - x| < \delta \Rightarrow |f(y) - f(x)| < \epsilon.$$

For such y one has $f(y) < f(x) + \epsilon$, whence $f(x) + \epsilon$ is the bound sought. \square

Continuing the proof of the Theorem, note that $c \neq a$ since the Lemma yields a bound on $\{f(y) \mid a \leq y < a + \delta\}$ for some δ . In particular $a + \frac{\delta}{2} \in X$ and $a < a + \frac{\delta}{2} < c$.

If $c < b$, apply the Lemma to c : For any choice of ϵ , e.g., $\epsilon = 1$, there is a $\delta > 0$ such that, for all $y \in [a, b]$,

$$|y - c| < \delta \Rightarrow f(y) < B_1,$$

for some bound B_1 . Let $\eta = \min\{\delta, c - a, b - c\}$. Now $c - \frac{\eta}{2} < c$ and is an element of $[a, b]$, whence $c - \frac{\eta}{2} \in X$ and there is a bound B_0 such that

$$\forall x \in [a, b] \left(x \leq c - \frac{\eta}{2} \Rightarrow f(x) < B_0 \right).$$

But $\left[c - \frac{\eta}{2}, c + \frac{\eta}{2}\right] \subseteq (c - \delta, c + \delta)$, so

$$\forall x \in [a, b] \left(x \in \left[c - \frac{\eta}{2}, c + \frac{\eta}{2}\right] \Rightarrow f(x) < B_1 \right).$$

Hence, if $x \in [a, b]$ is less than or equal to $c + \frac{\eta}{2}$, $f(x) < \max\{B_0, B_1\}$. Thus $c + \frac{\eta}{2} \in X$, contrary to the assumption that c is an upper bound of X . Hence $c = b$ and f is bounded on $[a, b]$.

Now define a new function on $[a, b]$ by

$$g(x) = \text{least upper bound of } \{f(y) \mid a \leq y \leq x\}.$$

g is defined because $\{f(y) \mid a \leq y \leq x\} \subseteq f([a, b])$ which we have just shown is bounded.

2.2.19 Exercise Show that g is continuous on $[a, b]$.

Finally, let c be the least upper bound of

$$Y = \{x \in [a, b] \mid g(x) < g(b)\}.$$

2.2.20 Exercise Show that $f(c) = g(b) = \text{maximum value of } f \text{ on } [a, b]$.

With this Exercise the reader has finished the alternative proof of the Extreme Value Theorem. \square

2.2.21 Remark I am beginning to think it might have been a mistake not to have given the usual proof using Cauchy sequences and the Bolzano-Weierstrass Theorem. These proofs of Theorems 2.2.11 and 2.2.12 can be motivated by two words: *continuous induction*. The Least Upper Bound Principle is a continuous analogue to the *Least Number Principle* in arithmetic whereby every non-empty set of natural numbers contains a least element. Contrapositive to the Least Number Principle is the *Strong Form of Mathematical Induction*, also called the *Principle of Complete Induction*, which is equivalent to the usual *Principle of Mathematical Induction* one learns in pre-Calculus courses and applies quite often in the Calculus course. There is an analogous principle based on the Least Upper Bound Principle called continuous induction. According to it, to prove a property P holds for all $x \in [a, b]$ one has only to show that it holds in some interval $[a, b_0)$ and that, if it holds in any interval $[a, b_0)$, it must hold in some interval extending $[a, b_0)$, either $[a, b]$ if $b = b_0$ or $[a, b_1)$ for some $b_1 > b_0$ if $b \neq b_0$. I leave the proof of the principle of continuous induction by appeal to the Least Number Principle as a nice exercise and invite the reader to apply it to either replace or explain the proof of the Uniform Continuity Theorem, which is coming up, in terms of such an induction.

Continuous induction has not been traditionally presented in Analysis courses and I confess not to have recognised it initially on presenting these proofs. When writing the next section, on similarly proving a theorem called the Strictly Increasing

Function Theorem, I added a comment on continuous induction as a heuristic to take the edge off what appears to be an overly complicated proof. It did not occur to me to carefully formulate and apply such a principle. It was only after completing the book when Robert B. Burckel mentioned it with respect to the proof given in Chap. 3, Sect. 3.10.2 of the Heine-Borel Theorem that the principle fully entered my consciousness. It is a venerable principle, going back at least to Lebesgue who outlined its use in proving the Heine-Borel Theorem in 1904⁷⁹ and more explicitly to Y.R. Chao who in 1919⁸⁰ seems first to have explicitly formulated and named a variant of the principle. Since then it has been repeatedly forgotten and rediscovered. I am of two minds on the use of continuous induction in proving these results. On the one hand, it seems to be more elegant than the approach I have taken. On the other, the details of the inductive proof are pretty much the same as those I've given. The difference is that in the induction step of the inductive proof, one goes from $[a, b_0]$ to $[a, b_1]$ while I go from $[a, c]$ to $[a, b_1]$, where c is the demarcation between constant validity of P and occasional non-validity of P . In the inductive proof one concludes P holds in $[a, b]$ by induction, while I conclude c must be b and P holds in $[a, b]$. As I like arguing from first principles, I haven't replaced my proofs by the more elegant approach. I thus leave the conversion of my proofs to applications of continuous induction to the more ambitious reader. For the curious, but less ambitious, reader I note that Pete L. Clark has written a very nice survey⁸¹ of continuous induction with several applications and bibliographic references.

2.2.22 Theorem (Uniform Continuity Theorem) *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$. Then: f is uniformly continuous on $[a, b]$.*

Proof Let $\epsilon > 0$ be given. Define

$$X = \{x \in [a, b] \mid \exists \delta > 0 \forall y \in [a, x] \forall z \in [a, x] (|y - z| < \delta \Rightarrow |f(y) - f(z)| < \epsilon)\}.$$

Trivially $a \in X$ since $y \in [a, a]$ and $z \in [a, a]$ imply $y = z = a$, whence $|f(y) - f(z)| = |f(a) - f(a)| = 0 < \epsilon$. As usual, we have a lemma extending the possible boundary of X :

2.2.23 Lemma *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$ and let $x \in [a, b]$. For any $\epsilon > 0$ there is a $\delta > 0$ such that, for all $y, z \in [a, b]$,*

$$|y - x| < \delta \ \& \ |z - x| < \delta \Rightarrow |f(y) - f(z)| < \epsilon.$$

Proof Using the continuity of f at x , choose $\delta > 0$ so that for all $y \in [a, b]$

$$|y - x| < \delta \Rightarrow |f(y) - f(x)| < \frac{\epsilon}{2}.$$

⁷⁹H. Lebesgue, *Leçons sur l'intégration et la recherche des fonctions primitives*, Gauthier-Villars, Paris, 1904, p. 105.

⁸⁰Y.R. Chao, "A note on 'Continuous mathematical induction'", *Bulletin of the American Mathematical Society* 26 (1919), pp. 17–18.

⁸¹Pete L. Clark, "The instructor's guide to real induction", online at <http://arxiv.org/abs/1208.0973>.

If both $|y - x| < \delta$ and $|z - x| < \delta$, then

$$\begin{aligned} |f(y) - f(z)| &= |f(y) - f(x) + f(x) - f(z)| \\ &\leq |f(y) - f(x)| + |f(x) - f(z)| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

□

Let c be the least upper bound of X and suppose by way of contradiction that $c < b$. Choose δ_1 according to the Lemma so that, for all $y, z \in [a, b]$,

$$|y - c| < \delta_1 \text{ \& } |z - c| < \delta_1 \Rightarrow |f(y) - f(z)| < \epsilon.$$

By the Lemma we know that $c > a$, so choose $\eta = \min\{\delta_1, c - a, b - c\}$ so that

$$(c - \eta, c + \eta) \subseteq [a, b],$$

and choose δ_0 so small that, for all $y, z \in [a, b]$,

$$y \leq c - \frac{\eta}{2} \text{ \& } z \leq c - \frac{\eta}{2} \Rightarrow |f(y) - f(z)| < \epsilon.$$

Finally, let $\delta = \min\{\delta_0, \delta_1, \frac{\eta}{2}\}$ so that for $y, z \in [a, b]$, if $y, z \leq c + \frac{\eta}{2}$,

$$|y - z| < \delta \Rightarrow y, z < c - \frac{\eta}{2} \text{ or } y, z \in (c - \eta, c + \eta).$$

Either possibility yields $|f(y) - f(z)| < \epsilon$. Thus $c + \frac{\eta}{2} \in X$, and the usual contradiction occurs. □

As with the Extreme Value Theorem, the Uniform Continuity Theorem fails for open intervals.

2.2.24 Exercise Let $f : (a, b) \rightarrow \mathbb{R}$ be uniformly continuous.

- i. Show that f is bounded on (a, b) : there is a B such that, for all $x \in (a, b)$, $|f(x)| < B$.
- ii. Show that f can be extended to a continuous function $\bar{f} : [a, b] \rightarrow \mathbb{R}$.
- iii. Show that $g(x) = \sin(1/x)$ is not uniformly continuous on $(0, 1)$.

[Part i is fairly straightforward. A nonstandard proof of part ii is also straightforward. The easiest standard proof of part ii uses the convergence of Cauchy sequences as the Completeness Axiom. A proof based directly on the Least Upper Bound Principle, like that of the Extreme Value Theorem, is a little trickier.]

2.2.25 Remark The nonstandard proofs given earlier for the Intermediate Value Theorem and the Extreme Value Theorem are fairly intuitive and have a heuristic value even for the mathematician unfamiliar with Nonstandard Analysis. There is also a nonstandard proof of the Uniform Continuity Theorem. Indeed, it is even more trivial

than the proofs given for Theorems 2.2.11 and 2.2.12, but it is not as intuitive in one particular and hence is of limited heuristic value. This proof rests on three facts:

- (1) A function $f[a, b] \rightarrow \mathbb{R}$ is continuous on $[a, b]$ iff, for every real $x \in [a, b]$ and every infinitesimal η for which $x + \eta \in [a, b]$, the difference $f(x + \eta) - f(x)$ is infinitesimal;
- (2) f is uniformly continuous on $[a, b]$ iff, for every number $x \in [a, b]$, real or nonstandard, and every infinitesimal η for which $x + \eta \in [a, b]$, the difference $f(x + \eta) - f(x)$ is infinitesimal; and
- (3) every nonstandard $\alpha \in [a, b]$ is infinitesimally close to a standard real $r \in [a, b]$.

Now, (1) is the usual interpretation of continuity in terms of infinitesimals and most mathematicians would accept it without hesitation. Point (3) has not entered universal consciousness, but when one considers that α defines a Dedekind cut, it becomes quite plausible. Point (2), though easy enough to prove to one familiar with the logical setting in which the existence of nonstandard numbers is established, is not intuitively obvious. It is essentially Cauchy's definition of continuity and only becomes clear when considering non-uniformly continuous functions like

$$f(x) = \frac{1}{x} \text{ on } (0, 1)$$

or

$$g(x) = x^2 \text{ on } [0, \infty).$$

[In the latter case, if η is a positive infinitesimal, $1/\eta \in [0, \infty)$ is infinite and

$$g\left(\frac{1}{\eta} + \eta\right) = \left(\frac{1}{\eta} + \eta\right)^2 = \frac{1}{\eta^2} + 2\eta \cdot \frac{1}{\eta} + \eta^2 = g\left(\frac{1}{\eta}\right) + 2 + \eta^2,$$

and the difference $2 + \eta^2 > 2$ is not infinitesimal.] On a closed, bounded interval, however, by virtue of (3), the nonstandard conditions for continuity and uniform continuity given by (1) and (2) are equivalent.

As mentioned earlier, the Intermediate Value Theorem and the Extreme Value Theorem will have a direct bearing on our later discussion of the Mean Value Theorem, the Uniform Continuity Theorem less so. Thus our digression to discuss these important theorems is not a digression from our main path, but is a small diversion from our immediate goal of discussing continuous curves.

Together, the Intermediate Value Theorem and the Extreme Value Theorem tell us that the range of a continuous function $f : [a, b] \rightarrow \mathbb{R}$ is a closed and bounded interval $[m, M]$, where m, M are the minimum and maximum values, respectively, attained by f on $[a, b]$. But this tells us nothing yet about the graph of f , i.e., the curve itself. Nor does it apply directly to more general curves with continuous parametrisations γ . It remains to see how well our formally defined curves match our intuition. If I is an interval and $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ is a continuous function, what

can be said about the curve $C = \gamma(I)$? Is C perfect and cohesive in Cantor's sense? Is it connected? Is it one-dimensional?

Definitive answers to some of these questions cannot be given without formal definitions of the concepts involved. Cantor's definition of a perfect set as one that is closed under taking limits and which has no isolated points requires a formal definition of limit, which we haven't given yet. And we never did answer the question here of what constituted one-dimensionality. Fortunately, some of the questions can be answered using only intuitive notions of the properties involved.

2.2.26 Examples Not every continuous curve is perfect.

- i. The quadratrix (Fig. 2.1), defined by

$$\gamma(t) = \left\langle (1-t) \tan \frac{t\pi}{2}, 1-t \right\rangle, \quad t \in [0, 1),$$

is not perfect because the point $E = \langle 2/\pi, 0 \rangle$ is the limit of $\gamma(t)$ as $t \rightarrow 1$, but E is not on the curve.

- ii. The logarithmic spiral (Fig. 2.18), defined by

$$\gamma(t) = \langle ae^{bt} \cos t, ae^{bt} \sin t \rangle, \quad t \in (-\infty, \infty),$$

for fixed $a, b > 0$ is not perfect because it does not contain the point $\langle 0, 0 \rangle$ which is the limit of $\gamma(t)$ as $t \rightarrow -\infty$.

- iii. The point, defined by $\gamma(t) = \langle a, b \rangle$ for t on any interval, is not perfect because it consists of an isolated point.

On the other hand, it can be shown that for a closed, bounded interval $I = [a, b]$ and any nonconstant $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$, the curve $\gamma(I)$ is perfect.

It can also be shown that every continuous curve is cohesive in Cantor's sense. This is a fairly easy consequence of the Uniform Continuity Theorem generalised to functions $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$.

2.2.27 Theorem *Let I be an interval and $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ be continuous. The curve $C = \gamma(I)$ is cohesive.*

Proof Let $P, Q \in C$ and $\epsilon > 0$ be given. We have to show that there exist P_0, P_1, \dots, P_n such that $P_0 = P$, $P_n = Q$, and, for $i = 0, 1, \dots, n-1$, $\text{dist}(P_i, P_{i+1}) < \epsilon$.

Choose $a, b \in I$ such that $\gamma(a) = P$ and $\gamma(b) = Q$. Relabelling P and Q if necessary we can assume $a < b$. Restricted to $[a, b]$, γ is continuous and hence uniformly continuous (as the reader will show in the next exercise). Thus there is a $\delta > 0$ such that for all $x, y \in [a, b]$, if $|x - y| < \delta$ then $|\gamma(x) - \gamma(y)| < \epsilon$. Let n be so large that $n > (b - a)/\delta$, i.e., $\delta > (b - a)/n$. For $i = 0, 1, \dots, n$, let $P_i = \gamma(a + i(b - a)/n)$. Note that

$$a + (i+1) \frac{b-a}{n} - \left(a + i \frac{b-a}{n} \right) = \frac{b-a}{n} < \delta,$$

whence $\text{dist}(P_i, P_{i+1}) < \epsilon$. \square

2.2.28 Exercise Prove the version of the Uniform Continuity Theorem appealed to in the proof of Theorem 2.2.27: If $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$ is continuous on $[a, b]$, then γ is uniformly continuous on $[a, b]$.

By the cohesiveness of continuous curves we have ruled out their having large gaps, but not isolated gaps like that of Fig. 2.23. For this — at least for curves that look like curves — we have to show that $C = \gamma(I)$ is connected.

2.2.29 Theorem *Let I be an interval and $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ be continuous. The curve $C = \gamma(I)$ is connected.*

Proof Suppose C were not connected, i.e., suppose $C \subseteq U \cup V$, where U, V are disjoint open sets and there are points P, Q of C with $P \in U$ and $Q \in V$. Let a, b be such that $P = \gamma(a)$, $Q = \gamma(b)$ and assume without loss of generality that $a < b$. Define a function g on $[a, b]$ by

$$g(t) = \begin{cases} -1, & \gamma(t) \in U \\ 1, & \gamma(t) \in V. \end{cases}$$

The claim is that g is continuous on $[a, b]$. To see this, let $\epsilon > 0$ be given.

For any $t \in [a, b]$, $\gamma(t) \in U$ or $\gamma(t) \in V$. Consider the case $\gamma(t) \in U$. Because U is open, there is some $\epsilon_0 > 0$ such that, for all points R in the plane, if $\text{dist}(\gamma(t), R) < \epsilon_0$, then $R \in U$. Now, γ is continuous, so there is a $\delta > 0$ such that, for all $t' \in [a, b]$,

$$\begin{aligned} |t - t'| < \delta &\Rightarrow |\gamma(t) - \gamma(t')| < \epsilon_0 \\ &\Rightarrow \gamma(t') \in U \\ &\Rightarrow g(t') = -1 \\ &\Rightarrow |g(t) - g(t')| = |-1 - (-1)| = 0 < \epsilon. \end{aligned}$$

Similarly, if $\gamma(t) \in V$, we can find δ such that

$$|t - t'| < \delta \Rightarrow |g(t) - g(t')| < \epsilon.$$

Thus g is continuous on $[a, b]$ with $g(a) = -1 < 0 < 1 = g(b)$. By the Intermediate Value Theorem there is some c between a and b at which $g(c) = 0$, which is not the case. \square

The establishment of a precise definition of the continuity of a curve allows us to make rigorous the first step in the informal argument outlined in the Preface for the truth of the Mean Value Theorem. This is the proof that, if $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$ is a continuous parametrisation of a curve, there is a number $c \in (a, b)$ at which the distance from $\gamma(c)$ to the line passing through $\gamma(a)$ and $\gamma(b)$ is maximum.

2.2.30 Lemma *Let L be a line in $\mathbb{R} \times \mathbb{R}$ and define $d_L(x, y)$ to be the distance from the point $P = \langle x, y \rangle$ to L . Then: d_L is a continuous function.*

Proof We have only defined continuity for functions of one variable, but the definition for functions of two variables is the same: f is continuous at a point $P_0 \in \mathbb{R} \times \mathbb{R}$ if for any $\epsilon > 0$ there is a $\delta > 0$ such that, for all P in the domain of f ,

$$\text{dist}(P, P_0) < \delta \Rightarrow |f(P) - f(P_0)| < \epsilon.$$

The function d_L is in fact uniformly continuous. To see this, let $\epsilon > 0$ be given and consider two points $P, Q \in \mathbb{R} \times \mathbb{R}$ as in Fig. 2.27. Let d_p, d_q denote the respective distances of P, Q from L , $a = |d_L(P) - d_L(Q)| = |d_p - d_q|$, and, for the sake of definiteness, let $d_p \geq d_q$ as in the figure. To get $|d_p - d_q| = a < \epsilon$, note that $a^2 + b^2 = c^2$, whence $a^2 \leq c^2$ and $a \leq c = \text{dist}(P, Q)$. Thus, for $\delta = \epsilon$, we have

$$\text{dist}(P, Q) < \delta \Rightarrow |d_L(P) - d_L(Q)| = a \leq \text{dist}(P, Q) < \delta = \epsilon. \quad \square$$

2.2.31 Lemma Let $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$ be a continuous parametrisation of a curve and let L be the line passing through $\gamma(a)$ and $\gamma(b)$. Define, for $t \in [a, b]$,

$$d_\gamma(t) = d_L(\gamma(t)) = d_L(x(t), y(t)).$$

Then: d_γ attains a maximum on $[a, b]$.

Proof d_γ is continuous: Let $\epsilon > 0$ be given, $t_0 \in [a, b]$. Choose δ_1 so that, for all $t \in [a, b]$,

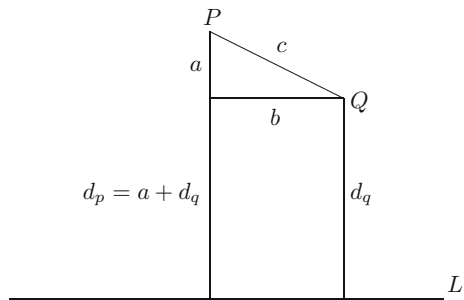
$$\text{dist}(\gamma(t), \gamma(t_0)) < \delta_1 \Rightarrow |d_L(\gamma(t)) - d_L(\gamma(t_0))| < \epsilon,$$

by the continuity of d_L ; and choose $\delta > 0$ so that, for all $t \in [a, b]$,

$$|t - t_0| < \delta \Rightarrow \text{dist}(\gamma(t), \gamma(t_0)) < \delta_1,$$

by the continuity of γ . The Extreme Value Theorem applies: For some $c \in [a, b]$ $d_\gamma(c)$ is maximum. Since $d_\gamma(a) = d_\gamma(b) = 0$, we can take c to be in the interior of the interval. \square

Fig. 2.27 Continuity of the distance function



To continue from here to prove the Mean Value Theorem, we need to guarantee that the curve $C = \gamma([a, b])$ has a tangent at all interior points $c \in (a, b)$, and that the tangent at a point of maximum distance on the curve from the line L passing through $\gamma(a)$ and $\gamma(b)$ is necessarily parallel to L . The first step in this is to define what a tangent line is. But before we do that there is one last property of curves to consider.

2.2.3 Peano's Space-Filling Curve

So far the only failure of the formal definition of a continuous curve has been the failure of curves defined on open or semi-open intervals to have endpoints as required by Cantor, but not so required by our intuition. It begins to look as if the formal definition of a continuous curve has captured the intuitive notion of curve. Alas, this is not the case. Whatever we mean by one-dimensionality, there are continuous curves that definitely are not one-dimensional. For, in 1890 Giuseppe Peano (1858–1932) proved the existence of continuous *space-filling curves*.

2.2.32 Theorem *The unit square $[0, 1] \times [0, 1]$ is a continuous curve, i.e., there is a continuous function $\gamma : [0, 1] \rightarrow \mathbb{R} \times \mathbb{R}$ such that $\gamma([0, 1]) = [0, 1] \times [0, 1]$.*

This came as a shock to mathematicians. It was known since Cantor that $[0, 1]$ and $[0, 1] \times [0, 1]$ could be put into one-to-one correspondence with each other, and thus there were “curves” γ for which $\gamma([0, 1])$ was $[0, 1] \times [0, 1]$. But, as shown by Eugen Netto, no such one-to-one correspondence could be continuous. The question of the nature and meaning of dimension arose. Peano's curve γ was not one-to-one and hence did not prove that $[0, 1]$ and $[0, 1] \times [0, 1]$ had the same dimension, and, indeed, it is not too difficult to prove by a simple appeal to connectivity that there cannot be a one-to-one continuous function γ such that $\gamma([0, 1]) = [0, 1] \times [0, 1]$: removing an interior point disconnects $[0, 1]$, but the removal of a single point will not disconnect $[0, 1] \times [0, 1]$.⁸² Still, the general problem of invariance of dimension had been raised and the result was only first proven in 1910 by Brouwer.

⁸²The one-to-one correspondence γ given at the end of the first section (p. 70, above) can be shown directly not to be continuous. The point $t_0 = .01$ is mapped by γ to the pair $\langle 0, .1 \rangle$. The points

$$t_n = .00 \underbrace{9999 \dots 99}_{2n}$$

can be chosen as close to t_0 as one wishes by choosing n large enough, yet

$$\begin{aligned} |\gamma(t_n) - \gamma(t_0)| &= |(\underbrace{.09 \dots 9}_n, \underbrace{.09 \dots 9}_n) - \langle 0, .1 \rangle| \\ &= \sqrt{(.09 \dots 9 - 0)^2 + (.09 \dots 9 - .1)^2} \\ &> .09 \dots 9 > .09, \end{aligned}$$

i.e., the points $\gamma(t_n)$ are bounded away from $\gamma(t_0)$.

Peano's curve was an important milestone in the history of mathematics, and, even though it will not be needed in our discussion of the Mean Value Theorem, it clearly belongs in any discussion of curves such as that we have been involving ourselves in. Thus, like many a less experienced writer, who doesn't know how to prune his creation, I have succumbed to the temptation to include a proof. The reader with no interest in this proof may safely skip ahead to the next section on page 79.

The construction of a space-filling curve is usually described geometrically⁸³ and the presentation of a rigorous proof can be a bit challenging combinatorially as one must translate the intuitive description into a sufficiently sharp analytic one. Moreover, such a proof relies on another deeper concept of analysis we haven't discussed yet in the present book, namely *uniform convergence*. Fortunately, Peano's original proof is much more elementary. And his paper is so simply written, with little extraneous material, that I have decided to present it in its entirety (after translation⁸⁴) here.

Peano's construction is very similar to the one-to-one correspondence given, but, instead of mapping $.r_0r_1r_2\dots$ to $\langle .r_0r_2\dots, .r_1r_3\dots \rangle$, he uses a clever device to occasionally replace r_n by its *complement*, which is like $9 - r_n$, but he uses the base 3 representation of numbers instead of the usual base 10 version for a reason explained at the end of his paper.

Without further ado, I present Peano's paper:

On a curve which fills an entire plane area.

by

G. PEANO in Turin.

In this note we determine two well-defined⁸⁵ and continuous functions x and y , of a (real) variable t , which, when t varies over the interval $(0, 1)$, takes each pair of values such that $0 \leq x \leq 1$, $0 \leq y \leq 1$. If, as is customary, one calls the set of points at which the coordinates are continuous functions of the variable, a *continuous curve* and one has thus an arc of a curve which passes through all the points of the square. Thus, being given an arc of a continuous curve, without making any other hypothesis, it is not always possible to contain it in an arbitrarily small area.

⁸³Cf., e.g., Hahn, *op. cit.*, pp. 85–87, or Bernard R. Gelbaum and John M.H. Olmsted, *Counterexamples in Analysis*, Holden-Day, Inc., San Francisco, 1964, pp. 133–134. The publication of Gelbaum and Olmsted has been taken over by Dover Publications and the book is still in print. The authors also cite a couple of variant constructions.

⁸⁴After making this translation, I was reminded by Ádám Besenyei that an excellent English translation can be found in: Hubert C. Kennedy (ed. and trans.), *Selected Works of Giuseppe Peano*, George Allen & Unwin Ltd, London, 1973. I bought a copy of this book decades ago and, being a logician, read some of the logical papers, storing the book on my general logic shelf. In my memory, the book was a selection of the logical papers of Peano and I thus neglected to consult it until receiving the reminder. Kennedy accompanies his translation with an excerpt from a later (1908) work of Peano in which a geometric construction is discussed.

⁸⁵Peano writes “uniformes”, which I take to mean “well-defined”. Kennedy translates this as “single-valued”, which is perhaps a more felicitous choice.

Let us adopt the number 3 as the number base; we call each of the numbers 0, 1, 2 a *cipher*⁸⁶; and consider an infinite sequence of ciphers a_1, a_2, a_3, \dots which we write⁸⁷

$$T = 0, a_1 a_2 a_3 \dots$$

(For the moment, T is only a sequence of ciphers).

If a is a cipher, designate by $\mathbf{k}a$ the cipher $2 - a$, *complementary* to a ; that is to say, put

$$\mathbf{k}0 = 2, \quad \mathbf{k}1 = 1, \quad \mathbf{k}2 = 0.$$

If $b = \mathbf{k}a$, we conclude $a = \mathbf{k}b$; we also have $\mathbf{k}a \equiv a \pmod{2}$.

Let $\mathbf{k}^n a$ denote the result of repeating the operation \mathbf{k} n times on a . If n is even, we have $\mathbf{k}^n a = a$; if n is odd, $\mathbf{k}^n a = \mathbf{k}a$. If $m \equiv n \pmod{2}$, we have $\mathbf{k}^m a = \mathbf{k}^n a$.

There correspond to the sequence T the two sequences

$$X = 0, b_1 b_2 b_3 \dots, \quad Y = 0, c_1 c_2 c_3 \dots,$$

where the ciphers b and c are given by the relations

$$\begin{aligned} b_1 &= a_1, \quad c_1 = \mathbf{k}^{a_1} a_2, \quad b_2 = \mathbf{k}^{a_2} a_3, \quad c_2 = \mathbf{k}^{a_1+a_3} a_4, \quad b_3 = \mathbf{k}^{a_2+a_4} a_5, \dots \\ b_n &= \mathbf{k}^{a_2+a_4+\dots+a_{2n-2}} a_{2n-1}, \quad c_n = \mathbf{k}^{a_1+a_3+\dots+a_{2n-1}} a_{2n}. \end{aligned}$$

Thus b_n , the n -th cipher of X , is equal to a_{2n-1} , the n -th cipher of odd rank of T , or to its complement, according as the sum $a_2 + \dots + a_{2n-2}$ of the ciphers of even rank which precede it is even or odd. Analogously for Y . We can thus write these relations in the form:

$$\begin{aligned} a_1 &= b_1, \quad a_2 = \mathbf{k}^{b_1} c_1, \quad a_3 = \mathbf{k}^{c_1} b_2, \quad a_4 = \mathbf{k}^{b_1+b_2} c_2, \dots, \\ a_{2n-1} &= \mathbf{k}^{c_1+c_2+\dots+c_{n-1}} b_n, \quad a_{2n} = \mathbf{k}^{b_1+b_2+\dots+b_n} c_n. \end{aligned}$$

If we are given the sequence T , then X and Y are consequently determined, and if we are given X and Y , then T is determined.

We call the *value* of the sequence T the quantity (analogous to a decimal number having the same notation)

$$t = \text{val. } T = \frac{a_1}{3} + \frac{a_2}{3^2} + \dots + \frac{a_n}{3^n} + \dots$$

To each sequence T corresponds a number t , and we have $0 \leq t \leq 1$. Conversely, the numbers t in the interval $(0, 1)$ are divided into two classes:

(α) The numbers, different from 0 and 1, which multiplied by a power of 3 yield an integer; they can be represented by two sequences, one

$$T = 0, a_1 a_2 \dots a_{n-1} a_n 222 \dots$$

where a_n is equal to 0 or to 1; the other

⁸⁶Kennedy uses the word “digit”, more in line with standard English usage. I tend to think of “digit” as referring to base 10 unless some modifier is added. In the present case this would result in “ternary digit”, which I didn’t like. So I stuck with the more literal “cipher”.

⁸⁷The European fashion is to use commas and periods in decimal representations where Americans use periods and commas, respectively. I have followed Peano more closely in these small details than Kennedy, for better or for worse.

$$T' = 0, a_1 a_2 \dots a_{n-1} a'_n 000 \dots$$

where $a'_n = a_n + 1$.

(β) The other numbers; they are represented by a unique sequence T .

Now the correspondence established between T and (X, Y) is such that if T and T' are two sequences of different form, but $\text{val. } T = \text{val. } T'$, and if X, Y are the sequences corresponding to T , and X', Y' those corresponding to T' , we have

$$\text{val. } X = \text{val. } X', \quad \text{val. } Y = \text{val. } Y'.$$

Indeed, consider the sequence

$$T = 0, a_1 a_2 \dots a_{2n-3} a_{2n-2} a_{2n-1} a_{2n} 2222 \dots$$

where a_{2n-1} and a_{2n} are not both equal to 2. This sequence represents a number of the class α . Letting

$$X = 0, b_1 b_2 \dots b_{n-1} b_n b_{n+1} \dots,$$

we have

$$b_n = \mathbf{k}^{a_2 + \dots + a_{2n-2}} a_{2n-1}, \quad b_{n+1} = b_{n+2} = \dots = \mathbf{k}^{a_2 + \dots + a_{2n-2} + a_{2n}} 2.$$

Letting T' be the other sequence coinciding with $\text{val. } T$,

$$T' = 0, a_1 a_2 \dots a_{2n-3} a_{2n-2} a'_{2n-1} a'_{2n} 0000 \dots$$

and

$$X' = 0, b_1 \dots b_{n-1} b'_n b'_{n+1} \dots$$

The first $2n - 2$ ciphers of T' coincide with those of T ; thus the first $n - 1$ ciphers of X' also coincide with those of X ; the others are determined by the relations

$$b'_n = \mathbf{k}^{a_2 + \dots + a_{2n-2}} a'_{2n-1}, \quad b'_{n+1} = b'_{n+2} = \dots = \mathbf{k}^{a_2 + \dots + a_{2n-2} + a'_{2n}} 0.$$

We distinguish two cases, according to whether $a_{2n} < 2$ or $a_{2n} = 2$.

If a_{2n} has the value 0 or 1, we have $a'_{2n} = a_{2n} + 1$, $a'_{2n-1} = a_{2n-1}$, $b'_n = b_n$,

$$a_2 + a_4 + \dots + a_{2n-2} + a'_{2n} = a_2 + \dots + a_{2n-2} + a_{2n} + 1,$$

whence

$$b'_{n+1} = b'_{n+2} = \dots = b_{n+1} = b_{n+2} = \dots = \mathbf{k}^{a_2 + \dots + a_{2n}} 2.$$

In this case the two sequences X and X' coincide in form and in value.

If $a_{2n} = 2$, we have $a_{2n-1} = 0$ or 1, $a'_{2n} = 0$, $a'_{2n-1} = a_{2n-1} + 1$, and on putting

$$s = a_2 + a_4 + \dots + a_{2n-2}$$

we have

$$\begin{aligned} b_n &= \mathbf{k}^s a_{2n-1}, \quad b_{n+1} = b_{n+2} = \dots = \mathbf{k}^s 2 \\ b'_n &= \mathbf{k}^s a'_{2n-1}, \quad b'_{n+1} = b'_{n+2} = \dots = \mathbf{k}^s 0. \end{aligned}$$

Now, seeing that $a'_{2n-1} = a_{2n-1} + 1$, the two fractions $0, a_{2n-2} 222 \dots$ and $0, a'_{2n-1} 000 \dots$ have the same value; and applying the same operation \mathbf{k}^s to the ciphers we obtain the two fractions $0, b_n b_{n+1} b_{n+2} \dots$ and $0, b'_n b'_{n+1} b'_{n+2} \dots$, which have likewise, as one can easily

see, the same value; thus the fractions X and X' , although of different forms, have the same value.

Analogously one proves that $\text{val. } Y = \text{val. } Y'$.

Thus if we set $x = \text{val. } X$, and $y = \text{val. } Y$, we conclude that x and y are two well-defined functions of the variable t over the interval $(0, 1)$. They are continuous; indeed if t tends to t_0 , terminating the first $2n$ ciphers in the development of t coincide with those of the development of t_0 , if t_0 is a β , or with those of one of the two developments of t_0 , if t_0 is an α ; and then the first n ciphers of x and y corresponding to t coincide with those of x and y corresponding to t_0 .

Finally to each pair (x, y) such that $0 \leq x \leq 1, 0 \leq y \leq 1$ corresponds at least a pair of sequences (X, Y) , which express the value; to (X, Y) corresponds a T , and to that its t ; thus we can always determine t in such a manner that the two functions x and y take arbitrarily given values in the interval $(0, 1)$.

One comes to the same conclusion if, in place of 3, one takes any odd number whatsoever for a number base. One can also take an even number for the base, but then it is necessary to establish a more complicated correspondence between T and (X, Y) .

One can form an arc of a continuous curve which entirely fills a cube. Make correspond to the fraction (in base 3)

$$T = 0, a_1 a_2 a_3 a_4 \dots$$

the fractions

$$X = 0, b_1 b_2 \dots, \quad Y = 0, c_1 c_2 \dots, \quad Z = 0, d_1 d_2 \dots$$

where

$$\begin{aligned} b_1 &= a_1, \quad c_1 = \mathbf{k}^{b_1} a_2, \quad d_1 = \mathbf{k}^{b_1+c_1} a_3, \quad b_2 = \mathbf{k}^{c_1+d_1} a_4, \quad \dots \\ b_n &= \mathbf{k}^{c_1+\dots+c_{n-1}+d_1+\dots+d_{n-1}} a_{3n-2}, \\ c_n &= \mathbf{k}^{d_1+\dots+d_{n-1}+b_1+\dots+b_n} a_{3n-1}, \\ d_n &= \mathbf{k}^{b_1+\dots+b_n+c_1+\dots+c_n} a_{3n}. \end{aligned}$$

One proves that $x = \text{val. } X, y = \text{val. } Y, z = \text{val. } Z$ are well-defined and continuous functions of the variable $t = \text{val. } T$; and if t varies between 0 and 1, x, y, z take on all the triples of values which satisfy $0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1$.

Mr. Cantor, (Crelle's Journal⁸⁸ 84, p. 242.) has demonstrated that one can establish a one-to-one correspondence between the points of a line and those of a surface. But Mr. Netto (Crelle's Journal 86, p. 263), and others have demonstrated that such a correspondence is necessarily discontinuous (see also G. Loria, *La definizione dello spazio ad n dimensioni ... secondo le ricerche di G. Cantor*, Giornale di Matematiche, 1877). In my note we demonstrate that one can establish well-definedness and continuity in one direction, that is to say, the points of the line can be corresponded to the points of a surface, in such a fashion that the points of the surface are a continuous function of the points of the line. But this correspondence is not at all one-to-one and onto, for to the points (x, y) of the square, if x and y are of [class] β , there rightly corresponds only one value of t , but if x , or y , or both of the two are of [class] α , the corresponding values of t are 2 or 4 in number.

⁸⁸More formally, *Journal für reine und angewandte Mathematik*. This journal was founded by August Crelle and is often called *Crelle's Journal* in his honour.

One has demonstrated that one can enclose an arc of a plane continuous curve in an arbitrarily small area:

- (1) If one of the functions, e.g., x coincides with the independent variable t ; one has then the theorem of integrability of continuous functions.
- (2) If the two functions are of bounded variation (Jordan, *Cours d'Analyse*, III, p. 599).
But, like the demonstration of the preceding example, this is not true if we suppose only the continuity of the functions x and y .

These x and y , continuous functions of the variable t , everywhere fail to have a derivative.

Turin, January 1890.⁸⁹

Following Peano's announcement of the existence of a space-filling curve, David Hilbert (1862–1943), a rising young German mathematician who would become the leading mathematician of the early decades of the 20th century, gave a geometric presentation of such a curve (see Fig. 2.28) at a meeting of the Society of German Natural Scientists and Physicians in Bremen. Felix Klein (1849–1925), the editor of the journal in which Peano's paper appeared, wrote to Hilbert on 23 November 1890:

Two additional wishes concerning the *Annalen*:

- (1) Could you not give us a note furnished with figures on the curve which you treated in Bremen. That you have *returned this matter to geometric intuition* is to me the essential thing. Indeed: I and probably many other mathematicians with me have not read the abstract presentation of Peano at all; however, with the figure, it becomes to me immediately accessible and I feel the whole importance of the matter.⁹⁰

Hilbert's note⁹¹ duly appeared in the 1891 volume of the same journal in which Peano's paper had appeared. Unlike Peano, Hilbert did not give a rigorous proof, just the geometric intuition behind his construction and a graphical display of the first few curves in his sequence. (See Fig. 2.28, for these curves.)

On the matter of the geometric approach, Peano's biographer tells us

In 1891 Hilbert published the first intuitive geometrical example of such a curve. His curve results as a limit of a sequence of curves. It is probable that Peano was led to the construction of his curve by such considerations. This is shown by his publication in the last edition (1908) of the *Formulario* of such a sequence of curves. He also had one of the curves in this sequence constructed on the terrace of the villa he purchased in the summer of 1891, where the curve showed up as black tiles on white. His 1890 publication, however, is purely analytic. Ugo Cassina has suggested that this is probably because he wished no doubt about the validity of his result and because he typically suppressed everything unnecessary to the goal set. "Besides," Cassina added, "the difficulty does not lie in becoming aware intuitively of the fact that a planar region can be conceived as the limit of a variable polygon, but in

⁸⁹G. Peano, "Sur une courbe, qui remplit toute une aire plane", *Mathematische Annalen* 36 (1890), pp. 157–160.

⁹⁰Günther Frei (ed.), *Der Briefwechsel David Hilbert – Felix Klein (1886–1918)*, Vandenhoeck & Ruprecht, Göttingen, 1985, pp. 70–71.

⁹¹David Hilbert, "Ueber die stetige Abbildung einer Linie auf ein Flächenstück", *Mathematische Annalen* 38 (1891), pp. 459–460.

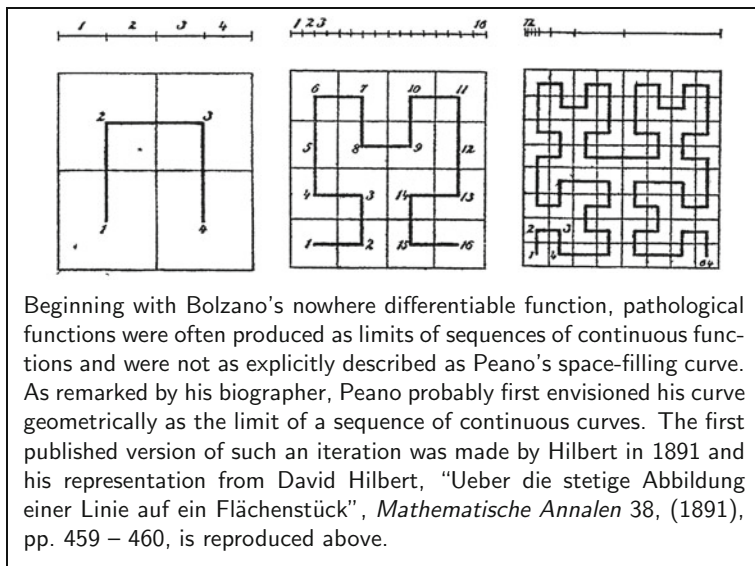


Fig. 2.28 Iterates of Hilbert's space filling curve

giving the explicit expression of the coordinates of a point of a planar region as *continuous* functions of a variable parameter in the interval."

Peano closes his note with the observation that the parametric functions are nowhere differentiable. We may add that this curve also has the property that, given any two points on the curve, the arc length between the two points is infinite. A curve often cited as having this property was invented by Helge von Koch in 1904.⁹²

Cassina, as quoted here, makes a good point. Geometrically, one presents the ranges of the curves $\gamma_0, \gamma_1, \gamma_2, \dots$, but not the parametrisations. It is not these ranges, but the functions $\gamma_0, \gamma_1, \gamma_2, \dots$ that tend to a limit. The same curves, under different parametrisations can have wildly different limits. If γ is the limit of Hilbert's parametrised curves and we define, for each γ_n a new parametrisation γ_n^* by

$$\gamma_n^* = \begin{cases} \gamma_n(nt), & 0 \leq t \leq \frac{1}{n+1} \\ \gamma_n\left(\frac{n}{n+1} + \frac{1}{n}\left(t - \frac{1}{n+1}\right)\right), & \frac{1}{n+1} < t \leq 1, \end{cases}$$

then the limit $\gamma^*(t) = \lim_{n \rightarrow \infty} \gamma_n^*(t)$ is discontinuous; indeed, one has

$$\gamma^*(t) = \begin{cases} \gamma(0), & t = 0 \\ \gamma(1), & t \neq 0. \end{cases}$$

⁹²Hubert C. Kennedy, *Peano, Life and Works of Giuseppe Peano*, D. Reidel Publishing Company, Dordrecht, 1980, p. 32.

The range of γ^* thus consists, not of the whole unit square, but of the two points $\gamma(0) = \langle 0, 0 \rangle$ and $\gamma(1) = \langle 1, 0 \rangle$.

An actual proof that Hilbert's curve, like Peano's, actually covers the unit square requires a careful presentation of the parametrisations, proof that the functions $\gamma_0, \gamma_1, \gamma_2, \dots$ converge uniformly to a continuous function γ , and then a proof that γ actually does map $[0, 1]$ onto $[0, 1] \times [0, 1]$. All of this is fairly routine and expositors do not always feel the need to present the details.^{93, 94}

Space-filling curves do not match our intuition of what a curve is or should be. Nonetheless, continuous curves like Peano's are accepted as curves in general mathematics. In Topology, one can work a bit harder and define one-dimensionality to refine the formal concept of curve even further, but in less specialised areas of mathematics, such as our discussion, a simpler refinement is often more useful — this is the notion of a *smooth curve*, which will be the topic of the next section.

2.3 Smooth Curves

2.3.1 Traditional Views of Tangents

We have already remarked in the Preface that it is not at all obvious what one should mean by the tangent to a curve. Euclid defines the tangent to a circle in Book III of *The Elements* as follows:

2. A straight line is said to **touch a circle** which, meeting the circle and being produced, does not cut the circle.⁹⁵

Heath offers little elucidation other than that there is a distinction between “meeting” the circle and “touching” it, and that the distinction was used by later geometers. The most important of these are Apollonius who determined the tangents to all the conic sections and Archimedes who found a single tangent to his spiral. A cursory

⁹³Whence, of course, follows Klein's preference for Hilbert's geometric presentation.

⁹⁴Two examples are the paper of Hahn and the book of Gelbaum and Olmsted cited in footnote 83 a few pages back. Hahn accompanies the pictures of some of the curves in Hilbert's sequence with the announcement, “It is now possible to give a rigorous proof that the successive motions considered here approach without limit a definite course, or curve, that takes the moving point through all the points of the large square in unit time”. Gelbaum and Olmsted give the parametrisation, but leave the details that the limit is a continuous function and that it fills the square as an exercise to the reader. E. Hairer and G. Wanner, *Analysis by Its History*, Springer-Verlag New York, Inc., New York, 1996, pp. 289–290 repeat Hilbert's graphical presentation and give the parametric representation for a more general construction, proving the continuity of the limit, but leaving unproven the more intuitive fact that the range of the function is the entire square. They also present Peano's construction geometrically as an exercise on page 298. A cursory check of my personal library found no fuller proof for the geometrical construction. Indeed, most of my textbooks on Analysis do not even mention the result.

⁹⁵Heath, *Elements*, *op. cit.*, vol. 2, p. 2.

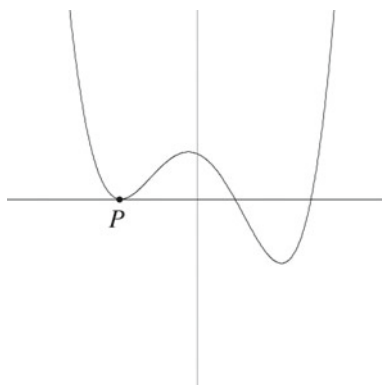


Fig. 2.29 A quartic

inspection of these later works reveals no new definition of “touching”, but Heath comments on the next definition in Euclid:

3. **Circles** are said to **touch one another** which, meeting one another, do not cut one another.⁹⁶

About this Heath says

Todhunter remarks that different opinions have been held as to what is, or should be, included in this definition, one opinion being that it only means that the circles do not cut in the neighbourhood of the point of contact, and that it must be shown that they do not cut elsewhere, while another opinion is that the definition means that the circles do not cut at all. Todhunter thinks the latter opinion correct. I do not think this is proved; and I prefer to read the definition as meaning simply that the circles meet at a point but do not cut *at that point*.⁹⁷

We have already seen a failing (see Figs. 1.5 and 1.6 of the Preface) of the definition of the tangent line as one which intersects the curve at only one point without crossing it, namely curves with points through which infinitely many tangents can pass. Figs. 2.29 and 2.30, give two more examples. In Fig. 2.29 the x -axis is clearly a tangent at P and yet clearly “cuts” the curve — but does not cut the curve *at* P . In Fig. 2.30, however, the x -axis does cross the curve at P , and yet one would like to consider this axis a tangent there for kinematic reasons: a particle travelling along the curve and allowed to go “off on a tangent” at P would follow the axis.

Another problematic curve is given in Fig. 2.31. Here I have simply taken a parabola, split it in two at the vertex, moved half of it over, and connected the two halves with a straight line. It is obvious that the straight line is the tangent at all points of the curve lying on it, yet it does not satisfy the usual condition of “touching” the curve. It, in fact, *coincides* with the curve for an entire interval, without “cutting” the curve.

⁹⁶*Ibid.*.

⁹⁷*Ibid.*, p. 3.

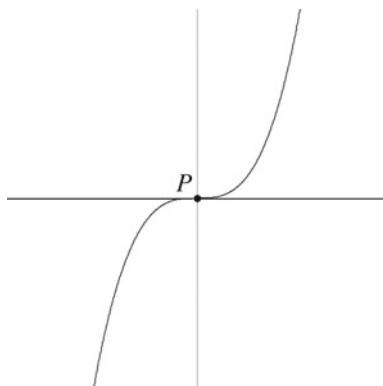


Fig. 2.30 A cubic

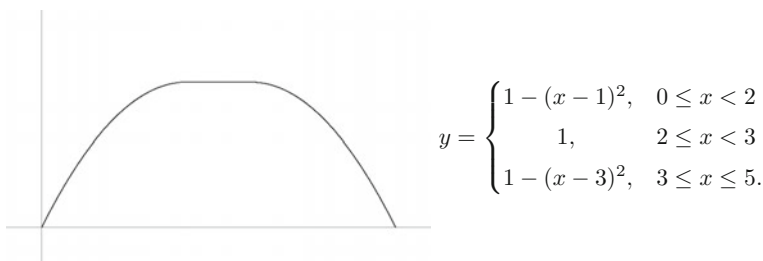


Fig. 2.31 A curve partially coinciding with a tangent

An even more problematic curve is Darboux's function graphed in Fig. 2.26. Again one would like to think of the x -axis as a horizontal tangent to the curve at the origin. It crosses the curve infinitely many times in any neighbourhood of the origin, but does it "cut" the curve at the origin? Put differently, does Darboux's curve *cross* the x -axis at the origin or just touch it there?

The intuitive geometric description of the tangent is too vague to determine definitively the tangency or non-tangency of lines in these questionable cases. In the more clear-cut ones, however, they are not totally useless. We can tentatively define the *Greek tangent* to a curve C at a point P on C to be the unique line, if it exists, which passes through P without crossing the curve. Further defining a continuous curve to be *smooth in the Greek sense* if every point other than an end point has a Greek tangent, we can almost get an easy proof of the Mean Value Theorem for curves that are smooth in the Greek sense. For, let $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$ be a continuous curve smooth in the Greek sense, and let $c \in (a, b)$ be a point on C , as given by Lemma 2.2.31, of maximum distance from the line connecting $\gamma(a)$ and $\gamma(b)$. The line through $\gamma(c)$ parallel to the line connecting $\gamma(a)$ and $\gamma(b)$ cannot cross the curve

at $\gamma(c)$ as those points of the curve on one side of the parallel are farther from the line connecting $\gamma(a)$ and $\gamma(b)$ than is $\gamma(c)$.^{98, 99}

In *La Géométrie*, Descartes introduced another definition of tangent and used it to determine the tangents to certain curves using geometric intuition and algebraic calculation. The geometric intuition is that if a circle and a line touch a curve at a given point P , the line is then simultaneously tangent to the circle and the curve at P . Hence the tangent to the curve at P is perpendicular to the normal to the circle. Descartes was rather pleased with himself over this method:

Finally, all other properties of curves depend only on the angles which these curves make with other lines. But the angle formed by two intersecting curves can be as easily measured as the angle formed by two straight lines, provided that a straight line can be drawn making right angles with one of these curves at its point of intersection with the other. This is my reason for believing that I shall have given here a sufficient introduction to the study of curves when I have given a general method of drawing a straight line making right angles with a curve at an arbitrarily chosen point upon it. And I dare say that this is not only the most useful and most general problem in geometry that I know, but even that I have ever desired to know.¹⁰⁰

The last sentence would no doubt be considered an exaggeration no matter what it referred to, but his excitement was understandable.

2.3.1 Remark Before leaving Descartes, note that the definition of the tangent as the line perpendicular to the normal of a circle touching the curve at a given point has its uses. Consider the cubic of Fig. 2.30: The circles of radius $1/2$ centred at the points $\langle 0, 1/2 \rangle$ and $\langle 0, -1/2 \rangle$ each touch the curve at P and have the x -axis as their tangents. It is easy to see that the cubic $y = x^3$ and, for example, the circle $x^2 + (y - 1/2)^2 = (1/2)^2$ have $P = \langle x, y \rangle = \langle 0, 0 \rangle$ in common. Verifying that they have no other common root requires a little work, but is not too hard:

$$\begin{aligned} x^2 + \left(y - \frac{1}{2}\right)^2 &= \frac{1}{4} \Rightarrow x^2 + y^2 - y = 0 \\ &\Rightarrow x^2 + (x^3)^2 - x^3 = 0, \text{ for } y = x^3 \\ &\Rightarrow x^6 - x^3 + x^2 = 0. \end{aligned} \tag{2.21}$$

(2.21) points to a double root at the origin. Eliminating this root results in the equation,

$$x^4 - x + 1 = 0. \tag{2.22}$$

⁹⁸I say this is “almost” a proof because we have not defined precisely what is meant by “crossing”. In algebraic terms we note that a line $Ax + By = C$ partitions the plane into three disjoint sets according as $Ax + By$ is $< C$, $= C$, or $> C$. A line may be said to cross the curve C at $P = \langle \alpha, \beta \rangle$ if $A\alpha + B\beta = C$ and in any neighbourhood of P there are points of the curve in each of the sets $\{\langle x, y \rangle \mid Ax + By < C\}$ and $\{\langle x, y \rangle \mid Ax + By > C\}$. Can we give a precise, purely geometric definition of the notion? How about the notion of two curves crossing each other?

⁹⁹Another problem is: how can we tell algebraically or analytically that a curve given by a continuous parametrisation γ is smooth in the Greek sense?.

¹⁰⁰Descartes, *op. cit.*, p. 95.

Three ways of verifying this to have no real roots come to mind. First, one can graph it on one's pocket calculator and notice that the graph does not appear to touch the x -axis.

Today we would probably use the Calculus to verify that our calculator isn't lying and differentiate: if $f(x) = x^4 - x + 1$ as in (2.22), then $f'(x) = 4x^3 - 1$ has the unique root $x = \sqrt[3]{1/4}$. $f''(x) = 12x^2$ is positive there, whence $f(\sqrt[3]{1/4})$ is a minimum and it happens to be positive.

Descartes was pre-Calculus, so he would have had to devise other methods. This he did in the third part of *La Géométrie*, wherein he enunciated *Descartes's Rule of Signs*: A polynomial $P(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ cannot have more positive roots than the number of sign changes in the sequence $a_n, a_{n-1}, \dots, a_1, a_0$. In the present case, this sequence is 1, 0, -1, 0, 1 and has only two sign changes, hence has at most two positive real roots. And $f(x)$ cannot have any negative roots since $f(-x) = x^4 + x + 1 > 1$ for $x > 0$. Moreover, $f(y+1) = y^4 + 4y^3 + 6y^2 + 3y + 1$ has no sign changes, whence no roots for $y > 0$, i.e., $f(x)$ has no roots for $x = y + 1 > 0 + 1 = 1$. Thus we have restricted any possible real root to the interval $(0, 1)$. And now the simplest thing to do is to make an *ad hoc* observation:

$$\begin{aligned} x \in (0, 1) &\Rightarrow 0 < x < 1 \text{ \& } 0 < 1 - x^3 < 1 \\ &\Rightarrow 0 < x(1 - x^3) < 1 \\ &\Rightarrow x - x^4 = x(1 - x^3) < 1 \\ &\Rightarrow 0 < 1 - (x - x^4) = x^4 - x + 1. \end{aligned}$$

2.3.2 Exercise Show that the circle of radius $1/2$ centred at $\langle 0, 1/2 \rangle$ intersects the parabola $y = x^2$ in only one point. Draw the same conclusion for this circle and Darboux's curve from Fig. 2.26. Conclude that Descartes should have accepted the x -axis as the tangent to Darboux's curve at $\langle 0, 0 \rangle$ had he been aware of the curve.

2.3.3 Exercise Show that the following circles all meet the curve $y = \sqrt[3]{|x|}$ at $\langle 0, 0 \rangle$ and nowhere else:

- i. $\left(x - \frac{1}{2}\right)^2 + y^2 = \frac{1}{4}$
- ii. $x^2 + (y + 1)^2 = 1$
- iii. $(x + 1)^2 + (y + 1)^2 = 2$.

What tangent lines do they suggest? Should the multiplicity of candidates for being the tangent line tell us that this curve has no tangent? Can you make a case for claiming that ii and iii cross the curve at $\langle 0, 0 \rangle$ and thus that the y -axis constitutes a Cartesian tangent to the curve? (Consider the curve $y^6 - x^2 = 0$.)

The implicit Cartesian definition of the tangent as the line perpendicular to the normal of a circle meeting the curve in only one point thus has something to offer. If we don't share his excitement today it is because the method can be computationally

horrendous and we now have much simpler methods. The more adventurous reader is referred elsewhere¹⁰¹ for the details of his method.

In January of 1638 Descartes received a letter from Fermat in which the latter laid out his method for finding maxima and minima, and showed how to use his own method for finding tangent lines. Fermat explained the procedure, but not the rationale:

We will express the maximum or minimum quantity in terms of a , by means of terms of any degree. We will then substitute $a + e$ for the primitive unknown a , and express the maximum or minimum quantity in terms containing a and e to any degree. We will *ad-equate*,¹⁰² to speak like Diophantus, the two expressions of the maximum or minimum quantity, and we will remove from them the terms common to both sides. Having done this, it will be found that on both sides, all the terms will involve e or a power of e . We will divide all the terms by e , or by a higher power of e , so that on at least one of the sides, e will disappear entirely. We will then eliminate all the terms where e (or one of its powers) still exists, and we will consider the others equal, or if nothing remains on one of the sides, we will equate the added terms with the subtracted terms, which comes to be the same. Solving this last equation will give the value of a , which will lead to the maximum or the minimum, in the original expression.¹⁰³

Translated into mathematical terms, given an expression $f(x)$, Fermat assumes f has a maximum at $x = a$ and $x = a + e$ and writes $f(a + e) \sim f(a)$.¹⁰⁴ For example, to find the maximum of $f(x) = x(b - x)$ he writes

$$(a + e)(b - (a + e)) \sim a(b - a) \\ ab - a^2 - ae + be - ae - e^2 \sim ab - a^2.$$

He then deletes the common terms

$$be - 2ae - e^2 \sim 0.$$

¹⁰¹*Ibid.*, pp. 95 ff. But see also Edwards, *op. cit.*, pp. 125–127.

¹⁰²The Latin original is “adæquentur”, later rendered into the French as “adégaler”. I suppose the most direct English translation would be “equate to”, but it is not clear that he really means “equate”. Thus historians of mathematics agree to keep the “ad”. The rest, i.e., what the term means, is hotly debated among the historians. Cf. Mikhail G. Katz, David M. Schaps, and Steven Schneider, “Almost equal: the method of adequality from Diophantus to Fermat and beyond”, [arXiv:1210.7750v1](https://arxiv.org/abs/1210.7750v1).

¹⁰³Pierre de Fermat, “Methodus ad disquirendum maximum & minimam”. Fermat did not publish the contents of this letter during his life, and it first appeared, in 1679 in Latin, in the *Varia opera mathematica* edited by his son Samuel de Fermat. A couple of centuries later, when it was translated into French for inclusion in the third volume (1896) of his collected works, *Œuvres de Fermat*, his antiquated notation was updated, the result being much more readable. In both these works, the letter to Descartes was accompanied by a number of later items on the method of maxima and minima. A translation into English of the modernised French translations of the letter to Descartes and its immediately following letter to Gilles Personne de Roberval (1602–1675) appeared in: Dirk Struik (ed.), *A Source Book in Mathematics, 1200–1800*, Harvard University Press, Cambridge (Mass.), 1969, pp. 222–227. The quotation reproduced above is from a more recent translation from the French edition by Jason Ross, which I found online. Ross translates all seven parts of Fermat’s method of maxima and minima.

¹⁰⁴ \sim is the symbol used in the French translation to stand for ad-equality.

The common terms are essentially $f(a)$ and he has basically formed $f(a + e) - f(a) \sim 0$. He now divides by e (or a higher power — whatever power is common to all the terms):

$$\frac{f(a + e) - f(a)}{e} = \frac{be - 2ae - e^2}{e} = b - 2a - e \sim 0. \quad (2.23)$$

And finally, he deletes all the remaining terms containing e , resulting in an equation

$$b - 2a = 0, \quad (2.24)$$

i.e., $a = b/2$ maximises $f(x) = x(b - x)$.

It is very hard not to recognise the difference quotient in (2.23) and the derivative $f'(a) = b - 2a$ in (2.24). That Fermat was not thinking in terms of the difference quotient (2.23) and its limit as $e \rightarrow 0$ becomes apparent when one continues to read his application to the construction of tangent lines, which strikes us today as roundabout. In later notes Fermat reveals that he is using a property of maxima and minima of continuous curves pointed out by Pappus:

...if one poses a question regarding given magnitudes which is satisfied in general by two points, then for the maximum or minimum values there would only be one point. It is for this reason that Pappus calls the smallest possible ratio for the question *minimum* and *singular* (that is, unique).¹⁰⁵

The maxima and minima are indeed unique and singular, as Pappus has said and as the ancients already knew... It follows that on both sides of the limit point, one could find an ambiguous equation; that the two ambiguous equations are then correlative, equal and alike.¹⁰⁶

The point is that if the maximum or minimum occurs at a , for the typical curve, $f(x)$ will not equal $f(a)$ nearby, but we will always have $f(x) < f(a)$ in the case of a maximum or $f(x) > f(a)$ for a minimum, provided $x \neq a$ is sufficiently close to a . Moreover, $f(x)$ will be paired with an $f(x')$ for x' on the opposite side of a . So he ad-equates the two expressions $f(a + e)$ and $f(a)$, removes common terms, and divides by e . Then

To find the maximum, we must equate the roots of the two equations...

Thus we must equate $a + e$ with a , whence $e = 0$...¹⁰⁷

Thus, Fermat is pursuing an algebraic solution justified by geometric considerations; he is not finding the slopes of secant lines and letting them rotate into the tangent's position. This would come later. When it did come, however, Fermat's method of finding maxima and minima would prove useful in proving the Mean Value Theorem.

¹⁰⁵“III. On the same method”, p. 5 of Ross, *op. cit.*

¹⁰⁶“IV. The method of maximum and minimum”, Ross, *op. cit.*, p. 7.

¹⁰⁷*Ibid.*, p. 9.

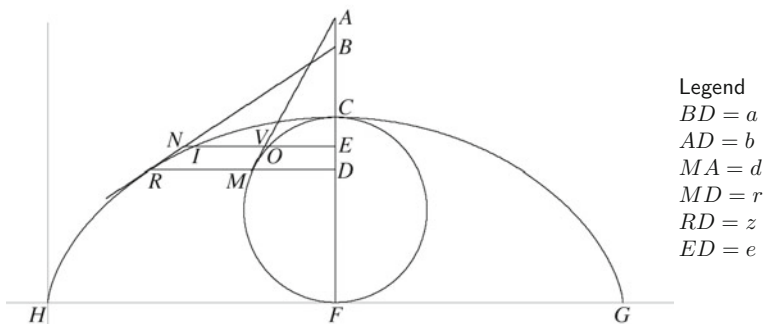


Fig. 2.32 The tangent of a cycloid

The impression to be had from Fermat's verbal description of his method of finding maxima and minima, as repeated above, is that it would apply only to polynomial expressions. This is not the case and, in a second note to bear the heading "On the same method", he explains how to find the tangents to the cissoid, the conchoid, the cycloid, and the quadratrix, giving the most detailed treatments for the cissoid and the cycloid. As we are already familiar with the cycloid, we shall consider how he determines the tangent to this curve.

Fermat begins with a cluttered diagram (Fig. 2.32). On this diagram, R represents the point at which we desire to draw the tangent line. RB is this tangent line. RD is the horizontal line through R . It intersects the circle inscribed in the centre of the cycloid at M and MA is the tangent to this circle at M . Finally, one draws a horizontal line from E to the tangent at a point E of distance e from D on the diameter CF of the circle. NE intersects the tangent RB at N , the cycloid at I , the circle's tangent MA at V , and the circle at O .

The tangent RB is determined by its slope $BD/RD = a/z$, which we hope to express in terms of known quantities. Given the cycloid, the circle, and the point R , these would be b, d, r , and z .

There are two pairs of similar triangles we can use to derive equations. From the pair RBD and NBE we have

$$\frac{NE}{RD} = \frac{BE}{BD}, \quad \text{i.e.,} \quad \frac{NE}{z} = \frac{a - e}{a}.$$

Thus

$$NE = \frac{za - ze}{a}. \quad (2.25)$$

And from VAE and MAD we have

$$\frac{EV}{MD} = \frac{AE}{AD}, \quad \text{i.e.,} \quad \frac{EV}{r} = \frac{b - e}{b}.$$

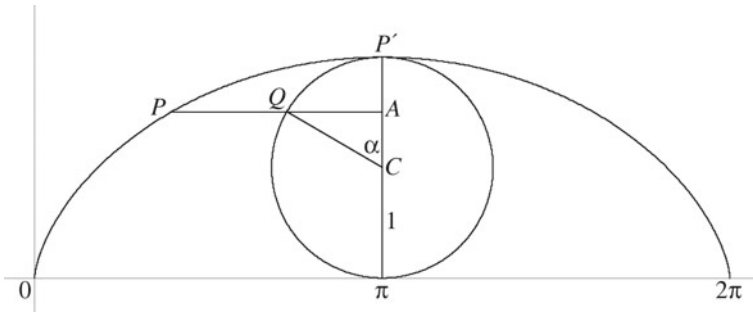


Fig. 2.33 A specific property

Thus

$$EV = \frac{rb - re}{b}. \quad (2.26)$$

At some point one must use some information about the cycloid itself. Fermat uses a “specific property of the curve”, one sufficiently non-obvious as to require isolation here as a lemma — or, better yet, as an exercise, which I state without loss of generality for a circle of radius 1:

2.3.4 Exercise Consider Fig. 2.33. Show that the segment PQ equals arc $QP' = \alpha$. [Hint: Use the parametrisation $P = \langle t - \sin t, 1 - \cos t \rangle$ to express PQ , QA , the height $1 + AC$ of A above the base line, and α in terms of t .]

In the present situation this means $RM = \text{arc } CM$ and $IO = \text{arc } OC$. Fermat now ad-equates

$$\begin{aligned} NE &= NO + OE \sim IO + OE \\ &\sim \text{arc } OC + OE \\ &\sim \text{arc } OC + EV. \end{aligned} \quad (2.27)$$

Now

$$\begin{aligned} \text{arc } OC &= \text{arc } CM - \text{arc } OM \\ &= RM - \text{arc } OM \\ &= RD - MD - \text{arc } OM \\ &= z - r - \text{arc } OM \\ &\sim z - r - VM \end{aligned} \quad (2.28)$$

as VM is very close to arc OM for e small, though Fermat omits this explanation.

Combining (2.25)–(2.28) we have

$$\begin{aligned}\frac{za - ze}{a} &\sim z - r - VM + \frac{rb - re}{b} \\ -\frac{z}{a}e &\sim -VM - \frac{r}{b}e.\end{aligned}\tag{2.29}$$

To determine VM , we use the similarity of VAE to MAD again:

$$\frac{AV}{AM} = \frac{AE}{AD}, \text{ i.e., } \frac{AV}{d} = \frac{b - e}{b},$$

whence

$$AV = \frac{db - de}{b}$$

and

$$VM = AM - AV = d - \frac{db - de}{b} = \frac{de}{b}.\tag{2.30}$$

Combining (2.29) and (2.30) we have

$$-\frac{z}{a}e \sim -\frac{d}{b}e - \frac{r}{b}e,$$

and dividing by $-e$ yields

$$\frac{z}{a} = \frac{d + r}{b}.$$

Fermat continues from here to construct the tangent line. Today we would do this by inverting the fractions to get

$$\text{slope of the tangent} = \frac{a}{z} = \frac{b}{d + r} = \frac{AD}{MA + MD}.$$

I think I have organised the details a little more clearly than Fermat, and it is definitely not as bad as the tangent determinations of Descartes, but it still isn't pretty. It is not the familiar use of the difference quotient in slightly disguised form as reading his introductory remark on his method of maxima and minima would have suggested to the modern reader. It is computational, but far from algorithmic. And:

2.3.5 Exercise What happens if the point D in Fig. 2.32 lies at or below the centre of the circle?

In a nice exposition of the history of the derivative, historian Judith Grabiner summarises Fermat's contribution:

...he did not explain why he could first divide by E (treating it as nonzero) and then throw it out (treating it as zero). Furthermore, he did not explain what he was doing as a special case of a more general concept, be it derivative, rate of change, or even slope of tangent. He did not even understand the relationship between his maximum-minimum method and the way one found tangents; in fact he followed his treatment of maxima and minima by saying

that the same method — that is, adding E , doing the algebra, then suppressing E — could be used to find tangents.

Though the considerations that led Fermat to his method may seem surprising to us, he did devise a method of finding extrema that worked, and it gave results that were far from trivial. For instance, Fermat applied his method to optics...

Though Fermat did not publish his method of maxima and minima, it became well known through correspondence and was widely used. After mathematicians had become familiar with a variety of examples, a pattern emerged from the solutions by Fermat's method to maximum-minimum problems.¹⁰⁸

By the 1650s there was any number of methods for finding tangents of and areas under curves. As regards tangents, two people in particular deserve mention — Johann Hudde (1628–1704) and René François de Sluse (1622–1685). Their actual methods need not be discussed here¹⁰⁹ as they are not directly related to the Mean Value Theorem, but they deserve to be noted because

...the principal significance of the rules of Sluse and Hudde lay in the fact that they provided general algorithms by which tangents to algebraic curves could be constructed in a routine manner. It was no longer necessary to resort to special devices adapted to particular curves, nor to give in every case a complete demonstration of the process. For these reasons, the rules of Sluse and Hudde were perhaps the first methods to exhibit fully the algorithmic approach that is a distinctive feature of the calculus....

The introduction in the 1650s of the algebraic rules of Hudde and Sluse was soon followed by infinitesimal derivations of these and similar methods. These newer derivations and methods owed more to the ideas of Fermat than those of Descartes, and involved the concept of a tangent line at the point P of a curve as the limiting position of a secant line PQ as Q approaches P along the curve.¹¹⁰

Judith Grabiner reports on the next stage in the development:

By the year 1660, both the computational and the geometric relationships between the problem of extrema and the problem of tangents were clearly understood; that is, a maximum was found by computing the slope of the tangent, according to the rule, and asking when it was zero. While in 1660 there was not yet a general concept of derivative, there was a general method for solving one type of geometric problem.¹¹¹

The two names to reckon with in the 1660s are Isaac Barrow (1630–1677) and Isaac Newton.

¹⁰⁸Judith V. Grabiner, "The changing concept of change: the derivative from Fermat to Weierstrass", *Mathematics Magazine* 56 (1983), pp. 195–206; here: p. 197. Grabiner is, of course, using " E " where Fermat used " e ".

¹⁰⁹Readable accounts of their contributions can be found in: Margaret E. Baron, *The Origins of the Infinitesimal Calculus*, Pergamon Press, Oxford, 1969, pp. 214–220; and Edwards, *op. cit.*, pp. 127–132.

¹¹⁰Edwards, *op. cit.*, pp. 131–132.

¹¹¹Grabiner, "Changing concept...", *op. cit.*, p. 198.

2.3.2 Isaac Barrow

In the mid-1660s Barrow lectured on Geometry at Cambridge, his lecture notes, the *Lectiones geometricæ*, first being published in Latin in 1670. His treatment of tangent and area problems is in the Greek geometric style, defining tangents as lines which touch the curve at single points. At the end of Lecture X, however, Barrow writes

Thus I have in some sort accomplished the chief Part of my proposed Design. As a Supplement to which, I shall annex our Method of determining Tangents by Calculation. Tho' I scarcely perceive the Use of so doing, considering the several Methods of this Nature now become common and published. I do this at least by the Advice of a Friend¹¹²; and indeed so much more willingly as it seems to be compendious and general with respect to what else I have handled. The Thing is thus.

Let AP, PM be right Lines given in Position (whereof PM cuts the proposed Curve in M,) and let MT touch the Curve in M, and cut the right Line AP in the Point T. Now to determine the length of the right Line PT, I suppose the Arch MN of the Curve to be indefinitely small, and draw the right Lines NQ, NR parallel to MP, AP; I call MP, m ; PT, t ; MR, a ; NR, e ; and give Names to other Lines useful to our purpose [See Barrow's Fig. 115 in Fig. 2.34.¹¹³], determin'd from the particular Nature of the Curve; and then compare MR, NR expressed by Calculation in an Equation, and by their means MP, PT themselves; observing the following Rules at the same Time.

1. I reject all the Terms in the Calculation, affected with¹¹⁴ any Power of a or e , or with the product of them; for these Terms will be equal to nothing.
2. After the Equation is formed, I reject all the Terms wherein are Letters expressing constant or known Quantities; or which are not affected with a , or e ; for these Terms brought over to one side of the Equation will be always equivalent to nothing.
3. I substitute a for m (MP), and t (PT) for e ; by which means the Quantity of PT will be found.

When any indefinitely small Particle of the Curve enters the Calculation, I substitute in its stead a Particle of the Curve properly taken; or any right Line equal to it, because of the indefinitely Smallness of the Part of the Curve.

All of this will appear more evident by the following Examples.

EXAMPLE I.¹¹⁵

Let ABH be a right Angle [As in Barrow's Fig. 116 in Fig. 2.34.], and let the Curve AMO be such, that drawing any right Line AK thro' A, cutting the right Line BH in K, and the Curve AMO in M, the Subtense AM may be equal to the Absciss BK; it is required to draw the Tangent (at M) of this Curve, or find the Value of the right Line PT.

Proceed according to the Directions above, and (drawing ANL) call AB, r , and AP, q . Then¹¹⁶ $AG = q - e$; also $QN = m - a$. Therefore it is $qq + ee - 2qe + mm + aa - 2ma = (AQq + QNq) = ANq$; that is, (rejecting according to the Rule above)

¹¹²Scholars have identified this friend as Newton, who helped prepare the work for publication.

¹¹³Note that the character that looks like an ℓ is the Q of the text.

¹¹⁴I.e., multiplied by.

¹¹⁵The following is a bit opaque and the reader may wish to skip ahead to the modern explanation following this quotation.

¹¹⁶The G here is clearly a misprint for Q.

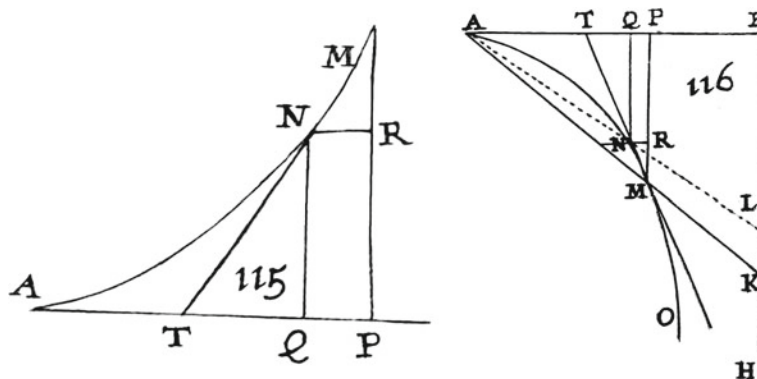


Fig. 2.34 Illustrations from Barrow

$qq - 2qe + mm - 2ma = \text{BL}q$. Again it is $\text{AQ}::\text{QN}::\text{AB}::\text{BL}$; that is $q - e : m - a :: r : \text{BL} = \frac{rm - ra}{q - e}$. Wherefore $\frac{rrmm + rraa - 2rrma}{qq + ee - 2qe} = \text{BL}q$. Or (casting away what is superfluous)¹¹⁷ $\frac{rrmm - 2rrmm}{qq - 2qe} = \text{BL}q = qq - 2qe + mm - 2ma$. Or $rrmm - 2rrma = q4 - 2q3e + qqmm - 2qqma - 2q3e + 4qqee - 2qmmme + 4qmae$, that is (rejecting as *per* Rule) $-2rrma = -4q3e - 2qqmma - 2qmmme$, or $rrma - qqma = 2q3e + qmmme$. Or at length substituting m for a , and t for e , it is¹¹⁸ $rrmm - qqmm = 2q3t - qmmmt$, or $\frac{rrmm - qqmm}{2q3 - qmm} = t = \text{PT}$.¹¹⁹

This excerpt can be understood with a great deal of patience and some guesswork about the notation, or with a little explanation. There are several obstacles for the modern reader. Barrow's book is geometrical in style and his analytic treatment here is not quite the modern presentation. He does not yet have the concept of function,

¹¹⁷There is a typo here: $-2rrmm$ should be $-2rrma$.

¹¹⁸There is a double typo here as the + between the two terms containing e accidentally changes to a -.

¹⁹Isaac Barrow (Edmund Stone trans.), *Geometrical Lectures: Explaining the Generation, Nature and Properties of Curve Lines*, London, 1735, pp. 171–175. This edition is available in facsimile online. The copy I downloaded, however, was very imperfectly done, some pages being repeated, and the fold-out plates scanned without being unfolded — whence not all the illustrations are available. One can, however, find all the illustrations online at ECHO (European Cultural Heritage Online) by searching, not for the *Lectiones geometricæ* of 1670, but for the larger work *Lectiones opticae & geometricæ* of 1674 in which the former is incorporated. Figure 2.34, combines screen captures of pieces of one of the plates (indexed by thumbnail 361 at ECHO) cleaned up with photo-retouching software.

A more recent annotated, but abridged, translation by J.M. Child, *The Geometrical Lectures of Isaac Barrow*, was published in 1916 by the Open Court Publishing Company (Chicago and London). This translation is available in several reprinted editions and can also be found online.

Struik, *op. cit.*, excerpts a couple of important passages from Barrow, including that portion of the above quotation omitting the Example.

Barrow illustrates his technique with five examples, of which I have cited the first, Child the fifth.

he does not emphasise the slope of the tangent, and his notation is archaic and not quite consistent. Finally, printing standards were not what they are today: the run-on structure of the final paragraph does nothing to help the reader along, but it is a positive advance of what had gone before, some lectures printed page after page in single unbroken paragraphs.

Historians of the Calculus explain Barrow's general remarks as follows. One has a curve $f(x, y) = 0$. To find the tangent at a point $\langle x, y \rangle$, one moves infinitesimally to a nearby point $\langle x - e, y - a \rangle$ on the curve:

$$f(x - e, y - a) = 0 = f(x, y).$$

One expands both sides and removes those terms containing no a or e (i.e., one subtracts $f(x, y)$ from both sides of the equation) (Barrow's Rule 2). Since a and e are infinitesimal, a^2 , e^2 , ae and all higher powers are infinitesimally small compared to a , e and can be deleted (Rule 1). This leaves an equation linear in a , e and one can solve for a/e or e/a . Referring to Barrow's Fig. 2.34, think of AP and MP as the axes, AP the y -axis (as $NR = e$), A on the positive side, and MP the x -axis, with M on the positive side. From the equation we determine e/a , but from the picture¹²⁰ we know the triangles TMP and NMR are adequately¹²¹ similar, whence

$$\frac{e}{a} \sim \frac{NR}{MR} \sim \frac{TP}{MP} \sim \frac{TP}{m}.$$

We know the ratio e/a and m , whence we know $TP = TP$ and can draw the tangent line. Barrow does not explicitly take the ratio here, but equivalently replaces e and a by $t = TP$ and m , respectively, in the linear equation. This is an improvement on Fermat in that he explicitly appeals to the infinitesimal nature of a and e in applying Rule 1 and implicitly appeals to the ad-equality of the ratios e/a and t/m in the application of Rule 3.

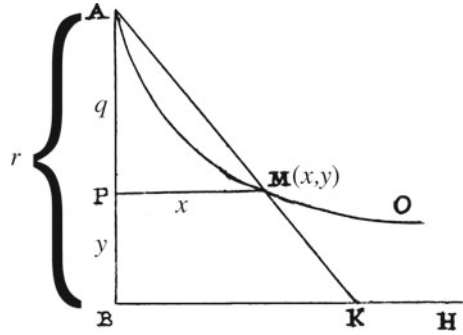
The example illustrates his method nicely if not completely clearly. First, there is no function f . The curve is defined as a locus and he derives an equation, but not the equation of the curve. The notation needs to be explained, and the equations laid out more readably.

The locus is described as follows. One takes a line BH and a point A not on BH but on the perpendicular to BH passing through B. Think of BH as the x -axis and AB as the y -axis, with B denoting the origin. (See Fig. 2.35, for a more familiar orientation.) The curve in question is the locus of points $M = \langle x, y \rangle$ such that when one extends the line AM to meet BH in a point K, then $AM = BK$. Where we would

¹²⁰His picture is imperfect here. The line TN is supposed to be tangent to the curve at M, not at N. The disposition of T, N, and M in his Fig. 2.35 is slightly better in that the tangent passes through M. Whether N lies on the curve or the tangent, however, is not discernible from these pictures. Fermat's Fig. 2.32, separating N from I and V from O is clearer in this respect. Barrow himself did better in his later Fig. 2.36 — cf. p. 96. Indeed, Struik reproduces Fig. 2.36 in place of Fig. 2.34 in his excerpt cited in the preceding footnote.

¹²¹See the preceding footnote.

Fig. 2.35 Simplified
Fig. 116 of Barrow



use x and y to determine the equation of the curve AMO, Barrow sets $AB = r$, $PM = m$, and $AP = q$. r is a constant, $q = r - y$ is a variable, as is $m = x$. Thus his equation is in the variables q, m with unspecified constant r .

To obtain an equation $f(q, m) = 0$ for the curve, note first that

$$AM = \sqrt{q^2 + x^2} = \sqrt{q^2 + m^2}. \quad (2.31)$$

Also note that, by the similarity of the triangles PAM and BAK,

$$\frac{BK}{PM} = \frac{AB}{AP},$$

i.e.,

$$\frac{BK}{m} = \frac{r}{q}, \text{ whence } BK = \frac{r}{q}m. \quad (2.32)$$

Plugging the values (2.31) and (2.32) into the locus equation $AM = BK$, we have

$$\frac{r}{q}m = \sqrt{q^2 + m^2}.$$

Thus

$$r^2m^2 = q^2(q^2 + m^2) = q^4 + q^2m^2,$$

and we have

$$f(q, m) = q^4 + q^2m^2 - r^2m^2 = 0$$

as the equation of the curve. If $\langle q - e, m - a \rangle$ is a nearby point on the curve, then

$$\begin{aligned} 0 &= f(q - e, m - a) = (q - e)^4 + (q - e)^2(m - a)^2 - r^2(m - a)^2 \\ &= q^4 - 4q^3e + 6q^2e^2 - 4qe^3 + e^4 + \\ &\quad (q^2 - 2qe + e^2)(m^2 - 2ma + a^2) - r^2(m^2 - 2ma + a^2). \end{aligned}$$

Applying Rule 1,¹²²

$$\begin{aligned} f(q - e, m - a) &= q^4 - 4q^3e + (q^2 - 2qe)(m^2 - 2ma) - r^2m^2 + 2r^2ma \\ &= q^4 - 4q^3e + q^2m^2 - 2q^2ma - 2qm^2e + 4qmea - \\ &\quad r^2m^2 + 2r^2ma \\ &= q^4 - 4q^3e + q^2m^2 - 2q^2ma - 2qm^2e - r^2m^2 + 2r^2ma, \end{aligned}$$

applying Rule 1 again. Subtracting $f(q, m)$ (i.e., applying Rule 2) yields

$$f(q - e, m - a) - f(q, m) = -4q^3e - 2q^2ma - 2qm^2e + 2r^2ma = 0,$$

whence

$$(2q^3 + qm^2)e = (r^2m - q^2m)a$$

and (in essence, Rule 3)

$$\frac{t}{m} = \frac{e}{a} = \frac{r^2m - q^2m}{2q^3 + qm^2}.$$

Thus,

$$t = \frac{r^2m^2 - q^2m^2}{2q^3 + qm^2},$$

which would agree with Barrow had not the plus sign in $2q^3e + qmme$ suddenly changed to a minus sign in the last sentence of the quotation.

Before discussing Barrow's derivation, the reader might want to check the result using ordinary Calculus:

2.3.6 Exercise Recalling that $m = x$, $q = r - y$, $f(q, m) = q^4 + q^2m^2 - r^2m^2$, verify that

$$\frac{dx}{dy} = \frac{dm}{-dq} = \frac{r^2m - q^2m}{2q^3 + qm^2}$$

by implicit differentiation.

Let us now consider Barrow's presentation of this example. He has the curve AMO and a point N on the curve infinitesimally close to M. Dropping the perpendicular NQ to AB, he has a right triangle AQN for which $AQ^2 + QN^2 = AN^2$. Moreover, $AN = BL$ by the defining property of the locus and

$$AQ^2 + QN^2 = AN^2 = BL^2,$$

telling us to read the q following AQ, QN, etc., in the equation¹²³

¹²²Perhaps we should use Fermat's adequality \sim here.

¹²³The parenthesis following AN q is merely a typographical error.

$$(AQq + QNq) = ANq = BLq$$

as squaring. This was a well-established practice at the time, and would give way to our modern exponential notation. A bit later Barrow approximates our modern notation in writing $q4e$ and $q3e$ for q^4e and q^3e , respectively. He nowhere uses 2 as an exponent, preferring qq , mm , etc. or AQq , etc. to represent the taking of squares. Exponential notation was still in its beginning stage of being used as an abbreviation, its functionality (indeed, functionality itself) not yet recognised.

Using the similarity of the triangles AQN and ABL he notes that

$$\frac{AQ}{QN} = \frac{AB}{BL},$$

in the quaint colonic notation $AQ:QN :: AB:BL$ which survives today in standardised tests of verbal skills:

$$\text{dog} : \text{puppy} :: \text{cat} : ?.$$

The point here is to express BL as a ratio as we did earlier for BK. In functional terms, he is deriving an algebraic expression for $f(q - e, m - a)$ directly without first determining $f(q, m)$ and then making the substitution as we did. The determination of $f(q, m)$ is implicit, however, in the application of Rule 2.

Barrow's second use of colons is the sort of thing that makes modern maths teachers cringe when they see it in their students' papers.

$$q - e : m - a :: r : BL = \frac{rm - ra}{q - e}$$

is to be read as

$$\frac{q - e}{m - a} = \frac{r}{BL} \text{ and therefore } BL = \frac{rm - ra}{q - e}.$$

The rest of Barrow's derivation is a straightforward algebraic computation augmented by Rules 1–3 until a homogeneous linear equation in a and e is established. The paragraph has more than its fair share of typographical errors, which do not enhance its readability.

The final step of replacing e by t and a by m would seem to be the most mysterious part of the procedure, suggesting to the modern reader the taking of the limit as $e \rightarrow t$ and $a \rightarrow m$. This, of course, is not the case, as t and m are fixed finite values while e and a are infinitesimal variables. Barrow is, as indicated earlier,¹²⁴ using the supposed similarity of the triangles NRM and TPM and the consequent equation of proportionality, $e/a = t/m$.

¹²⁴Cf. his comment following Rule 3 on page 91, above.

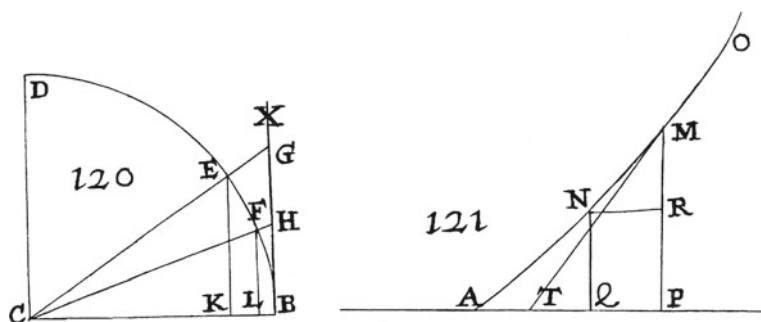


Fig. 2.36 Barrow's figures for tangent

The other examples of the algebraic determination of tangents given in the *Lectioes geometricae* are of somewhat more interesting curves. The first, which we have discussed here, is at first sight a strange little curve.

2.3.7 Exercise For the sake of definiteness choose r to be 2 in Barrow's example, so that

$$f(q, m) = q^4 + q^2 m^2 - 4m^2.$$

- Graph $f(q, m) = 0$ on the qm -plane.
- Define $g(x, y) = f(2 - y, x)$ and graph $g(x, y) = 0$ in the xy -plane.
- Writing $q = r \cos \theta$, $m = r \sin \theta$, express $f(q, m) = 0$ in polar coordinates and graph the resulting equation.

The result of the Exercise shows us that Barrow's Example I is a rotated and translated version of the polar curve $r = c \tan \theta$ for some constant c . His fifth example replaces r and θ by y and x : in other words, he shows the derivative to $y = c \tan x$ to be $y' = c \sec^2 x$.¹²⁵ In doing so, he introduced two figures — 120 and 121 reproduced in Fig. 2.36. The thing to notice is that his Fig. 121 is essentially the same as his Fig. 115 (in Fig. 2.34), albeit more carefully drawn: That N is on the curve and not on the tangent line is clearly evident in the new image; this was not at all evident in Fig. 115.¹²⁶

In the *Lectioes geometricae*, the illustrations are not presented in the text where they are referred to, but are collected in special folded sheets. Figures 115, 116, 120, and 121 are all on the same sheet in close proximity, with Fig. 115 situated directly above 120, which is directly to the left of 121. This effectively renders Fig. 115 redundant, a less well-drawn copy of 121.

The distinction between the point N on the curve and the point, say I , of intersection of the tangent line MT and the line NR , which was clear enough on Fermat's cluttered

¹²⁵The other curves for which he finds tangents are two versions of the *folium of Descartes*, with equations $x^3 + y^3 = c$ and $x^3 + y^3 = cxy$, and the quadratrix.

¹²⁶Cf. footnote 120.

diagram¹²⁷ is missing in Barrow's 115 and 116, but visible in Fig. 121. Like Fermat's reference to "ad-equate" in place of "equate", he was consciously assuming the difference between two quantities,

$$\frac{e}{a} = \frac{NR}{MR} \text{ and } \frac{IR}{MR} = \frac{TP}{MP} = \frac{t}{m},$$

to be negligible, i.e., infinitesimal. This makes NI/MR infinitesimal and NI thus to be an infinitesimal of higher order than NR and MR. This puts Rule 3, the identification of e/m with t/m , on par with Rule 1, the removal of higher order infinitesimals. The language of higher order infinitesimals would be developed by Leibniz. The importance of the higher infinitesimality of IN would emerge in the work of Lagrange.

A few parting words about Barrow might be in order. I begin with a quotation from Margaret Baron:

In conclusion it is perhaps worth saying once again that Barrow's *Geometrical Lectures* should be viewed, not as an isolated study, but as the culmination of all the seventeenth-century geometrical investigations leading to the calculus. In this context the work represents the most detailed and systematic treatment of these properties of curves such as tangents, arcs, areas and so on, which, in the hands of Newton and Leibniz, led so rapidly to the invention of the calculus. By the use of modern notation, it is, of course, possible to transform the geometrical results arrived at by Barrow into standard differentials and definite integrals and Child has drawn up a formidable array of results which he obtained by so doing. Moreover, Barrow was able to integrate the concepts of time and motion with those of space in the manner suggested by Torricelli, Galileo and Roberval, and thus to move nearer to Newton's fluxions.¹²⁸

Child's translation of Barrow's geometrical lectures was not undertaken so much to translate the work from Latin into English as to translate the geometric theorems presented therein into analytic terms and thereby prove that Barrow had pretty much invented the Calculus before Newton and Leibniz and to make the case that their works were highly dependent on Barrow's. Like any conclusion in the history of mathematics, this has been disputed by other historians.

Baron continues

Mathematical invention is a process of continuous change and development rather than something which takes place at a given point in time, but if it be considered necessary to draw a line between those mathematicians of the seventeenth century who "had the calculus" and those who "had not" the line would inevitably exclude Barrow on the grounds that he exhibited no calcular rules and used no specialised notation or symbolism. The claim made by Child that Barrow privately made use of notation, rules and symbols and that he turned these over to Newton whilst preferring to publish his own work in purely geometrical language, cannot be considered seriously. Barrow was a skilful geometer, not only in the purely formal

¹²⁷Figure 2.32. Fermat's and Barrow's labelling differ:

Barrow	M	N	R	P	T	—
Fermat	R	N	E	D	B	I

¹²⁸Baron, *op. cit.*, p. 251.

sense, but also in his intuitive appreciation, through the concepts of time and motion, of the properties of curves, tangents and areas. His approach to the study of curves was made possible by the acceptance, in his thinking processes, of Cavalierian indivisibles, and there is no evidence that he evolved, or indeed felt any need for, any kind of analytical procedure.¹²⁹

The infinitely small had been around for some time in the form of *indivisibles*, the exact nature of which varied from person to person. Bonaventura Cavalieri (1598–1647), mentioned by Baron, was particularly adept at using them, but his key contribution based on indivisibles was in finding areas and, as much fun as it is, it is not at first sight really relevant here and, indeed, Cavalieri’s infinitesimals are not particularly relevant to the present book. One aspect of his work, however, will be considered in the next chapter, in Sect. 3.2.5. For now, I must content myself with suggesting the reader look up some account of Cavalieri’s method.¹³⁰

2.3.3 *Transition to Newton and Leibniz*

Barrow, as Baron says, did not invent the Calculus. But perhaps one should draw two lines, on either side of Barrow separating the predecessors to Barrow — Fermat, Hudde, de Sluse, Torricelli, Roberval, Cavalieri, etc. — from Barrow and Barrow from his successors — Newton and Leibniz. Child did do the transformations referred to by Baron and found one can read into Barrow most of the rules which, analytically expressed, would be used by Newton and Leibniz in constructing the differentiation algorithm which turned the analytic *art* into a *calculus*, one so powerful it became *the Calculus*. Moreover, Barrow derived geometrically a version of the Fundamental Theorem of the Calculus by which the area and tangent problems are inverse to one another and, in the hands of Newton and Leibniz, the integral calculus became at least semi-algorithmic as well. Barrow, however, was a geometrician and, as evidenced by his remark on having had to be persuaded to include some examples of the analytic determination of tangents, was not interested in the analytic development of his results. This was where Newton and Leibniz came in. Historians debate on how much they owe to Barrow. Child, in his translation of Barrow’s geometrical lectures and a translation of Leibniz’s early mathematical manuscripts¹³¹ attempted to prove that they both owed almost all to Barrow’s work, but most historians consider this view extreme.

Grabiner explains the difference nicely:

¹²⁹*Ibid.*, pp. 251–252.

¹³⁰Excerpts from Cavalieri’s work can be found in: David Eugene Smith (ed.), *A Source Book in Mathematics*, 1929 (reprinted: Dover Publications, Inc., New York, 1959, pp. 605–609); Struik, *op. cit.*, pp. 209–219; and Stedall (*op. cit.*, pp. 62–65. Accounts can also be found in Edwards (*op. cit.*, pp. 104–109) and Baron (*op. cit.*, pp. 122–135). By far the most complete discussion in English however is Kirsti Andersen, “Cavalieri’s Method of Indivisibles”, *Archive for History of Exact Sciences* 31(1985), pp. 291–367.

¹³¹J.M. Child (ed. and trans.), *The Early Mathematical Manuscripts of Leibniz*, The Open Court Publishing Company, Chicago and London, 1920.

In the latter third of the seventeenth century, Newton and Leibniz, each independently, invented the calculus. By “inventing the calculus” I mean that they did three things. First, they took the wealth of methods that already existed for finding tangents, extrema, and areas, and they subsumed all these methods under the heading of two general concepts, the concepts which we now call **derivative** and **integral**. Second, Newton and Leibniz each worked out a notation which made it easy, almost automatic, to use these general concepts...Third, Newton and Leibniz each gave an argument to prove what we now call the Fundamental Theorem of the Calculus: the derivative and the integral are mutually inverse. Newton called our “derivative” a *fluxion* — a rate of flux or change; Leibniz saw the derivative as a ratio of infinitesimal differences and called it the *differential quotient*. But whatever terms were used, the concept of derivative was now embedded in a general subject — the calculus — and its relationship to the other basic concept, which Leibniz called the integral, was now understood.¹³²

Another thing Newton and Leibniz did, not nearly as successfully, was to provide justifications for the ad-equations. Barrow’s analytic determination of the tangent made a double advance on Fermat by explicitly introducing infinitesimals into the discussion, something we now recognise as the equivalent of taking limits, and by drawing attention to the *characteristic triangle* MTP or the “triangle” MNR from which we get the slope of the tangent. But he left unexplained why MNR could be taken as similar to MTP.

Once one has the characteristic triangle in mind I suppose it is inevitable to view the tangent line through a point P on a curve C as the limiting position of secant lines passing through P and a second point N on the curve as N nears P . Or, one might view the tangent as the secant line for N infinitesimally close to P . That the emphasis would fall on the slope of the line, i.e., on $\tan(\angle MTP)$ as opposed to its reciprocal $\cot(\angle MTP)$ as calculated by Barrow is perhaps a little less inevitable and may be attributable to the desire to represent the tangent line in functional form, $y = mx + b$, where m is the slope. Or, it may be due to the notion of rates of change and the habit of choosing $x(t) = t$ wherever possible. For whatever reason, before Newton and Leibniz any property of the tangent line was used to find it, and after Newton and Leibniz everyone calculated the slope of the tangent line via the difference quotient

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x},$$

for very small Δx . Exactly how Δx was equivalent to 0 was variously explained and not satisfactorily so until the 19th century.

2.3.4 Newton

Newton was the first to “invent” the Calculus. The story is well-known. He was studying in Cambridge when the plague arrived in England in 1665 and he retreated to his family farm in Woolsthorpe where he worked out the Calculus and Physics in

¹³²Grabiner, “Changing concept...”, *op. cit.*, p. 199.

one *annus mirabilis*. He then kept as quiet as possible about it for as long as possible. He did share a few results with colleagues, but it was decades before he published the Calculus in definitive Newtonian form. By then, of course, Leibniz had rediscovered everything and published it all.

A proper discussion of Newton can only be given by a dedicated Newton scholar. His publications do not reflect the time or even order of discovery. It is well-known, for example, that he used the Calculus to arrive at the results of his *Principia*, but there is no trace of this in the book, in which these results are established by classical geometrical methods. With respect to tangents, Baron informs us that

In the first stages of Newton's investigations into the properties of curved lines he relied mainly on Descartes' tangent method. He took over Descartes' characteristic symbolism...as well as the method of equal roots. This he incorporated with *Hudde's Rule*.¹³³

Once he began to develop the Calculus, however, he switched over to infinitesimals, and later to some vague notion of limit:

It is well known that at first the method employed by Newton involved fixed infinitesimals. But in the Introduction to the *Quadratura curvarum*, published in 1704, Newton aimed to develop his theory without the use of infinitely small quantities. Both in his *Principia* and *Quadratura curvarum*, he used "prime and ultimate ratios," which involve the concept of limits, though in a form differing from that adopted by mathematicians later. These prime and ultimate ratios do not contemplate primarily one constant which one variable approaches. The prime and ultimate ratios are ratios of two quantities just springing into being or else vanishing. Only secondarily does Newton, in applying his theory to finding the fluxion¹³⁴ of x^n , for example, consider in the right member of his equations what we would call the limit of a ratio. Newton was really considering the ratio of two quantities, each of which was approaching the limit zero, rather than the limit of one quantity that was the ratio of two quantities.¹³⁵

The works cited are the *Tractatus de quadratura curvarum* [*Treatise on the quadrature of curves*] (1704) and the *Philosophiæ naturalis principia mathematica* [*Mathematical principles of natural philosophy*] (1687), both usually referred to by the abbreviated titles given in the quotation. The *Quadratura curvarum* is Newton's most famous work on the Calculus and the *Principia*, of course, his most famous and important work.

Newton's first work on the Calculus to be disseminated was *De analysis per æquationes numeri terminorum infinitas* of 1669, not published until 1711. This dealt mainly with infinite series and term-by-term differentiation and integration and is of no importance for our present purposes. This was followed in 1671 by *De methodus fluxionum et serierum infinitarum*, first published in English translation by John Colson as *The Method of Fluxions and Infinite Series with its Application to the Geometry of CURVE-LINES* in 1736. The *Principia*, which could have been the

¹³³Baron, *op. cit.*, p. 257.

¹³⁴The fluxion is essentially the derivative of one of the variables with respect to time — more anon.

¹³⁵Florian Cajori, "Newton's fluxions", in: David Eugene Smith (ed.), *Sir Isaac Newton, 1727–1927; A Bicentenary Evaluation of His Work*, The Williams & Wilkins Company, Baltimore, 1928, p. 193.

first publication of the Calculus, was published in 1687, but Newton replaced the Calculus, which he had used to obtain his results, by classical geometrical reasoning. Thus, the *Quadratura curvarum*, written a few years later and tacked on to Newton's *Opticks* became in 1704 his first published account of his treatment of the Calculus.

English translations of these works are all available online or in any university library, and substantial excerpts are to be found in most source books. I quote from an early translation of the *De methodus fluxionum*:

Now those Quantities which I consider as gradually and indefinitely increasing, I shall hereafter call *Fluents*, or *Flowing Quantities*, and shall represent them by the final Letters of the Alphabet v, x, y , and z ; that I may distinguish them from other Quantities, which in Equations are to be consider'd as known and determinate, and which therefore are represented by the initial Letters a, b, c &c. And the Velocities by which every Fluent is increased by its generating Motion, (which I may call *Fluxions*, or simply Velocities or Celebrities,) I shall represent by the same Letters pointed thus $\dot{v}, \dot{x}, \dot{y}$, and \dot{z} . That is, for the Celerity of the Quantity v I shall put \dot{v} , and so for the Celebrities of the other Quantities x, y , and z , I shall put \dot{x}, \dot{y} , and \dot{z} respectively.¹³⁶

We can think of the fluents v, x, y , and z as functions $v(t), x(t), y(t)$ and $z(t)$ of time t , and their fluxions as their instantaneous rates of change: $\dot{v} = dv/dt$, $\dot{x} = dx/dt$, $\dot{y} = dy/dt$, and $\dot{z} = dz/dt$. Having introduced such, Newton next considers several problems, offering examples of each type with their solutions, and then follows up with a demonstration of the solution:

PROB. I.

The Relation of the Flowing Quantities to one another being given, to determine the Relation of their Fluxions.

SOLUTION.

1. Dispose the Equation, by which the given Relation is express'd, according to the Dimensions of some one of its flowing Quantities, suppose x , and multiply its Terms by any Arithmetical Progression, and then by $\frac{\dot{x}}{x}$. And perform this Operation separately for every one of the flowing Quantities. Then make the Sum of the Products equal to nothing, and you will have the Equation required.
2. EXAMPLE I. If the Relation of the flowing Quantities x and y be $x^3 - ax^2 + axy - y^3 = 0$; first dispose the Terms according to x , and then according to y , and multiply them in the following manner.

Mult	x^3	$-ax^2$	$+axy$	$-y^3$	$-y^3$	$+axy$	$-ax^2$
by	$\frac{3\dot{x}}{x}$	$\cdot \frac{2\dot{x}}{x}$	$\cdot \frac{\dot{x}}{x}$	$\cdot 0$	$\frac{3\dot{y}}{y}$	$\cdot \frac{\dot{y}}{y}$	$\cdot 0$
makes	$3\dot{x}x^2 - 2ax\dot{x} + a\dot{x}y$				$-3\dot{y}y^2 + a\dot{y}x$		
	*				*		

The Sum of the Products is $3\dot{x}x^2 - 2ax\dot{x} + a\dot{x}y - 3\dot{y}y^2 + a\dot{y}x = 0$, which Equation gives the Relation between the Fluxions \dot{x} and \dot{y} . For if you take x at pleasure, the Equation $x^3 - ax^2 + axy - y^3 = 0$ will give y . Which being determined, it will be $\dot{x} : \dot{y} :: 3y^2 - ax : 3x^2 - 2ax + ay$.¹³⁷

¹³⁶Isaac Newton (John Colson ed. and trans.), *The Method of Fluxions and Infinite Series with its Application to the Geometry of CURVE-LINES*, 1736, p. 20.

¹³⁷*Ibid.*, p. 21.

Overall one should recognise the method from Calculus as differentiating $f(x(t), y(t)) = 0$ to obtain $df/dt(x(t), y(t)) = 0$ for $f(x, y) = x^3 - ax^2 + axy - y^3$ and then determining dx/dy . The reference to an arbitrary arithmetical progression indicates that he is relying on the methods of Hudde and de Sluse.

After presenting a few additional examples, he comes to the demonstration. Here he resorts to infinitesimals by introducing the “indefinitely small Quantity” o , which we may take to be the differential dt and considers the “Moments” $\dot{v}o$, etc., which we may consider to be the differentials $dv = v'(t)dt$, etc.

DEMONSTRATION of the Solution.

13. The Moments of flowing Quantities, (that is, their indefinitely small Parts, by the accession of which, in indefinitely small portions of Time, they are continually increased,) are as the Velocities of their Flowing or Increasing.
14. Wherefore if the Moment of any one, as x , be represented by the Product of its Celerity \dot{x} into an indefinitely small Quantity o (that is, by $\dot{x}o$.) the Moments of the others y, z , will be represented by $\dot{y}o, \dot{z}o$; because $\dot{v}o, \dot{x}o, \dot{y}o$, and $\dot{z}o$ are to each other as $\dot{v}, \dot{x}, \dot{y}$, and \dot{z} .
15. Now since the Moments, as $\dot{x}o$ and $\dot{y}o$ are the indefinitely little accessions of the flowing Quantities x and y , by which those Quantities are increased through the several indefinitely little intervals of Time; it follows, that those Quantities x and y , after any indefinitely small interval of Time, become $x + \dot{x}o$ and $y + \dot{y}o$. And therefore the Equation, which at all times indifferently expresses the Relation of the flowing Quantities, will as well express the Relation between $x + \dot{x}o$ and $y + \dot{y}o$, as between x and y : So that $x + \dot{x}o$ and $y + \dot{y}o$ may be substituted in the same Equation for those Quantities, instead of x and y .
16. Therefore let any Equation $x^3 - ax^2 + axy - y^3 = 0$ be given, and substitute $x + \dot{x}o$ for x , and $y + \dot{y}o$ for y , and there will arise

$$\left. \begin{aligned} & x^3 + 3\dot{x}ox^2 + 3\dot{x}^2oox + \dot{x}^3o^3 \\ & - ax^2 - 2a\dot{x}ox - a\dot{x}^2oo \\ & + axy + a\dot{x}oy + a\dot{y}ox + a\dot{x}\dot{y}oo \\ & - y^3 - 3\dot{y}oy^2 - 3\dot{y}^2ooy - \dot{y}^3o^3 \end{aligned} \right\} = 0.$$

17. Now by Supposition $x^3 - ax^2 + axy - y^3 = 0$, which therefore being expunged, and the remaining Terms being divided by o , there will remain $3\dot{x}x^2 + 3\dot{x}^2ox + \dot{x}^3oo - 2a\dot{x}x - a\dot{x}^2o + a\dot{x}y + a\dot{y}x + a\dot{x}\dot{y}o - 3\dot{y}y^2 - 3\dot{y}^2oy - \dot{y}^3oo = 0$. But whereas o is supposed to be infinitely little, that it may represent the Moments of Quantities; the Terms that are multiply'd by it will be nothing in respect of the rest. Therefore I reject them, and there remains $3\dot{x}x^2 - 2a\dot{x}x + a\dot{x}y + a\dot{y}x - 3\dot{y}y^2 = 0$, as above in Examp. I.
18. Here we may observe, that the Terms that are not multiply'd by o will always vanish, as also those Terms that are multiply'd by o of more than one Dimension. And that the rest of the Terms being divided by o , will always acquire the form that they ought to have by the foregoing Rule: Which was the thing to be proved.¹³⁸

The method here is virtually the same as Fermat's calculation of

$$\frac{f(A + E) - f(A)}{E}$$

¹³⁸*Ibid.*, pp. 24–25.

and the subsequent elimination of all terms containing E , or Barrow's similar procedure. The big difference is the explicit reference to the infinitesimal nature of o as justification.

Newton gives several examples of the use of this technique before going on to discuss related problems, the third of which is

PROB. III.

To determine the Maxima and Minima of Quantities.

1. When a Quantity is the greatest or the least that it can be, at that moment it neither flows backwards or forwards. For if it flows forwards, or increases, that proves it was less, and will presently be greater than it is. And the contrary if it flows backwards, or decreases. Wherefore find its Fluxion, by Prob. 1. and suppose it to be nothing.
2. EXAMP. 1. If in the Equation $x^3 - ax^2 + axy - y^2 = 0$ the greatest Value of x be required; find the Relation of the Fluxions of x and y , and you will have $3\dot{x}x^2 - 2a\dot{x}x + a\dot{x}y - 3\dot{y}y^2 + a\dot{y}x = 0$. Then making $\dot{x} = 0$, there will remain $-3\dot{y}y^2 + a\dot{y}x = 0$ or $3y^2 = ax$. By the help of this you may exterminate either x or y out of the primary Equation, and by the resulting Equation you may determine the other, and then both of them by $-3y^2 + ax = 0$.¹³⁹

His next problem is the construction of tangent lines.

PROB. IV.

To draw Tangents to Curves.

First Manner.

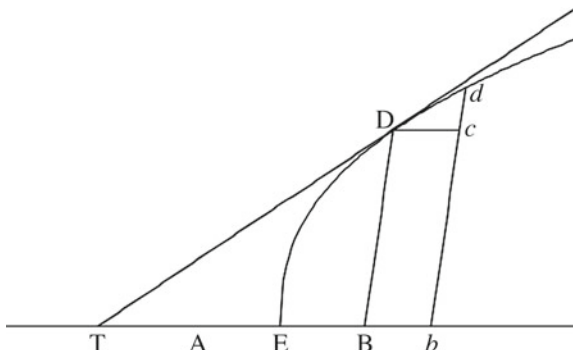
1. Tangents may be variously drawn, according to the various Relations of Curves to right Lines. And first let BD be a right Line, or Ordinate, in a given Angle to another right Line AB, as a Base or Absciss, and terminated at the Curve ED. Let this Ordinate move¹⁴⁰ through an indefinitely small Space to the place bd , so that it may be increased by the Moment cd , while AB is increased by the Moment Bb , to which Dc is equal and parallel. [See Fig. 2.37.] Let Dd be produced till it meets with AB in T, and this Line will touch the Curve in D or d ; and the Triangles dcD , DBT will be similar.¹⁴¹ So that it is $TB : BD :: Dc$ (or Bb) : cd .
2. Since therefore the Relation of BD to AB is exhibited by the Equation, by which the nature of the Curve is determined; seek for the Relation of Fluxions, by Prob. 1. Then take TB to BD in the Ratio of the Fluxion of AB to the Fluxion of BD, and TD will touch the Curve in the point D.
3. EX. 1. Calling $AB = x$, and $BD = y$, let their Relation be $x^3 - ax^2 + axy - y^3 = 0$. And the Relation of the Fluxions will be $3\dot{x}x^2 - 2a\dot{x}x + a\dot{x}y - 3\dot{y}y^2 + a\dot{y}x = 0$.

¹³⁹*Ibid.*, p. 44.

¹⁴⁰For Newton the curve is traced out as the point of intersection of two non-rotating lines moving along a pair of axes. The axes need not meet at right angles and the lines, called the *abscissa* and *ordinate*, need not be vertical and horizontal, but must remain parallel to the axes. The abscissa is parallel to the y -axis and its coordinate is x , while the ordinate is parallel to the x -axis with coordinate y .

¹⁴¹Newton is, of course, being only approximate here. If cd is the moment, d lies on the curve infinitesimally close to the tangent line, but not *on* the tangent line. The line through D and d will cross the curve not touch it.

Fig. 2.37 Newton's tangent construction



So that $\dot{y} : \dot{x} :: 3xx - 2ax + ay : 3y^2 - ax :: BD (y) : BT$. Therefore $BT = \frac{3y^3 - axy}{3x^2 - 2ax + ay}$. Therefore the point D being given, and thence DB and AB, or y and x, the length BT will be given, by which the Tangent TD is determined.¹⁴²

I find it amusing to contemplate the full title of the English edition of the *De methodus fluxionum* and the implied promise of the last line. I present it below without the numerous variations in font size, letting the frequent changes in style testify to its garishness:

THE
METHOD of FLUXIONS
AND
INFINITE SERIES;
WITH ITS
Application to the Geometry of CURVE- LINES.

By the INVENTOR
Sir ISAAC NEWTON, *K^t*.
Late President of the Royal Society.

*Translated from the AUTHOR's LATIN ORIGINAL
not yet made publick.*

To which is subjoin'd,
A PERPETUAL COMMENT upon the whole Work,
Consisting of
ANNOTATIONS, ILLUSTRATIONS, and SUPPLEMENTS,
In order to make this Treatise
A compleat Institution for the use of LEARNERS.

The “learners” in question are almost certainly not undergraduates. University instruction in mathematics of the day was simply not up to this level. Moreover, by all accounts, Newton was not a successful lecturer, having few students if any, often returning to his office early or occasionally speaking to an empty hall when students failed to show up. And one can see from the above passages, despite the care to logic,

¹⁴²Newton, *op. cit.*, p. 46.

why students would have a hard time understanding him. Today we would begin with the simplest case of a curve given by the graph of a function $y = f(x)$, and form the difference quotient

$$\frac{f(x + o) - f(x)}{o}. \quad (2.33)$$

Newton started off with a curve $f(x, y) = 0$ and calculated the mysterious

$$\frac{f(x + \dot{x}o, y + \dot{y}o) - f(x, y)}{o},$$

which we would represent as

$$\frac{f(x(t + o), y(t + o)) - f(x(t), y(t))}{o},$$

i.e., as

$$\frac{g(t + o) - g(t)}{o},$$

where $g(t) = f(x(t), y(t))$. Unfortunately, he did not yet have functional notation and launched right into multivariable calculus.

As for tangents, starting with $y = f(x)$, after simplifying (2.33), expunging o , whether by ignoring infinitesimal differences *à la* Newton or taking the limit as we do today, we would get the slope $f'(x)$ of the tangent and be led in the two variable case to finding $dy/dx = (dy/dt)/(dx/dt)$. Newton, not having the single variable case causing one to standardise on the slope, was slightly inconsistent. Thus, for his Example I, in Problem I he calculated the slope's reciprocal $\dot{x}/\dot{y} = dx/dy$ and in paragraphs 1 and 2 of Problem IV he describes the process in terms of finding this reciprocal, and then, in considering this Example, immediately finds the slope \dot{y}/\dot{x} itself. I want to add the simultaneous placement of d on the curve and the tangent line as another source of possible confusion, but his "learners" would be astronomers, physicists, mathematicians, and other learned scholars already familiar with ad-equality or the dismissal of infinitesimals, but not Newton's systematic method for solving problems involving curves and certainly not the Leibnizian d -notation, and this simultaneous occupation of two places by d would not have been confusing to them.

As already mentioned, the *De methodus fluxionum* was privately disseminated and a number of mathematicians learned the fluxional calculus before the book's publication in 1736. The first public announcement by Newton of the method was in 1687 in the first edition of the *Principia* and he still used infinitesimals. Within a few years he rejected them in favour of a not very clearly presented notion of limit in the *Quadratura curvarum*, composed in the years 1691–1692 and published in 1704 — his first publication on the Calculus. It was also the last to be written and the most widely read of his accounts, thus perhaps the definitive version of his theory. The second English translation of 1745 bears the subtitle "The Treatises themselves,

translated into English, with a large Commentary; in which the Demonstrations are supplied where wanting, the Doctrine illustrated, and the whole accommodated to the Capacities of Beginners, for whom it is chiefly designed". Reading Newton's opening words confirms that it is the Commentary that is chiefly designed for beginners:

INTRODUCTION to the Quadrature of Curves.

1. I Consider¹⁴³ mathematical Quantities in this Place not as consisting of very small Parts; but as describ'd by a continued Motion. Lines are describ'd, and thereby generated not by the Apposition of Parts, but by the continued Motion of Points¹⁴⁴; Superficies's¹⁴⁵ by the Motion of Lines; Solids by the Motion of Superficies's; Angles by the Rotation of the Sides; Portions of Time by a continual Flux: and so in other Quantities. These Geneses really take Place in the Nature of Things, and are daily seen in the Motion of Bodies. And after this Manner the Ancients, by drawing moveable right Lines along immoveable right Lines, taught the Genesis of Rectangles.
2. Therefore considering that Quantities, which increase in equal Times, and by increasing are generated, become greater or less according to the greater or less Velocity with which they increase and are generated; I sought a Method of determining Quantities from the Velocities of the Motions or Increments, with which they are generated; and calling these Velocities of the Motions or Increments *Fluxions*, and the generated Quantities *Fluents*, I fell by degrees upon the Method of Fluxions, which I have made use of here in the Quadrature of Curves, in the Years 1665 and 1666.
3. Fluxions are very nearly as the Augments of the Fluents generated in equal but very small Particles of Time, and, to speak accurately, they are the *first Ratio* of the nascent Augments; but they may be expounded by any Lines which are proportional to them.¹⁴⁶
4. Thus if the Area's ABC, AB DG [See Fig. 2.38.¹⁴⁷] be described by the Ordinates BC, BD moving along the Base AB with an uniform Motion, the Fluxions of these Area's shall be to one another as the describing Ordinates BC and BD, and may be expounded by these Ordinates, because that these Ordinates are as the nascent Augments of the Area's.¹⁴⁸
5. Let the Ordinate BC advance from its Place into any new Place *bc*. Complete the Parallelogram BCEb, and draw the right Line VTH touching the curve in C, and meeting

¹⁴³The paragraph opens with a normal sized "I", followed by a large drop cap "I", and "consider" subsequently capitalised. I decided this required too much effort to duplicate completely.

¹⁴⁴Newton is here taking a very Aristotelian view of the line.

¹⁴⁵I.e., surfaces.

¹⁴⁶"First ratio" and "nascent" are not exactly defined here. Their meaning will emerge when examples are discussed. In a couple of pages the "first ratio" will be called the "prime ratio". My interpretation of this paragraph is that, for any fluents v , w , $\dot{v}/\dot{w} \approx \Delta v/\Delta w$, that is, $\dot{v}/\dot{w} \approx (\dot{v}o)/(\dot{w}o)$.

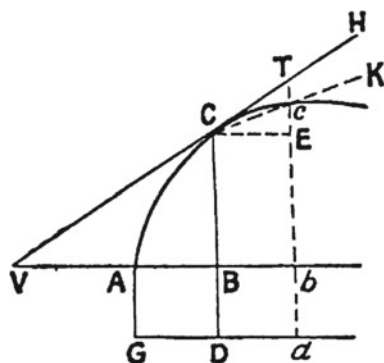
¹⁴⁷This image is taken from: Florian Cajori, *A History of the Conceptions of Limits and Fluxions in Great Britain from Newton to Woodhouse*, The Open Court Publishing Company, Chicago and London, 1919, p. 42, and is a clean reproduction of Newton's original.

¹⁴⁸In light of paragraph 3, he seems to be suggesting something like

$$\frac{d \int_a^x f(t) dt}{dx} \bigg/ \frac{d \int_a^x g(t) dt}{dx} = \frac{f(t)}{g(t)},$$

where the curve ACc is given by $y = f(x)$ and GDa by $y = g(x)$. Paragraphs 3 and 4 are unhelpful in the extreme. He thus seems to be asserting the Fundamental Theorem of the Calculus via some clumsy reference to proportion.

Fig. 2.38 Newton's prime and ultimate ratios



the two lines bc and BA produc'd in T and V : and Bb , Ec , and Cc will be the Augments now generated of the Absciss AB , the Ordinate BC and the Curve Line ACc ¹⁴⁹; and the Sides of the Triangle CET are in the *first Ratio* of these Augments considered as nascent, therefore the Fluxions of AB , BC and AC are as the Sides CE , ET and CT of that Triangle CET , and may be expounded by these same Sides, or, which is the same thing, by the Sides of the Triangle VBC , which is similar to the Triangle CET .

6. It comes to the same Purpose to take the Fluxions in the *ultimate Ratio* of the evanescent Parts. Draw the right Line Cc , and produce it to K . Let the Ordinate bc return into its former place BC , and when the Points C and c coalesce, the right Line CK will coincide with the Tangent CH , and the evanescent Triangle CEc in its ultimate Form will become similar to the Triangle CET , and its evanescent Sides CE , Ec and Cc will be *ultimately* among themselves as the sides CE , ET and CT of the other Triangle CET , are, and therefore the Fluxions of the Lines AB , BC and AC are in this same Ratio. If the Points C and c are distant from one another by any small Distance, the right Line CK will likewise be distant from the Tangent CH by a small Distance. That the right Line CK may coincide with the Tangent CH , and the ultimate Ratios of the Lines CE , Ec and Cc may be found, the Points C and c ought to coalesce and exactly coincide. The very smallest Errors in mathematical Matters are not to be neglected.¹⁵⁰

The last sentence is occasionally interpreted as the complete and final rejection of infinitesimals on Newton's part. Newton is here fumbling for a notion of limit: What are the ratios in the triangle CEc when C and c finally coalesce? He expressed himself more clearly elsewhere. In the second edition of the *Principia* of 1713 we read:

Quantities, as also ratios of quantities, which constantly tend toward equality in any finite time, and before the end of that time approach each other more nearly than [with] any given difference whatever, become ultimately equal...¹⁵¹

The objection is that there is no ratio of evanescent quantities, which obviously, before they have vanished, is not ultimate; when they have vanished, there is none. But also by the same like argument it may be contended that there is no ultimate velocity of a body arriving

¹⁴⁹Thus, these will be our dx , dy , and ds .

¹⁵⁰John Stewart, *Sir Isaac Newton's Two Treatises of the Quadrature of Curves, and Analysis by Equations of an infinite Number of Terms, explained*, London, 1745, pp. 1–2. The second treatise referred to in the title is *De analysi*.

¹⁵¹This is remarkably close to our modern definition of "limit".

at a certain position; for before the body attains the position, this is not ultimate; when it has attained [it], there is none. And the answer is easy: By ultimate velocity I understand that with which the body is moved, neither before it arrives at the ultimate position and the motion ceases, nor thereafter, but just when it arrives; that is, that very velocity with which the body arrives at the ultimate position and with which the motion ceases. And similarly for the motion of evanescent quantities is to be understood the ratio of the quantities, not before they vanish, nor thereafter, but [that] with which they vanish. And likewise the first nascent ratio is the ratio with which they begin. And the prime and ultimate amount is to be [that] with which they begin and cease (if you will, to increase and diminish). There exists a limit which the velocity may attain at the end of the motion, but [which it may] not pass. This is the ultimate velocity. And the ratio of the limit of all quantities and proportions, beginning and ceasing, is equal...

The ultimate ratios in which quantities vanish, are not really the ratios of ultimate quantities, but the limits toward which the ratios of quantities, decreasing without limit, always approach; and to which they can come nearer than any given difference, but which they can never pass nor attain before the quantities are diminished indefinitely.¹⁵²

Newton uses the word “limit” here in the sense of the word “bound”. The curves familiar in those days were well-behaved with little oscillation. As one approached a point on a curve from the left or from the right, once one got close enough, the curve was monotone. Thus, as x approached a from the left, say, the values $f(x)$ either approached their limit from above or from below, but never passed it. Today, as with Darboux’s function of Fig. 2.26, we drop the clause disallowing the quantities from attaining or surpassing their limits.

Newton clearly has a conception of limit, but has been having difficulty expressing it. He hasn’t isolated its defining characteristic.

Following the introductory remarks in the *Quadratura curvarum*, Newton gives a few examples to illustrate his concepts. The one most anthologised is his differentiation of powers of x :

11. Let the Quantity x flow uniformly,¹⁵³ and let it be proposed to find the Fluxion of x^n . In the same Time that the Quantity x , by flowing, becomes $x + o$, the Quantity x^n will become $\overline{x + o}^n$, that is, by the Method of infinite Series’s,¹⁵⁴ $x^n + nox^{n-1} + \frac{n^2-n}{2}oox^{n-2} + \&c$. And the Augments o and $nox^{n-1} + \frac{n^2-n}{2}oox^{n-2} + \&c$. are to one another as 1 and $nx^{n-1} + \frac{n^2-n}{2}ox^{n-2} + \&c$.
Now let these Augments vanish, and their ultimate Ratio will be 1 to nx^{n-1} .
12. By like ways of reasoning, the Fluxions of Lines, whether right or curve in all Cases, as likewise the Fluxions of Superficies’s, Angles and other Quantities, may be collected by the Method of *prime* and *ultimate* Ratios. Now to institute an Analysis after this manner in finite Quantities and investigate the *prime* or *ultimate* Ratios of these finite Quantities when in their nascent or evanescent State, is consonant to the Geometry of the Ancients: and I was willing to show that, in the Method of Fluxions, there is no necessity of introducing Figures infinitely small into Geometry. Yet the Analysis

¹⁵²Smith, *Source Book...*, *op. cit.*, pp. 617–618. Cf. also Struik, *op. cit.*, pp. 299–300.

¹⁵³That is, let x be a constant multiple of time, so that $y = x^n$, being a function of time, is in fact a function of x .

¹⁵⁴Newton had extended the Binomial Theorem to the case of arbitrary rational exponents. For n not a positive integer, however, the expansion is an infinite series. Thus, Newton is here differentiating x^n for arbitrary rational n . For more information, I refer the reader to Smoryński, *Treatise*.

may be performed in any kind of Figures, whether finite or infinitely small, which are imagin'd similar to the evanescent Figures; as likewise in these Figures, which, by the Method of Indivisibles, use to be reckoned as infinitely small, provided you proceed with due Caution.¹⁵⁵

I imagine the modern mathematician reacting to paragraph 11 in a manner not unlike the way one typically reacts to fingernails on a chalkboard. However lacking in rigour, Newton has nevertheless essentially defined the derivative of x^n as

$$\lim_{o \rightarrow 0} \frac{(x + o)^n - x^n}{o} = nx^{n-1}.$$

When I say this, I do so as a mathematician in identifying things that, however different, are abstractly the same. Newton did not have our conceptual framework. He did not have the concept of function. Thus, 11 does not read: Let $f(x) = x^n$; then $f'(x) = nx^{n-1}$. x^n was not a function of x , but another flowing quantity y varying with time and the two quantities were related by an equation $y = x^n$. They had fluxions \dot{x} and \dot{y} , which we think of as dx/dt and dy/dt , as we regard x and y as functions of t . Time is somewhere behind the scenes in Newton, but not as an explicit variable t . The closest he comes to this is when he assumes x to “flow uniformly” — in essence making x stand in for t .

2.3.5 Leibniz

Leibniz independently discovered the Calculus in the 1670s, making his first discoveries around 1672 and publishing his first paper on the subject a little over a decade later in 1684. Where today mathematical papers have abstracts following the titles, in those days the fashion was to incorporate the abstract into the title: “Nova methodus pro maximis et minimis, itemque tangentibus, quæ nec fractas nec irrationales quantitates moratur, et singulare pro illis calculi genus”¹⁵⁶ [“A new method for maxima and minima as well as tangents, which is impeded neither by fractional nor irrational quantities, and a remarkable type of calculus for this”]. Leibniz, however, never published a systematic account of the whole, spreading his work out over numerous short papers and correspondence with others.

While Leibniz’s early publications are important for the history of mathematics, his earlier unpublished manuscripts may offer more insight. From a manuscript of 1677 we get the following announcement, which may also serve as a partial review of the history of tangent finding methods:

¹⁵⁵Stewart, *op. cit.*, p. 4.

¹⁵⁶*Acta Eruditorum* 3 (1684), pp. 467–473. A German translation of this and several further papers of Leibniz was published by Gerhardt Kowalewski as number 162 in Ostwald’s series of scientific classics in 1908. A partial English translation appears in Smith’s source book, *op. cit.*, a full translation in Struik, *op. cit.*, and a nearly full translation in Stedall, *op. cit.*

Fermat was the first to find a method which could be made general for finding the straight lines that touch analytical curves. Descartes accomplished it in another way, but the calculation that he prescribes is a little prolix. Hudde has found a remarkable abridgment by multiplying the terms of the progression by those of the arithmetical progression. He has only published it for equations in one unknown; although he has obtained it for those in two unknowns. Then the thanks of the public are due to Sluse; and after that, several have thought that this method was completely worked out.¹⁵⁷ But all these methods that have been published suppose that the equation *has been reduced* and cleared of fractions and irrationals; I mean of those in which the variables occur. I however have found means of obviating these useless reductions, which make the calculation increase to a terrible degree, and oblige us to rise to very high dimensions, in which case we have to look for a corresponding depression with much trouble; instead of all this, everything is accomplished at the first attack.¹⁵⁸

This method has more advantage over all the others that have been published, than that of Sluse has over the rest, because it is one thing to give a simple abridgment of the calculation, and quite another thing to get rid of reductions and depressions. With respect to the publication of it, on account of the great extension of the matter which Descartes himself has stated to be the most useful part of Geometry, and of which he has expressed the hope that there is more to follow — in order to explain myself shortly and clearly, I must introduce some *fresh characters*, and give to them a *new Algorithm*, that is to say, altogether special rules, for their addition, subtraction, multiplication, division, powers, roots, and also for equations.¹⁵⁹

The “fresh characters” are the differentials:

Explanation of the characters.

Suppose there are several curves, as CD, FE, HJ, connected with one and the same axis AB by ordinates drawn through one and the same point B, to wit, BC, BF, BH. The tangents CT, FL, HM to these curves cut the axis in the points T, L, M [See Fig. 2.39.]; the point A in the axis is fixed, and the point B changes with the ordinates. Let $AB = x$, $BC = y$, $BF = w$, $BH = v$; also let the ratio of TB to BC be called that of dx to dy , and the ratio of LB to BF that of dx to dw , and the ratio of MB to BH that of dx to dv . Then if, for example, y is equal to vw , we should say dvw instead of dy , and so on for all other cases. Let a be a constant straight line; then if y is equal to a , that is if CD is a straight line parallel to AB, dy or da will be equal to 0, or equal to zero. If the magnitude dx/dw comes out negative, then FL, instead of being drawn toward A, above B, will be drawn in the contrary direction, below B.¹⁶⁰

Leibniz follows this with a list of the computation rules for differentials: for a constant,

if $y = v \pm w \pm a$ then $dy = dv \pm dw$

if $y = avw$ then $dy = avdw + awdv$,

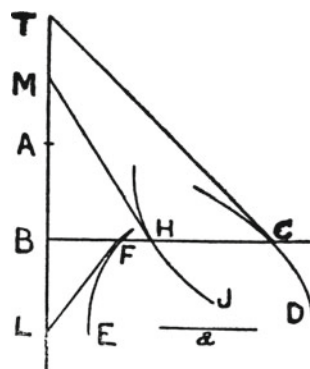
¹⁵⁷Newton had not yet published on the subject and consequently is not mentioned. It is not clear why Barrow is not mentioned. Child is of the opinion that Leibniz was hiding his dependence on Barrow, but it could also be the similarity of Barrow's analytical method to Fermat's.

¹⁵⁸“Reductions” obviously refers to simplifying the equations for which tangents are sought. The process, especially in clearing surds, results in polynomials of higher degrees and the introduction of extra possibilities for the tangent. Presumably “depressions” refers at least in part to the elimination of false solutions. The point of this passage is that the method he is introducing is more direct and eliminates this excess work.

¹⁵⁹Child, *Early Mathematical Manuscripts...*, *op. cit.*, pp. 131–132.

¹⁶⁰*Ibid.*, pp. 132–133.

Fig. 2.39 Leibniz's diagram



etc. The proofs for *infinitely small* quantities dx, dy are given in a revision of the manuscript written sometime in the period 1677–1680.

Edwards makes the interesting point that for Newton the fluxion \dot{x} , or derivative, was the same sort of quantity as x , while for Leibniz the fundamental notion was the differential dx ($=$ Newton's $\dot{x}o$). Where we see a derivative, Leibniz saw “‘merely’ a geometrically significant quotient”.¹⁶¹ For Leibniz, once dx was given, dy was defined by setting dx/dy equal to the reciprocal of the slope of the tangent line.¹⁶²

With Newton and Leibniz we finally have a calculus making routine the determination of the tangent for any analytic curve. As Leibniz emphasised, Hudde and de Sluse could do so for curves $f(x, y) = 0$ for polynomials $f(x, y)$. When f involved rational functions or roots, one had to do some preliminary work eliminating the fractions or roots under their method. This was no longer necessary with the Newton-Leibniz algorithm. Moreover, as new functions like $\sin x$ or $\ln x$ were differentiated, they could be combined with the old functions and the new combinations differentiated by the new method. A definite advance had been made with respect to *finding* tangents. But what about *defining* tangents? The passages quoted contain no definition.

Each author hinted at the modern definition. I have already quoted one translation of Newton above on page 107:

If the points C and c are distant from one another by any small Distance, the right Line CK will likewise be distant from the Tangent CH by a small Distance. That the right Line CK may coincide with the Tangent CH ...the Points C and c ought to coalesce and exactly coincide.

The meaning of this passage is perhaps more easily seen in Smith's translation:

¹⁶¹Edwards, *op. cit.*, p. 266. Edwards offers, incidentally, an excellent discussion of Leibniz's early papers as well as the published ones.

¹⁶²Think of AB and BC as the x - and y -axes, respectively, in Fig. 2.39.

If the points are distinct from each other by an interval, however small, the secant will be distant from the tangent by a small interval. That it may coincide with the tangent and the last ratio be found, the two points must unite and coincide altogether.¹⁶³

A comparison with the Latin original reveals that Smith's is a loose translation, Newton himself referring to the specific points and lines of Fig. 2.38 as in Struik's translation. Nonetheless, he makes perfectly clear the fact that Newton is stating that the tangent is the limiting position of the secant lines as b moves closer and closer to B .

Leibniz unabashedly embraces the infinitely small and instead of referring to motion or limits appeals to infinitesimals. In his first published paper on the Calculus we read

We have only to keep in mind that to find a *tangent* means to draw a line that connects two points of the curve at an infinitely small distance, or the continued side of a polygon with an infinite number of angles, which for us takes the place of the *curve*.¹⁶⁴

A formal definition of the tangent line ought to be possible from either of these remarks. What is lacking is in Newton's case a clear and precise definition of limit and in Leibniz's case a solid justification of the use of infinitesimals. Leibniz attempted such a justification after receiving some harsh criticism, but his attempt was ultimately unsuccessful. Both approaches received stinging criticism in 1734 by George Berkeley, Bishop of Cloyne, (1685–1753) who, irked by mathematicians criticising theologians for their weak reasoning exposed the equal or even greater weakness of the reasoning resorted to by mathematicians in explaining the new calculus. Berkeley's tract, *The Analyst; or, a Discourse Addressed to an Infidel Mathematician*, is a carefully reasoned critique of the flaws in the arguments put forward by the mathematicians of the day.

2.3.6 Bumps in the Road

A discussion of Berkeley is not immediately relevant to our discussion of the Mean Value Theorem, but it is interesting, it sheds some light on the shortcomings of the work cited so far, and, of all the criticisms levelled against the Calculus, Berkeley's was the most influential. Hence I have decided to enter into a digression on Berkeley. The reader who is pressed for time may prefer to skip this and jump ahead to the next subsection on page 123, below.

Berkeley was certainly not the first to criticise analytical practice. "The philosopher Thomas Hobbes raised the first doubts, to be echoed by many others later, on the use of infinitely small or indivisible quantities".¹⁶⁵ In response to some criticisms by Bernard Nieuwentijt (1654–1718), Leibniz began

¹⁶³Smith, *Source Book*, *op. cit.*, p. 617.

¹⁶⁴Struik, *op. cit.*, p. 276.

¹⁶⁵Stedall, *op. cit.*, p. 66. Stedall includes (p. 69) an excerpt from Hobbes, *Six lessons to the Professors of Mathematics*, 1656, p. 46, criticising in the plainest language the use of infinitesimals

When my infinitesimal calculus, which includes the calculus of differences and sums, had appeared and spread, certain over-precise veterans began to make trouble; just as once long ago the Sceptics opposed the Dogmatics, as is seen from the work of Empiricus [*sic*: Epicurus is meant.] against the mathematicians (i.e., the dogmatics), and such as Francisco Sanchez, the author of the book *Quod nihil scitur*, brought against Clavius; and his opponents to Cavalieri, and Thomas Hobbes to all geometers, and just lately such objections as are made against the quadrature of the parabola by Archimedes by that renowned man, Dethlevus Cluver. When then our method of infinitesimals, which had become known by the name of the calculus of differences, began to be spread abroad by several examples of its use, both of my own and also of the famous brothers Bernoulli, and more especially by the elegant writings of that illustrious Frenchman, the Marquis d'Hospital, just lately a certain erudite mathematician, writing under an assumed name in the scientific *Journal de Trevoux*, appeared to find fault with this method. But to mention one of them by name, even before this there arose against me in Holland Bernard Nieuwentijt, one indeed really well equipped both in learning and ability, but one who wished rather to become known by revising our methods to some extent than by advancing them.¹⁶⁶

Though milder in tone than the comments by Hobbes and dripping with less sarcasm than Berkeley's tract, and as entertaining as it is, today such writing would be deemed inappropriate, particularly the *ad hominem* remark of the last cited sentence. But even the opening claim to martyrdom would raise the issue of paranoia, and an author would be ill-advised to submit such remarks for publication.

Berkeley too does not restrain himself from making personal attacks. Still, Berkeley does offer some cogent criticism, and overall Robert Woodhouse's assessment is probably fair:

The name of Berkeley has occurred more than once in the preceding pages: and I cannot quit this part of my subject without commending the analyst and the subsequent pieces, as forming the most satisfactory controversial discussion of pure science, that ever yet appeared: into what perfection of perspicuity and of logical precision, the doctrine of fluxions may be advanced, is no subject of consideration: But, view the doctrine as Berkeley found it, and its defects in metaphysics and logic are clearly made out.

If, for the purpose of habituating the mind to just reasoning, (and mental discipline is all the good the generality of students derive from the mathematics)¹⁶⁷ I were to recommend a book, it should be the *Analyst*. Even those who still regard the doctrine of fluxions as clearly and firmly established by their immortal inventor, may read it, not unprofitably, since, if it does not prove the cure of prejudice, it will be at least the punishment.¹⁶⁸

Berkeley begins with a brief description of the method of fluxions as it is described in the *Quadratura curvarum*:

III. The Method of Fluxions is the general Key, by help whereof the modern Mathematicians unlock the secrets of Geometry, and consequently of Nature. And as it is that which hath enabled them so remarkably to outgo the Ancients in discovering Theorems and

(Footnote 165 continued)

by John Wallis (1616–1703) whose *Arithmetica infinitorum* of 1655 provided the spark that ignited Newton.

¹⁶⁶Child, *Early Mathematical Manuscripts...*, *op. cit.*, pp. 145–146.

¹⁶⁷And not even this when courses are watered down and only drill is offered.

¹⁶⁸Robert Woodhouse, *The Principles of Analytical Calculation*, University of Cambridge Press, Cambridge, 1803, pp. xvii–xviii.

solving Problems, the exercise and application thereof is become the main, if not sole, employment of all those who in this Age pass for profound Geometers. But whether this Method be clear or obscure, consistent or repugnant, demonstrative or precarious, as I shall inquire with the utmost impartiality, so I submit my inquiry to your own Judgment, and that of every candid Reader. Lines are supposed to be generated by the motion of Points, Plains by the motion of Lines, and Solids by the motion of Plains. And whereas Quantities generated in equal times are greater or lesser, according to the greater or lesser Velocity, wherewith they increase and are generated, a Method hath been found to determine Quantities from the Velocities of their generating Motions. And such Velocities are called Fluxions: and the Quantities generated are called flowing Quantities. These Fluxions are said to be nearly as the Increments of the flowing Quantities, generated in the least equal Particles of time; and to be accurately in the first Proportion of the nascent, or in the last of the evanescent, Increments. Sometimes, instead of Velocities, the momentaneous Increments or Decrements of undetermined flowing Quantities are considered, under the Appellation of Moments.¹⁶⁹

Berkeley launches his criticism on general epistemological principles:

- IV. By Moments we are not to understand finite Particles. These are said not to be Moments, but Quantities generated from Moments, which last are only the nascent Principles of finite Quantities. It is said, that the minutest Errors are not to be neglected in Mathematics: that the Fluxions are Celebrities, not proportional to the finite Increments though ever so small; but only to the Moments or nascent Increments, whereof the Proportion alone, and not the Magnitude is considered. And of the aforesaid Fluxions there be other Fluxions, which Fluxions of Fluxions are called second Fluxions. And the Fluxions of these second Fluxions are called third Fluxions: and so on, fourth, fifth, sixth, &c. *ad infinitum*. Now as our Sense is strained and puzzled with the perception of Objects extremely minute, even so the Imagination, which Faculty derives from Sense, is very much strained and puzzled to frame clear Ideas of the least Particles of time, or the least Increments generated therein: and much more so to comprehend the Moments, or those Increments of the flowing Quantities in *statu nascenti*, in their very first origin or beginning to exist, before they become finite Particles. And it seems still more difficult, to conceive the abstracted Velocities of such nascent imperfect Entities. But the Velocities of the Velocities, the second, third, fourth and fifth Velocities, &c. exceed, if I mistake not, all Humane Understanding. The further the Mind analyseth and pursueth these fugitive Ideas, the more it is lost and bewildered; the Objects, at first fleeting and minute, soon vanishing out of sight. Certainly in any Sense a second or third Fluxion seems an obscure Mystery. The incipient Celerity of an incipient Celerity, the nascent Augment of a nascent Augment, *i. e.* of a thing which hath no Magnitude: Take it in which light you please, the clear Conception of it will, if I mistake not, be found impossible, whether it be so or no I appeal to the trial of every thinking Reader. And if a second Fluxion be inconceivable, what are we to think of third, fourth, fifth Fluxions, and so onward without end?¹⁷⁰

Paragraph IV is a philosophical assault on the method of fluxions and instances his opposition to abstract entities and his belief that Geometry should deal only with the immediately perceivable:

...he rejects infinitesimals on the grounds that they are simply incomprehensible:

¹⁶⁹George Berkeley, *The Analyst; or, a Discourse Addressed to an Infidel Mathematician.*, London, 1734, pp. 6–7.

¹⁷⁰*Ibid.*, pp. 8–9.

Axiom. No reasoning about things whereof we have no idea. Therefore no reasoning about Infinitesimals.

Nor can it be objected that we reason about Numbers w^{ch} are only words & not ideas, for these Infinitesimals are words of no use if not supposed to stand for Ideas.

Much less infinitesimals of infinitesimals &c.

Berkeley argues here that we can frame no idea of infinitesimals and, thus, that there is no legitimate purpose served by introducing signs such as dx or $o\dot{x}$ into mathematical discourse. His criticisms clearly depend upon the “axiom” that no word is to be used without an idea. When Berkeley later repudiates this axiom it might be thought that he is no longer entitled to this kind of critique of infinitesimals. I think, however, that this conclusion is unduly hasty.^{171, 172}

Infinitesimals were already being hotly debated before Berkeley wrote *The Analyst*. Newton had used them before replacing them by his references to nascent and evanescent augments and prime and ultimate ratios, finding, he believed, greater rigour in the replacement’s implied limit concept. Others accepted them wholeheartedly, but disagreed on their exact nature. Nieuwentijt and Leibniz embraced them, but while Nieuwentijt was willing to accept infinitesimal objects infinitely small in comparison with finite numbers, he was unwilling to accept things infinitely small compared to them. Leibniz, on the other hand, had no qualms about positing a whole hierarchy of infinitesimals: There were first order infinitesimals, which were infinitely small with respect to finite numbers; then second order infinitesimals infinitely small with respect to first order infinitesimals; and so on. When Comparing finite quantities, one could ignore infinitesimal differences; when comparing first order infinitesimal quantities, one could ignore second and higher order infinitesimal differences; and so on. An exact account of infinitesimals was lacking and their use was informal, intuitive, and most decidedly non-rigorous.

Berkeley now carried his criticism of fluxions and their iterations over to infinitesimals and higher infinitesimals in paragraphs V and VI. His argument is again not that a justification for their use is lacking, but that a justification must be lacking because they make no sense:

Now to conceive a Quantity infinitely small, that is, infinitely less than any sensible or imaginable Quantity, or than any the least finite Magnitude, is, I confess, above my Capacity. But to conceive a Part of such infinitely small Quantity, that shall be still infinitely less than it, and consequently though multiply’d infinitely shall never equal the minutest finite Quantity, is, I suspect, an infinite Difficulty to any Man whatsoever; and will be allowed such by those who candidly say what they think; provided they really think and reflect, and do not take things upon trust.¹⁷³

¹⁷¹Douglas M. Jesseph, *Berkeley’s Philosophy of Mathematics*, University of Chicago Press, Chicago, 1993, pp. 158–159. I confess to having given this book only a quick and superficial reading, but it strikes me as offering an excellent in-depth discussion of Berkeley’s criticism of the Calculus. Another source worthy of mention is Cajori, *A History of the Conceptions of Limits...*, *op. cit.*.

¹⁷²However “unduly hasty”, the rejection of Berkeley’s critique follows from Berkeley’s own Lemma cited in paragraph XII of *The Analyst* — cf. p. 117, below.

¹⁷³Berkeley, *op. cit.*, p. 10.

Berkeley was no enemy of mathematics, nor even of the method of fluxions itself. His goal was not to criticise analysis *per se*, but to deflate the mathematicians who criticised theological argumentation by showing that the analysts were guilty of the same crimes against reason they accused the theologians of. The hint of this in the last clause of the above citation is replaced by more direct statements of this thesis in paragraphs VII and VIII, for example:

- VII. All these Points, I say, are supposed and believed by certain rigorous Exactors of Evidence in Religion, Men who pretend to believe no further than they can see. That Men, who have been conversant only about clear Points, should with difficulty admit obscure ones might not seem altogether unaccountable. But he who can digest a second or third Fluxion, a second or third Difference, need not, methinks, be squeamish about any Point in Divinity... But with what appearance of Reason shall any Man presume to say, that Mysteries may not be Objects of Faith, at the same time that he himself admits such obscure Mysteries to be the Object of Science?¹⁷⁴

Berkeley is on firmer ground in paragraph IX when he attacks Newton's derivation of the product formula for differentiation, stated in terms of moments, in the *Principia*:

The main Point in the method of Fluxions is to obtain the Fluxion or Momentum of the Rectangle or Product of two indeterminate Quantities. Inasmuch as from thence are derived Rules for obtaining the Fluxions of all other Products and Powers; be the Coefficients or the Indexes what they will, integers or fractions, rational or surd. Now this fundamental Point one would think should be very clearly made out, considering how much is built upon it, and that its Influence extends throughout the whole Analysis. But let the Reader judge. This is given for Demonstration. Suppose the Product or Rectangle AB increased by continual Motion: and that the momentaneous Increments of the Sides A and B are a and b . When the Sides A and B were deficient, or lesser by one half of their Moments, the Rectangle was $A - \frac{1}{2}a \times B - \frac{1}{2}b$, i.e. $AB - \frac{1}{2}aB - \frac{1}{2}bA + \frac{1}{4}ab$. And as soon as the Sides A and B are increased by the other two halves of their Moments, the Rectangle becomes $A + \frac{1}{2}a \times B + \frac{1}{2}b$ or $AB + \frac{1}{2}aB + \frac{1}{2}bA + \frac{1}{4}ab$. From the latter Rectangle subduct the former, and the remaining difference will be $aB + bA$. Therefore the Increment of the Rectangle generated by the intire [*sic*] Increments a and b is $aB + bA$. *Q.E.D.* But it is plain that the direct and true Method to obtain the Moment or Increment of the Rectangle AB , is to take the Sides as increased by their whole Increments, and so multiply them together, $A + a$ by $B + b$, the product whereof $AB + aB + bA + ab$ is the augmented Rectangle; whence if we subduct AB , the Remainder $aB + bA + ab$ will be the true Increment of the Rectangle, exceeding that which was obtained by the former illegitimate and indirect Method by the Quantity ab . And this holds universally be the Quantities a and b what they will, big or little, Finite or Infinitesimal, Increments, Moments or Velocities. Nor will it avail to say that ab is a Quantity exceeding small: Since we are told that *in rebus mathematicis errores quàm minimi non sunt contemnendi*.^{175, 176}

Newton's argument is a bit of sleight of hand, but not the good kind where he dazzles us with a clever trick we would never have thought of. No, he tries to pull the

¹⁷⁴*Ibid.*, p. 12.

¹⁷⁵This is the Latin original of Newton's remark cited above that "The very smallest Errors in mathematical Matters are not to be neglected".

¹⁷⁶Berkeley, *op. cit.*, pp. 14–16.

wool over our eyes by calculating a different expression. Berkeley was not alone in criticising Newton on this; no less a mathematician than William Rowan Hamilton wrote later, in 1862, to no less an admirer of Newton than Augustus de Morgan:

His mode of getting rid of ab appeared to me long ago (I must confess it) to involve so much of *artifice*, as to deserve to be called *sophistical*; although I should not like to say so publicly. He subtracts, you know $\left(A - \frac{1}{2}a\right)\left(B - \frac{1}{2}b\right)$ from $\left(A + \frac{1}{2}a\right)\left(B + \frac{1}{2}b\right)$; whereby, of course, ab disappears in the result. But by *what right*, or *what reason* other than to give an unreal air of *simplicity* to the calculation, does he *prepare* the *products* thus?¹⁷⁷

Newton's trick survives today in the form of the following exercise.

2.3.8 Exercise Let f be a function defined in some interval containing the number a .

- i. Show: If $f'(a)$ exists, then

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a-h)}{2h} = f'(a).$$

- ii. Show by example that

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a-h)}{2h}$$

can exist even when $f'(a)$ does not.

So Newton is, in effect, calculating a generalised derivative which may exist when the derivative doesn't, but will always give the correct value when it does exist. He may not have provided any justification for his procedure, but some justification does exist.

Berkeley's next mathematical assault is a more direct frontal attack. He has just said that Newton's derivation of the product formula succeeds by ignoring an error and thus is in violation of one of Newton's own principles. He now claims that Newton's argument is self-contradictory and conclusions based on contradictions carry no weight.

- XII. From the foregoing Principle so demonstrated, the general Rule for finding the Fluxion of any Power of a flowing Quantity is derived. But, as there seems to have been some inward Scruple or Consciousness of defect in the foregoing Demonstration, and as this finding the Fluxion of a given Power is a Point of primary Importance, it hath therefore been judged proper to demonstrate the same in a different manner independent of the foregoing Demonstration. But whether this other Method be more legitimate and conclusive than the former, I proceed now to examine; and in order thereto shall premise the following Lemma. "If with a View to demonstrate any Proposition, a certain Point is supposed, by virtue of which certain other points are attained; and such supposed Point be it self afterwards destroyed or rejected by contrary Supposition; in that case, all the other Points, attained thereby and consequent thereupon, must also be destroyed and rejected, so as from thence forward to be no more supposed or applied in the Demonstration." This is so plain as to need no Proof.

¹⁷⁷Smith, *Source Book*, *op. cit.*, p. 631.

XIII. Now the other Method of obtaining a Rule to find the Fluxion of any Power is as follows. Let the Quantity x flow uniformly, and be it proposed to find the Fluxion of x^n . In the same time that x by flowing becomes $x + o$, the Power x^n becomes $\bar{x} + o|n$, i. e. by the Method of infinite Series $x^n + nox^{n-1} + \frac{n(n-1)}{2}oox^{n-2} + \&c.$ and the Increments o and $nox^{n-1} + \frac{n(n-1)}{2}oox^{n-2} + \&c.$ are one to another as 1 to $nx^{n-1} + \frac{n(n-1)}{2}ox^{n-2} + \&c.$ Let now the Increments vanish, and their last Proportion will be 1 to nx^{n-1} . But it should seem that this reasoning is not fair or conclusive. For when it is said, let the Increment vanish, i. e. let the Increments be nothing, or let there be no Increments, the former Supposition that the Increments were something, or that there were Increments, is destroyed, and yet a Consequence of that Supposition, i. e. an Expression got by virtue thereof, is retained. Which, by the foregoing Lemma, is a false way of reasoning. Certainly when we suppose the Increments to vanish, we must suppose their Proportions, their Expressions, and everything else derived from the Supposition of their Existence to vanish with them.¹⁷⁸

Berkeley was right about the product formula yielding the formula for differentiating powers x^n — for positive integers n . For other rational exponents a bit more is required. Newton's new derivation is not an "inward Scruple or Consciousness of defect" about the earlier proof, but an attempt to unify the treatment of the differentiation of x^n for all rational n . This new derivation is far from rigorous and the necessary rigour was a long time coming.¹⁷⁹

Berkeley's objection to the derivation is, however, completely off the mark. Newton in no way violates Berkeley's Lemma. He is not first taking o not to be 0 so he can divide by o and obtain a new equation, and then changing his mind and saying o is 0 in the resulting equation. Newton does indeed assume o is not 0 in calculating

$$\frac{(x + o)^n - x^n}{o} = nx^{n-1} + \frac{n(n-1)}{2}ox^{n-2} + \dots,$$

but he now wants to claim that, the values on the left always equalling those on the right, the two expressions will share the same limit as $o \rightarrow 0$ and that the limit on the right can be calculated by setting o equal to 0 in *that* expression. He wants to say this, but the limit concept has not yet crystallised sufficiently for him to say this clearly.

Newton, Leibniz, and Berkeley agreed on the value of the Calculus and on the truth of its results. All three also believed that any result obtained by the method of fluxions or the use of infinitesimals could be rigorously verified by the old Greek methods. What they disagreed on was the justification of the new procedures. Newton was evidently not one to worry himself much on the matter. When, for example, he extended the Binomial Theorem from positive integral exponents to arbitrary rational exponents by guessing the form for an expansion of $\sqrt{1 - x^2}$, he did not prove the result, but checked it by formally multiplying the resulting series by itself to get

¹⁷⁸*Ibid.*, pp. 19–21.

¹⁷⁹I give a fairly complete account of the history of the Binomial Theorem in: Smoryński, *Treatise*, *op. cit.* It might be added that the two proofs are from two different works of Newton's and thus their simultaneous existence, even should the results have been on equal footing, would not necessarily have been proof of anything more than variety.

$1 - x^2$. He also formally applied the familiar old algorithm for finding square roots and obtained the same infinite series. As to the question of fluxions and its use of infinitesimals, he merely said that infinitesimals weren't needed, that they could be replaced by sufficiently small finite quantities, and mumbled something about "limits". He made no attempt at a rigorous justification.

Berkeley tried to explain the success by a theory of compensating errors. I confess to find this too absurd to have read this part of his essay, though it may well be the case that he is able to treat one or two special cases successfully.

Leibniz is the one who thought deeply about the matter. In one of his manuscripts only published posthumously, he wrote

It has been proposed to me several times to confirm the essentials of our calculus by demonstrations, and here I have indicated below its fundamental principles, with the intent that any one who has the leisure may complete the work. Yet I have not seen up to the present anyone who would do it. For what the learned Hermann has begun in his writings, published in my defence against Nieuwentiit [*sic*], is not yet complete.

For I have, beside the mathematical infinitesimal calculus, a method also for use in Physics, of which an example was given in the *Nouvelles de la République des Lettres*; and both of these I include under the Law of Continuity; and adhering to this, I have shown that the rules of the renowned philosophers Descartes and Malebranche were sufficient in themselves to attack all problems on Motion.

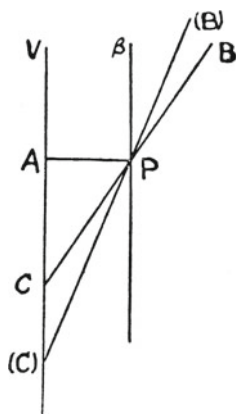
I take for granted the following postulate:

In any supposed transition, ending in any terminus, it is permissible to institute a general reasoning, in which the final terminus may also be included.

For example, if A and B are any two quantities, of which the former is the greater and the latter is the less, and while B remains the same, it is supposed that A is continually diminished, until A becomes equal to B; then it will be permissible to include under a general reasoning the prior cases in which A was greater than B, and also the ultimate case in which the difference vanishes and A is equal to B. Similarly, if two bodies are in motion at the same time, and it is assumed that while the motion of B remains the same, the velocity of A is continually diminished until it vanishes altogether, or the speed of A becomes zero; it will be permissible to include this case with the case of the motion of B under one general reasoning. We do the same thing in geometry, when two straight lines are taken, produced in any manner, one VA being given in position or remaining in the same site, the other BP passing through a given point P, and varying in position while the point P remains fixed; at first indeed converging toward the line VA and meeting it in the point C; then, as the angle of inclination VCA [*sic*, this should read PCA] is continually diminished, meeting VA in some more remote point (C), until at length from BP, through the position (B)P, it comes to βP , in which the straight line no longer converges toward VA, but is parallel to it, and C is an impossible or imaginary point (Fig. 2.40). With this supposition it is permissible to include under some one general reasoning not only all the intermediate cases such as (B)P but also the ultimate case βP .

Hence also it comes to pass that we include as one case ellipses and the parabola, just as if A is considered to be one focus of an ellipse (of which V is the given vertex), and this focus remains fixed, while the other focus is variable as we pass from ellipse to ellipse, until at length (in the case when the line BP, by its intersection with the line VA, gives the variable focus) the focus C becomes evanescent¹⁸⁰ or impossible, in which case the ellipse passes into a parabola. Hence it is permissible with our postulate that a parabola should

¹⁸⁰Child adds a footnote here explaining that "evanescent" should be read "vanishing into the far distance".

Fig. 2.40 Leibniz's diagram

be considered with ellipses under a common reasoning. Just as it is common practice to make use of this method in geometrical constructions, when they include under one general construction many different cases, noting that in a certain case the converging straight line passes into a parallel straight line, the angle between it and another straight line vanishing.¹⁸¹

The Law of Continuity is not very clearly expressed. The simplest interpretation would be that, if a property held for all values of o as $o \rightarrow 0$ then it must also hold at 0. Obviously, this does not hold for all properties — for example for $P(o) : o \neq 0$. Presumably it holds in some sense for some crucial properties, but its value as a postulate rests on a delineation of the properties to which the Law applies. Without such the Law of Continuity can be no more than a heuristic useful in finding solutions, but not in establishing them.

In Leibniz's early publications he seemed to believe in the existence of infinitesimals, and a number of his followers accepted them. In a letter to François Pinsson published in 1701, he surprised many of his followers:

For instead of the infinite or infinitely small, one takes quantities as large, or as small, as necessary in order that the error be smaller than the given error, so that one differs from Archimedes's style only in the expressions, which are more direct in our method and more conform to the art of invention.¹⁸²

Another quote:

It will be sufficient if, when we speak of infinitely great (or more strictly unlimited), or of infinitely small quantities (i. e., the very least of those within our knowledge), it is understood that we mean quantities that are indefinitely great or indefinitely small, i. e., as great as you please, or as small as you please, so that the error that any one may assign may be less than a certain assigned quantity. Also, since in general it will appear that, when any small error is assigned, it can be shown that it should be less, it follows that the error is absolutely nothing; an almost exactly similar kind of argument is used in different places by Euclid,

¹⁸¹Child, *Early Mathematical Manuscripts...*, *op. cit.*, pp. 146–148.

¹⁸²Quoted in translation in H.J.M. Bos, *Differentials, Higher Order Differentials and the Derivative in the Leibnizian Calculus*, dissertation, Rijksuniversiteit te Utrecht, 1973, p. 73.

Theodosius and others; and this seemed to them to be a wonderful thing, although it could not be denied that it was perfectly true that, from the very thing that was assumed as an error, it could be inferred that the error was non-existent. Thus, by infinitely great and infinitely small, we understand something indefinitely great, or something indefinitely small, so that each conducts itself as a sort of class, and not merely as the last thing of a class. If any one wishes to understand these as the ultimate things, or as truly infinite, it can be done, and that too without falling back upon a controversy about the reality of extensions, or of infinite continuums in general, or of the infinitely small, ay, even though he think that such things are utterly impossible; it will be sufficient simply to make use of them as a tool that has advantages for the purpose of the calculation, just as the algebraists retain imaginary roots with great profit. For they contain a handy means of reckoning, as can manifestly be verified in every case in a rigorous manner by the method already stated.¹⁸³

Phrases like Newton's "come nearer than any given difference" (p. 106, above) or Leibniz's "one takes quantities as large, or as small, as necessary in order that the error be smaller than the given error" (in the quotation from the letter to Pinsson), especially the latter, are awfully close to the modern ϵ - δ definition of limit and one wonders how it took so long for that definition to crystallise. For it seems only first to have been used and almost expressed by Bolzano in 1816, then by Cauchy in 1821, finally to be expressed clearly and unambiguously in modern form sometime in the mid-19th century by Weierstrass and his students (like Heine, cited in Sect. 1). Did Berkeley's critique accelerate or decelerate the process?

The immediate effect of Berkeley's attack on infidel mathematicians and their limits and infinitesimals seems to have been beneficial. Two years after the appearance of *The Analyst* no fewer than four textbooks on the method of fluxions were published in England, more than all the previously existing expositions put together. Also, Colin Maclaurin's (1698–1746) later text of 1742 on fluxions was a response to Berkeley. And there were other published responses to Berkeley. On the continent the debate over infinitesimals raged on. While analysis continued to grow and develop, its foundations remained shaky.

Our modern foundation for the Calculus rests on two pillars — the familiar ϵ - δ definitions of limit (pointwise, uniform, etc.) and a characterisation of the real number field (completeness, archimedean order). The latter only began to emerge in the mid-1810s in the work of Bolzano and reached final form in 1872 with an explosion of papers offering different methods of reaching a common solution. The former, however, was seemingly within reach of everyone — or, at least, anyone who chanced upon the key phrases of Newton and Leibniz, who was not attracted by the lure of infinitesimals, and who knew to ignore everything else.

Perhaps, without a clear concept of the completeness of the real numbers, the problem was just too intractable. In any event, the serious attempts to found the Calculus rigorously in the 18th century went in a different direction. I would even venture to say they went astray in that the approaches were doomed to failure.

The two commonly cited attempts to found the Calculus without recourse to limits were the *residual analysis* of John Landen (1719–1790) and the formal power series of Joseph Louis Lagrange (1736–1813). Landen wrote two books on the subject, a

¹⁸³Child, *Early Mathematical Manuscripts...*, *op. cit.*, p. 150.

short volume of 44 pages, *A Discourse Concerning the Residual Analysis* (1758), and a longer exposition, *The Residual Analysis* (1764). Lagrange also wrote two books on his approach, *Théorie des fonctions analytiques, contenant les principes du calcul différentiel dégagés de toute considération d'infiniment petits ou d'évanouissans, de limites ou de fluxions, et réduits à l'analyse algébrique des quantités finies* [*Theory of Analytic Functions...*] (1797), usually referred to as *Théorie des Fonctions*, and *Leçons sur le calcul des fonctions* [*Lessons on the Calculus of Functions*] (1806).

Landen's starting point was Newton's Binomial Theorem,

$$(1+x)^q = 1 + qx + \frac{q(q-1)}{2}x^2 + \frac{q(q-1)q-2}{6}x^3 + \dots,$$

for rational q . The right-hand-side generally requires $|x| < 1$ to guarantee convergence, a fact widely recognised but not emphasised. His approach was to consider and simplify an expression,

$$\frac{f(x) - f(y)}{x - y},$$

obtaining an equation that generally held for $x \neq y$ and claiming it must therefore hold for $x = y$ as well. Using a simple algebraic identity

$$\frac{u^{m/n} - v^{m/n}}{u - v} = u^{m/n-1} \frac{1 + \frac{v}{u} + \left(\frac{v}{u}\right)^2 + \dots + \left(\frac{v}{u}\right)^{m-1}}{1 + \left(\frac{v}{u}\right)^{m/n} + \left(\frac{v}{u}\right)^{2m/n} + \dots + \left(\frac{v}{u}\right)^{(n-1)m/n}} \quad (2.34)$$

and an assumed expansion

$$(1+x)^{m/n} = 1 + ax + bx^2 + cx^3 + \dots,$$

he derived

$$\frac{m}{n}(1+x)^{m/n-1} = a + 2bx + 3cx^2 + \dots$$

by algebraic manipulation. Multiplication of both sides by $1+x$ allowed him to determine a, b, c, \dots in succession. Following this he proceeded to apply (2.34) to a variety of tangent and max/min problems. I forego discussion of these matters, referring the interested reader to the literature.¹⁸⁴

Lagrange's approach was to assume every function f could be expanded into a power series:

$$f(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \dots$$

¹⁸⁴Landen's *Discourse* is available online and a print edition by Gale, 2010, of Book I of *The Residual Analysis* exists. Additionally, excerpts from *Discourse* are reproduced in Struik, *op. cit.*, pp. 386–388 and Stedall, *op. cit.*, pp. 398–401. I also refer to Smoryński, *Treatise, op. cit.*, pp. 148–151, for a detailed account of Landen's "proof" of the Binomial Theorem.

He defined the derivative by $f'(a) = a_1$, wrote f' as a power series

$$f'(x) = b_0 + b_1(x - a) + b_2(x - a)^2 + \dots,$$

noted that $f''(a) = b_1$ and that f'' could be expanded into a power series itself, etc., and went on to calculate

$$f'(a) = a_1, \quad f''(a) = 2a_2, \quad f'''(a) = 6a_3, \quad \dots$$

Lagrange's approach was popular for a while,¹⁸⁵ but its limitations were startlingly revealed by Cauchy in his *Résumé des leçons* in 1823 where he produced an analytically expressible function that equalled its power series expansion at exactly one point and deduced from it that two distinct analytic functions could have the same power series expansions (though they could not, of course, both equal these expansions). Again I refer the interested reader to the literature.¹⁸⁶

Landen's work is of fleeting importance, of interest today only as an example of the nonlinear development of mathematics: Mathematics is not an unbroken progressive development; it occasionally goes down blind alleys, and the residual analysis was one of them. Lagrange, though he based his approach on the false assumption that every function could be expanded into a power series and used this "fact" in proving theorems, produced results of lasting importance. These results required new proofs, but he made the discoveries and provided starting points for some of these new proofs. One of his results was the Mean Value Theorem, which contribution will be discussed in the next chapter. For now, we skip ahead to Lagrange's successors and the definition of the derivative.

2.3.7 The Derivative Defined

The modern definition of the derivative is given in terms of limits, which are themselves defined in terms of approximations. The first serious analyses of approxima-

¹⁸⁵His term "derivative" and notation f' for the derivative are still used today.

¹⁸⁶Both books by Lagrange are available online. English translations of excerpts from *Théorie des fonctions* can be found in Struik, *op. cit.*, pp. 389–391, and Stedall, *op. cit.*, pp. 404–406. Other discussions of Lagrange's approach can be found in Edwards, *op. cit.*, pp. 296–299, and Smoryński, *Formalism, op. cit.*, pp. 127–135. This last reference, incidentally, includes in Exercise 6.6 of Chapter II, pp. 184–185, an outline of Cauchy's result mentioned above.

tions were performed independently by Lagrange¹⁸⁷ and d'Alembert.¹⁸⁸ And three decades later, in his two books on founding the Calculus on power series, Lagrange presented such a treatment for these series.

Unlike Lagrange, d'Alembert was a firm believer in founding the Calculus on the notion of limit. In volume IX (1765) of Diderot's *Encyclopédie*, d'Alembert writes

LIMIT (*Mathematics*). One says that a magnitude is the *limit* of another magnitude, when the second may approach the first more closely than by a given quantity, as small as one wishes, moreover without the magnitude which approaches being allowed ever to surpass the magnitude that it approaches; so that the difference between such a quantity and its *limit* is absolutely unassignable...

The theory of *limits* is the foundation of the true justification of the differential calculus. See DIFFERENTIAL, FLUXION, EXHAUSTION, INFINITE. Strictly speaking, the *limit* never coincides, or never becomes equal to the quantity of which it is the *limit*, but the latter approaches it ever more closely, and may differ from it as little as one wishes.¹⁸⁹

This is no improvement on Newton's proclamation about limits of half a century earlier.¹⁹⁰

Oddly enough, despite his distaste for limits, Lagrange came closer to our modern definition of limit than d'Alembert even though both men all but proved certain limits to exist using their numerical analyses of convergence. Lagrange, in fact, came close to the definition of continuity:

And, Lagrange said, "The course of the curve will necessarily be *continuous* from this point; thus it will, little by little, approach the axis before cutting it, and approach it, consequently, within a quantity less than any given quantity." This characterization of continuity appears geometric. But Lagrange rendered it algebraic: "So we can always find an abscissa h corresponding to an ordinate less than any given quantity; and then all smaller values of h correspond also to ordinates less than the given quantity." This is a far cry from "insensible degrees" or "infinitely small changes." But it is not far from this characterization of continuity at $h = 0$ to the Bolzano-Cauchy definitions of continuity in general. Even though Lagrange himself did not take his characterization to be the defining property of continuous function, he had for the first time stated, in terms of inequalities, what Cauchy and Bolzano later recognized as such.¹⁹¹

Taking their cue from Lagrange, Bolzano and Cauchy defined continuity, as we saw in the preceding section, and gave rigorous ϵ - δ proofs of limit theorems, albeit not always using this notation.

It is high time we defined the notions of limit and derivative.

¹⁸⁷J.L. Lagrange, "Sur la résolution des équations numériques, et additions au mémoire sur la résolution des équations numériques", *Mémoires de l'Académie...Berlin* 23 (1767), pp. 311–352 and 24 (1768), pp. 111–180; reprinted in volume 2 of *Oeuvres de Lagrange*, Gauthier-Villars, Paris, 1867–1882.

¹⁸⁸"Réflexions sur les suites et sur les racines imaginaires", in: J. d'Alembert, *Opuscules mathématiques*, vol. 5, Briasson, Paris, 1768, pp. 171–215. An annotated English translation of the relevant portions can be found in Smoryński, *Treatise*, *op. cit.*, pp. 182–188.

¹⁸⁹English translation from: Stedall, *op. cit.*, pp. 297–298. Stedall includes also excerpts on limits from Wallis, Newton, Maclaurin, and Cauchy.

¹⁹⁰Cf. p. 95, above.

¹⁹¹Judith V. Grabiner, *Origins*, *op. cit.* p. 95.

2.3.9 Definition Let f be a function defined everywhere in an interval I with the possible exception of a point a , with $a \in I$. A number L is the *limit of f as x approaches a* , written

$$\lim_{x \rightarrow a} f(x) = L,$$

if, for any $\epsilon > 0$ there is a $\delta > 0$ such that for all $x \in I$,

$$0 < |x - a| < \delta \Rightarrow |f(x) - L| < \epsilon.$$

The main difference between this and the definition of the continuity of f at a is the clause $0 < |x - a|$, i.e., $x \neq a$, in the premise of the final implication. For, it is not assumed that $f(a)$ is defined, and, in any event, any possible value of f at a is irrelevant in determining how f behaves as x approaches a .

The quintessential functional limit is the derivative:

2.3.10 Definitions Let I be an interval, $a \in I$, and $f : I \rightarrow \mathbb{R}$. f is *differentiable at a* if

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

exists. This limit, when it exists, is called the *derivative of f at a* and is denoted $f'(a)$. The function f' mapping a to $f'(a)$ is called the *derivative of f* .

The first to take the hint and give rigorous ϵ - δ proofs in print, as far as I know, was Bernard Bolzano in his *Der binomische Lehrsatz, und als Folgerung aus ihm der polynomische, und die Reihen, die zur Berechnung der Logarithmen und Exponentialgrößen dienen, genauer als bisher erwiesen*¹⁹² [The Binomial Theorem, and as a Consequence from it the Polynomial Theorem and the Series which serve for the Calculation of Logarithmic and Exponential Quantities, proved more strictly than before]. He did not formulate it as explicitly as we do today, and it looks a lot like he is dealing with infinitesimals. However, if one looks carefully at the details of his treatment, one sees he is using standard ϵ - δ arguments.

Bolzano is quite adamant about his avoidance of infinitesimals:

I have, not only in this work, already generally made the rule, by which instead of the so-called *infinitely small quantities* I have used with equal success the concept of *quantities which can be made smaller than any given value* or, (to avoid monotony, I sometimes call, albeit less accurately) the quantities *which can be made as small as one wishes*. Hopefully, one will not misjudge the difference between quantities of this sort and those which one otherwise imagines under the scope of the infinitely small. The demand, to imagine a quantity (I mean a variable one) which can always be made smaller than one has already taken it to be, and generally smaller than any given value, really contains nothing that can be offensive to anyone. Must one not see, rather, that frequently there are such quantities in space as well as in time? Against this the thought of a quantity which can not only be *assumed* smaller, but rather already *is* smaller than any mere given one but also [than] any *supposed*, i.e.,

¹⁹²Prague, 1816; English translation in: Russ, *op. cit.*.

thinkable, quantity; should this not be contradictory? So reads the customary definition of the infinitely small.¹⁹³

Convention. To indicate a quantity which can become smaller than any given value, we choose the symbols ω , Ω or something similar.¹⁹⁴

The “something similar” is generally an ω or Ω with a superscript — a numeral written directly above the omega. The two variants have distinct uses: ω is a variable quantity we can make as small as we wish at will and Ω is a variable quantity that can be made as small as desired as a consequence of our choices for the sizes of various ω ’s. Having stated these conventions, he proves for these variable quantities versions of closure properties of infinitesimals used in proving various limit theorems. For example, the theorem to the effect that the limit of a sum is the sum of the limits, which traditionally followed from the fact that the sum of infinitesimals is itself infinitesimal, is rendered:¹⁹⁵

Lemma. If each of the quantities $\omega, \overset{(1)}{\omega}, \overset{(2)}{\omega}, \dots, \overset{(m)}{\omega}$ can become as small as one wishes while the (finite) number of them does not change, then their algebraic *sum* or *difference* is also a quantity which can become as small as one wishes, i.e.

$$\omega \pm \overset{(1)}{\omega} \pm \overset{(2)}{\omega} \pm \dots \pm \overset{(m)}{\omega} = \Omega.$$

He proves this by showing the sum Ω can be made less than D (think: ϵ) in absolute value by choosing each ω less than $D/(m+1)$ (think: δ) in absolute value.

With respect to differentiation, in Sect. 23 he determines the derivative of $f(x) = x^n$ for arbitrary real a and $x > 0$.¹⁹⁶

Lemma. The quantity

$$\frac{(x + \omega)^n - x^n}{\omega}$$

can be brought as close to the value nx^{n-1} as one wishes, if ω is taken small enough; n and x may denote whatever desired, if only x is not $= 0$; i.e.,

$$\frac{(x + \omega)^n - x^n}{\omega} = nx^{n-1} + \Omega.$$

His proof is not perfect: His treatment of the case for irrational n is garbled, but the argument is overall correct. He shows that the difference,

$$\frac{(x + \omega)^n - x^n}{\omega} - nx^{n-1},$$

can be made smaller than any given $D (= \epsilon)$ by showing how small ω needs to be taken (i.e., he finds δ).

¹⁹³*Ibid.*, p. v; Russ, p. 158.

¹⁹⁴*Ibid.*, p. 15; Russ, p. 173.

¹⁹⁵*Ibid.*, p. 15; Russ, p. 173.

¹⁹⁶*Ibid.*, p. 20; Russ, p. 176.

The ω, Ω notation is a convenient shorthand for our use of δ, ϵ , respectively. However, when things get complicated it is not sufficient as dependencies can get confused. Bolzano's next big result, that, if f_0, f_1, \dots converges to f , then f'_0, f'_1, \dots converges to f' , is simply false.

Bolzano does not offer a definition of the derivative nor even introduce the term in his first paper of 1816. In his later unpublished "Functionenlehre", he would do so, even defining one-sided derivatives. By then Cauchy had published his two famous textbooks on Analysis, treating differentiation in the *Résumé des leçons* in 1823.

As with continuity, Cauchy did not define differentiability at a point, but on an interval. And, as with continuity, he built some uniformity into the definition. Today we define differentiability on an interval as follows.

2.3.11 Definition Let I be an interval and $f : I \rightarrow \mathbb{R}$. f is *differentiable on I* if, for every $x \in I$, f is differentiable at x .

Cauchy's introduction of the derivative in the *Résumé des leçons* gives a relatively loose definition:

THIRD LESSON
Derivatives of Functions of a single Variable

WHEN the function $y = f(x)$ remains continuous between two limits of the variable x , and when one assigns to this variable a value between the limits at hand and confers an infinitely small increment to this variable, an infinitely small increment of the function itself is produced. Consequently, if one puts $\Delta x = i$, the two terms of the *ratio of differences*

$$\frac{\Delta y}{\Delta x} = \frac{f(x+i) - f(x)}{i}$$

will be infinitely small quantities. But, whereas when these two terms approach indefinitely and simultaneously the limit zero, the ratio itself will be able to converge towards another limit, either positive, or negative. This limit, when it exists, has a determinate value, for each particular value of x ; but this varies with x . Thus, for example, if one takes $f(x) = x^m$, m denoting a whole number, the ratio of the infinitely small differences becomes

$$\frac{(x+i)^m - x^m}{i} = mx^{m-1} + \frac{m(m-1)}{1.2}x^{m-2}i + \dots + i^{m-1}$$

and it has for a limit the quantity mx^{m-1} , that is to say, a new function of the variable x . It will be the same in general; however, the form of the new function which serves up the limit of the ratio $\frac{f(x+i)-f(x)}{i}$ depends on the form of the proposed function $y = f(x)$. To indicate this dependence, we give the new function the name of *derived function*, and we denote it, with the aid of an accent, by the notation

$$y' \text{ or } f'(x).$$

Several pages later, when it comes time to prove the theorem on which his proof of the Mean Value Theorem depends, he introduces ϵ and δ .¹⁹⁷

¹⁹⁷Cauchy, *Résumé*, *op. cit.*, p. 9. After making many of my translations from the *Résumé* for this book, I learned of a complete translation of the work by Dennis M. Cates. There are two versions of

Denote by δ, ε , two very small numbers, the first being chosen of such kind which, for the numerical value¹⁹⁸ of i less than δ , and for any value whatsoever of x between the limits x_0, X , the ratio

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

always lies above $f'(x) - \varepsilon$, and below $f'(x) + \varepsilon$.¹⁹⁹

The thing to note is that, given ε , the same δ is claimed to work for all x in the given interval:

2.3.12 Definition Let I be an interval and $f : I \rightarrow \mathbb{R}$. f is *uniformly differentiable* on I with derivative f' if for all $\varepsilon > 0$ there is a $\delta > 0$ such that for all $x \in I$ and all h ,

$$0 < |h| < \delta \text{ \& } x+h \in I \Rightarrow \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| < \varepsilon.$$

A differentiable function is readily seen to be continuous, but it does not imply the same for the derivative (Cf. Remark 2.2.6 or Exercise 2.3.20.). Cauchy's uniform differentiability, however, does.

2.3.13 Lemma Let f be uniformly differentiable on I . Then: f' is uniformly continuous on I .

Proof Let $\varepsilon > 0$ be given, choose $\delta > 0$ so that for all $y \in I$,

$$0 < |h| < \delta \text{ \& } y+h \in I \Rightarrow \left| \frac{f(y+h) - f(y)}{h} - f'(y) \right| < \frac{\varepsilon}{2},$$

and note

$$\begin{aligned} |f'(x+h) - f'(x)| &= \left| f'(x+h) - \frac{f(x) - f(x+h)}{-h} - \frac{f(x+h) - f(x)}{-h} - f'(x) \right| \\ &\leq \left| f'(x+h) - \frac{f(x) - f(x+h)}{-h} \right| + \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2}. \end{aligned}$$

□

(Footnote 197 continued)

this translation, an expensive annotated edition, *A Guide to Cauchy's Calculus; A Translation and Analysis of Calculus Infinitesimal*, Fairview Academic Press, Walnut Creek (California), 2011, and a more affordable student edition, *Cauchy's Calculus Infinitesimal; A Complete English Translation*, same publisher, 2012. In comparing my translations with his, I find the differences minor and have kept my own. Nonetheless, I shall give page references to the less expensive copy which is now in my possession. The reference in the present case is to p. 7.

¹⁹⁸I.e., the absolute value.

¹⁹⁹Cauchy, *Résumé*, *op. cit.*, p. 27; Cates, *op. cit.*, p. 23.

2.3.14 Definition Let I be an interval and $f : I \rightarrow \mathbb{R}$. f is *continuously differentiable* on I if f is differentiable on I and f' is continuous on I .

An immediate corollary to Lemma 2.3.13 is that a uniformly differentiable function is continuously differentiable. Thus

$$\text{uniform differentiability} \Rightarrow \text{continuous differentiability} \Rightarrow \text{differentiability}.$$

The converse implications fail in general.

2.3.15 Exercise i. Show that Darboux's function,

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

is everywhere differentiable, but f' is not continuous at $x = 0$.

ii. Show that $f(x) = 1/x$ is continuously differentiable on $(0, 1)$, but f' is not uniformly continuous there, whence f is not uniformly differentiable on this interval. [In part i you may assume the usual rules for differentiation already to have been established.]

With a little more theory, it can be shown that the uniform continuity of f' on a closed, bounded interval $[a, b]$ entails the uniform differentiability of f on $[a, b]$.²⁰⁰

With three distinct candidates for a notion of differentiability to choose from — differentiability on an interval, continuous differentiability on an interval, and uniform differentiability on an interval — the question arises: Which notion is fundamental and which are variants — stronger forms or weaker generalisations of the concept? Obviously, the nomenclature gives away the conventional answer: Differentiability is the fundamental concept, while continuous and uniform differentiability are modifications. The reason for this is theoretical. In practice, thanks to the Mean Value Theorem, most results of interest depend only on differentiability, although this was not always the case.

In 1816 Bolzano's interest in differentiation was primarily in differentiating a few specific functions. His paper on the Binomial Theorem does not even refer to the derivative by name, much less establish any of its properties. In contrast, Cauchy set out immediately to provide rigorous proofs for those properties of the derivative Lagrange had discovered, and to derive some of his own. This does not include Lemma 2.3.13, which he was apparently unaware of. His statement of the Mean Value Theorem, for example, explicitly assumes the continuity of f' as an added assumption to the uniform differentiability of f . Indeed, his proof of the theorem from which he derived the Mean Value Theorem itself relies explicitly on the continuity of f' — as we will see in the next chapter. Many of the early proofs of theorems about differentiable functions were valid only for uniformly differentiable functions and, were it not for the validity of the Mean Value Theorem for the broader class of

²⁰⁰Cf. pp. 301–304, below, for details.

differentiable functions, we might well regard uniform differentiability today as the fundamental concept, terming it “differentiability” and calling differentiability itself by some derivative name like “weak differentiability” or “generalised differentiability”.

Some of the more interesting of these theorems about differentiable functions will be discussed in the next two sections. For now we consider only a few simpler properties. In a course on the Calculus, after defining differentiation, one generally differentiates a few simple functions, derives the rules for differentiating functions built up from simpler ones by the arithmetic operations and composition, and then applies differentiation to find tangent lines and maxima and minima.

The student who has had an honest Calculus course in which he or she was required to perform some simple ϵ - δ proofs should have no difficulty proving some of the basic computation rules:

$$\begin{aligned} f'(x) &= 0 \text{ for any constant function } f \\ f'(x) &= 1 \text{ for } f(x) = x \\ (f \pm g)'(x) &= f'(x) \pm g'(x) \\ (cf)'(x) &= cf'(x) \text{ for any constant } c \\ (f \cdot g)'(x) &= f'(x)g(x) + f(x)g'(x) \\ (1/f)'(x) &= \frac{-f'(x)}{f(x)^2}, \text{ whenever } f(x) \neq 0 \\ (f/g)'(x) &= \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}, \text{ whenever } g(x) \neq 0. \end{aligned}$$

More difficult are the rules for differentiating the trigonometric, exponential, and logarithmic functions, the demonstrations of which few students can master in a first Calculus course. Likewise, the derivation of the Chain Rule,

$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x),$$

and the existence assertion of the Inverse Function Theorem,

$$(f^{-1})'(y) = \frac{1}{f'(x)} = \frac{1}{f'(f^{-1}(y))}, \text{ when } y = f(x) \text{ is one-to-one and } f'(x) \neq 0,$$

may well require greater experience in formulating proofs. The reader who has just completed a Calculus course may find this a perfect occasion to review these proofs, carrying out the proofs for the easy algebraic rules cited above and looking up the demonstrations for the transcendental functions, the Chain Rule, and the Inverse Function Theorem in his or her Calculus text.²⁰¹ For my own part, I have resisted the temptation to repeat the proofs here as the details do not bear directly on our main topic, the Mean Value Theorem, which, believe it or not, we are rapidly approaching.

²⁰¹Particularly nice proofs of the Chain Rule and the Inverse Function Theorem can be found in Jan Mikusiński and Piotr Mikusiński, *An Introduction to Analysis: From Number to Integral*, John Wiley & Sons, New York, 1993, pp. 123–124 and 132–133, respectively.

The value of having an official definition of the derivative is not so much in being able to prove that this or that function is differentiable with such and such a derivative. For a century and a half, from Newton to Cauchy, mathematicians could find derivatives without a formal definition or rigorously proven theorems about derivatives. Without such a definition, however, one could not specify exactly the conditions under which theorems held. In the words of Niels Henrik Abel (1802–1829), theorems could “admit exceptions”, and it is precise definitions and strict rigour in proofs that explain and, with luck, delineate these exceptions.

Historically, the exceptions were rare. Just as one had had a paucity of curves for a couple of millennia, the new analysis had a restricted collection of functions. Initially they were algebraic or trigonometric and were quite smooth, continuous and differentiable at all but a few isolated points. As the concept of function crystallised and the stock of functions grew, it was realised that this was false. The first example was given in the 1830s by Bolzano in his “Functionenlehre”,²⁰² wherein he constructed a continuous function which failed to have a derivative anywhere, though he only proved it to fail to have a derivative on a dense set of points. Bolzano’s work, however, was unfinished and his result was not published until the 20th century. In the 1850s Bernhard Riemann (1826–1866) produced examples of functions which failed to have derivatives on dense sets of points, and around 1860 Charles Cellierier (1818–1889) produced a (*continuous*) *nowhere differentiable function*, as continuous functions possessing no derivatives at any point are now called. His example,

$$C(x) = \sum_{n=0}^{\infty} \frac{1}{a^n} \sin(a^n x),$$

where $a > 1000$ is an even integer, was not published until 1890. In the meantime, in 1872, Weierstrass startled the mathematical world when he announced in a lecture that

$$W(x) = \sum_{n=0}^{\infty} b^n \cos(a^n x \pi)$$

is nowhere differentiable when a is an odd positive integer, $0 < b < 1$, and $ab > 1 + \frac{3}{2}\pi$.²⁰³ His proof was not published until 1895, by which time his followers had published a number of such examples. And by then Peano had published (1890) his continuous space-filling curve $\gamma_P(t) = \langle x_P(t), y_P(t) \rangle$ and announced that neither of its component functions x_P nor y_P was differentiable anywhere in their domain.²⁰⁴

2.3.16 Exercise Prove Peano’s assertion.

Obviously, although one can construct these functions without a definition of differentiation, one would be hard put to prove their nowhere differentiability without

²⁰²Russ, *op. cit.*, pp. 487–489, 507–508.

²⁰³In one course, at least, given in 1874, he said that the conditions could be relaxed to a being an integer > 1 and $ab > 1$, but that the proof was more difficult under these more general conditions.

²⁰⁴Cf. Sect. 2.2.3, above.

such a definition. And one similarly needs such a definition to prove the various properties of derivatives as those used in the familiar procedures for finding tangents and solving maxima/minima problems. Our problem is, of course, to give a formal definition of the tangent to a curve in order to give scope and meaning to the Mean Value Theorem, and to justify the methods of finding maxima/minima and thereby lend validity to the proof of the Mean Value Theorem outlined in the Preface.

Let us begin with the merely technical problem of maxima and minima.

The crucial result is quite simple.

2.3.17 Lemma *Let $f : [a, b] \rightarrow \mathbb{R}$ assume a maximum or a minimum at $c \in (a, b)$ and assume $f'(c)$ exists. Then: $f'(c) = 0$.*

That some form of the Lemma holds is intuitively obvious. We have already quoted Newton (p. 102, above) on this. In Europe, the principle seems first to have been enunciated by Johannes Kepler (1571–1630). The principle, in fact, is said to be found in a work of Bhāskara II (1114–1185) dating from 1150.²⁰⁵ However, prior to Cauchy there was no attempt to determine the conditions on f under which the conclusion held. For Cauchy, of course, f had to be uniformly differentiable in an interval and his result was not as general as that above.

The proof of Lemma 2.3.17 reduces to another pair of Lemmas.

2.3.18 Lemma *Let $f : I \rightarrow \mathbb{R}$ be differentiable at $c \in I$ and suppose $f'(c) > 0$. There is a $\delta > 0$ such that, for all $x \in I$,*

$$c - \delta < x < c \Rightarrow f(x) < f(c) \text{ and } c < x < c + \delta \Rightarrow f(c) < f(x).$$

Proof Let $\epsilon = f'(c)/2$ and choose $\delta > 0$ so that, for $x \in I$,

$$0 < |x - c| < \delta \Rightarrow \left| \frac{f(x) - f(c)}{x - c} - f'(c) \right| < \epsilon.$$

For $c < x < c + \delta$, one has

$$-\frac{f'(c)}{2} < \frac{f(x) - f(c)}{x - c} - f'(c) < \frac{f'(c)}{2},$$

whence

$$0 < \frac{f'(c)}{2} < \frac{f(x) - f(c)}{x - c}.$$

²⁰⁵Bibhutibhushan Datta and Avadesh Narayan Singh (Kripa Shankar Shukla, reviser), “Use of Calculus in Hindu mathematics”, *Indian Journal of History of Science* 19, No. 2 (1984), pp. 95–109; here: p. 98. Mediaeval Hindu mathematicians, particularly in the Kerala region, were several centuries ahead of the Europeans in many areas, including the beginnings of the infinitesimal calculus. In the last few decades some primary sources have been published in English translation, but not enough yet for one to develop an accurate picture of their state of knowledge. The internet is rife with references to the Hindu origins of the above Lemma and the Mean Value Theorem, but they tend to offer no details. We discuss the matter in greater detail in Sect. 2.3 of Chap. 3.

$x - c$ being positive, multiplication by it will preserve the inequality:

$$0 < f(x) - f(c), \text{ i.e. } f(c) < f(x).$$

For $c - \delta < x < c$, we again have

$$0 < \frac{f(x) - f(c)}{x - c},$$

but $x - c$ is negative, whence multiplication by it reverses the inequality:

$$0 > f(x) - f(c),$$

i.e. $f(c) > f(x)$. □

2.3.19 Lemma Let $f : I \rightarrow \mathbb{R}$ be differentiable at $c \in I$ and suppose $f'(c) < 0$. There is a $\delta > 0$ such that, for all $x \in I$,

$$c - \delta < x < c \Rightarrow f(c) < f(x) \text{ and } c < x < c + \delta \Rightarrow f(x) < f(c).$$

I leave the proof as an exercise to the reader.

Proof of Lemma 2.3.17. Let $c \in (a, b)$ be where f assumes a maximum in $[a, b]$. If $f'(c) > 0$, choose $\delta > 0$ according to Lemma 2.3.18. Because c is an interior point there is some $x \in I$ satisfying $c < x < c + \delta$. But for such x , $f(x) > f(c)$, contradicting the maximality of f at c . Similarly Lemma 2.3.19 tells us that $f'(c)$ cannot be less than 0. As $f'(c)$ exists by assumption, we must have $f'(c) = 0$.

The proof for f assuming the minimum value is similar. □

Note that Lemma 2.3.18, for example, says that if $f'(c) > 0$, then for x, y sufficiently close to c , one has

$$x < c < y \Rightarrow f(x) < f(c) < f(y).$$

One sometimes expresses this in words as *f is increasing at c*, which is not the same as saying that *f is increasing in a neighbourhood of c*,

$$x < y \Rightarrow f(x) < f(y)$$

for x, y sufficiently close to c , as the result of the following exercise shows.

2.3.20 Exercise Define

$$f(x) = \begin{cases} x + x^2 \sin\left(\frac{1}{x^2}\right), & x \neq 0 \\ 0, & x = 0. \end{cases}$$

- i. Show directly using the definition of the derivative that $f'(0) = 1 > 0$.

- ii. Show that $f'(1/\sqrt{2n\pi}) = 1 - 2\sqrt{2n\pi} < 0$ for positive integers n .
- iii. Use Lemma 2.3.19 to conclude that, for any $\delta > 0$, there are $x, y \in (0, \delta)$ such that $x < y$ and $f(x) > f(y)$.

Lemma 2.3.17 is the key lemma needed to complete the proof of the Mean Value Theorem. However, we have raised the issue of the relation between the derivative's sign and whether or not the function is increasing in an interval and we might as well answer it here. One would expect, from Lemma 2.3.18 that, if the derivative is always positive in an interval, the function is increasing at every point and thus must be increasing in a more global sense. This is true, but the proof is subtle.

2.3.21 Corollary (Strictly Increasing Function Theorem) *Let $f : I \rightarrow \mathbb{R}$ be differentiable on the interval in question and suppose $f'(x) > 0$ for all $x \in I$. Then f is strictly increasing on I : for all $x, y \in I, x < y \Rightarrow f(x) < f(y)$.*

Proof We use the same sort of continuous induction used in the proof of the Intermediate and Extreme Value Theorems and the Uniform Continuity Theorem.

Let $x \in I$ be any element of I other than the right endpoint if I has one, and define

$$X = \{z \in I \mid x < z \text{ \& } \forall y \in I (x < y < z \Rightarrow f(x) < f(y))\}.$$

By Lemma 2.3.18, X is nonempty: for some $\delta > 0, x + \delta \in X$.

If X is unbounded, then for all $y \in I$ there is some $z \in X$ such that $y < z$. If $x < y$, then $x < y < z$ and it follows that $f(x) < f(y)$.

If X is bounded, it has a least upper bound z_0 . Let $x < y < z_0$ and choose $z \in X$ such that $y < z$. Then

$$x < y < z \in X \Rightarrow f(x) < f(y).$$

Either $z_0 \in I$ or $z_0 \notin I$.

If $z_0 \notin I$, then $z_0 > y$ for all $y \in I$ and we have shown that, for all $y \in I$, if $x < y$ then $f(x) < f(y)$.

Thus, assume $z_0 \in I$. Then $z_0 \in X$.

We first apply Lemma 2.3.18 to conclude $f(x) < f(z_0)$: Choose $\delta > 0$ according to the Lemma so that $x < z_0 - \delta$ and for all $y \in I$

$$z_0 - \delta < y < z_0 \Rightarrow f(y) < f(z_0). \quad (2.35)$$

But z_0 is the least upper bound of X , whence there is some $z \in X$ satisfying $y < z < z_0$. This means, for $x < z_0 - \delta < y < z < z_0$, we have $f(x) < f(y)$. Combined with (2.35) this yields $f(x) < f(z_0)$.

If z_0 is the right endpoint of I , we have shown, for all $y \in I$, that $x < y \Rightarrow f(x) < f(y)$.

If z_0 is an interior point of I , we apply the second part of Lemma 2.3.18 to obtain a contradiction. Choose δ so small that $(z_0, z_0 + \delta) \subseteq I$ and for all y ,

$$\begin{aligned} z_0 < y < z_0 + \delta &\Rightarrow f(z_0) < f(y) \\ &\Rightarrow f(x) < f(y), \text{ by the above.} \end{aligned}$$

If we choose δ small enough so that $z_0 + \delta$ is also an interior point, we see that $z_0 + \delta \in X$, contrary to the assumption that z_0 is an upper bound on X . \square

By Corollary 2.3.21, we know that f is strictly increasing on (a, b) if $f'(x) > 0$ for all $x \in (a, b)$ and, similarly, f is strictly increasing on $[a, b]$ if $f'(x) > 0$ on $[a, b]$. In the latter case, the conclusion still holds for f continuous on $[a, b]$ if we weaken the differentiability requirement to assuming $f'(x)$ exists and is positive for all $x \in (a, b)$. For, if $f(a) > f(x)$ for some $a < x \in (a, b)$, one can choose $\delta > 0$ so small that, for $y \in (a, a + \delta)$, $|f(x) - f(a)| < \frac{1}{2}(f(a) - f(x))$. But then $x < y$ and $f(y) > f(x)$. Likewise, $f(x) < f(b)$ for $x < b$.

2.3.22 Remark The proof given of Corollary 2.3.21 is surprisingly complicated and it seems to have taken some time for mathematicians to realise that it was not an immediate consequence of Lemma 2.3.18. That something has to be added may be seen by considering the simple example,

$$f(x) = \begin{cases} x, & 0 < x < 1 \\ x - 1, & 1 < x < 2. \end{cases}$$

Here $f'(x) = 1 > 0$ everywhere in the domain of f , yet $f(1/2) > f(5/4)$. The problem here, of course, is that the domain of f is not one interval, but two disjoint intervals. That said, one might acknowledge that a proof is necessary but still question the need for anything as inelegant as the proof given here, with its cases and subcases. I have chosen the current proof using the Least Upper Bound Principle as a sort of induction principle because it fits in with earlier proofs along these lines and is thus a natural choice. There are slicker proofs. In the next chapter we will encounter Weierstrass's more direct proof, which proof relies on the Extreme Value Theorem, which we proved by appeal to the Least Upper Bound Principle. Weierstrass's proof evolved into a very simple proof by appeal to the Mean Value Theorem and will be given a few pages from now (p. 139, below). An alternative reduction not as dependent on the Least Upper Bound Principle is to assume f is uniformly differentiable. We will encounter this more than once in the next chapter when we discuss the history of the Mean Value Theorem.

2.3.23 Remark With Exercise 2.3.20 we showed that it does not follow from the assumption $f'(c) > 0$ that f is increasing in any neighbourhood around c . If, however, we assume f' to be continuous in some interval around c , then Lemma 2.2.17 (p. 62, above) tells us that $f'(x) > 0$ everywhere in some interval $(c - \delta, c + \delta)$, whence Corollary 2.3.21 tells us f is strictly increasing there.

The negative derivative has its analogous result:

2.3.24 Corollary (Strictly Decreasing Function Theorem) *Let $f : I \rightarrow \mathbb{R}$ be differentiable on the interval in question and suppose $f'(x) < 0$ for all $x \in I$. Then f is strictly decreasing on I : for all $x, y \in I, x < y \Rightarrow f(y) < f(x)$.*

This can be given a proof analogous to that of Corollary 2.3.21 or it can be reduced to the application of that Corollary to $g(x) = -f(x)$. Again, if $I = [a, b]$ is closed and f is continuous on I , $f'(x)$ need only be assumed to exist and be negative for $x \in (a, b)$.

2.3.25 Exercise Paired with the Strictly Increasing Function Theorem is the (*Weakly*) *Increasing Function Theorem*: Let $f : I \rightarrow \mathbb{R}$ be differentiable in the interval in question and suppose $f'(x) \geq 0$ for all $x \in I$. Then f is increasing on I : for all $x, y \in I$, $x < y \Rightarrow f(x) \leq f(y)$.

- i. Prove this.
- ii. State and prove an analogous (*Weakly*) *Decreasing Function Theorem*.
- iii. Prove the *Constant Function Theorem*: Let $f : I \rightarrow \mathbb{R}$ be differentiable in the interval in question and suppose $f'(x) = 0$ for all $x \in I$. Then f is constant on I .

[Darboux's function shows that the analogue to Lemma 2.3.18 for $f'(x) \geq 0$ fails. Thus, instead of modifying the proof of the Strictly Increasing Function Theorem, reduce the Increasing Function Theorem to it by considering the functions $f_n(x) = f(x) + x/n$. Use a similar function for ii. The same can be done for iii, or one can reduce it directly to i and ii.]

Getting back on track, I note that Lemma 2.3.17 is the obvious lemma to use to find the point on a smooth curve C of maximum distance from a given line in the proof of the Mean Value Theorem as outlined in the Preface. To apply it we need a formula for the distance from a point to a line. Such was given first by Ludwig Otto Hesse (1811–1874):

2.3.26 Lemma *Let the line L have the equation $Ax + By + C = 0$ and let $\langle \alpha, \beta \rangle$ be any point in the plane. The distance from $\langle \alpha, \beta \rangle$ to L is given by*

$$d_L(\alpha, \beta) = \frac{|A\alpha + B\beta + C|}{\sqrt{A^2 + B^2}}. \quad (2.36)$$

Proof There are three cases to consider.

Case 1. $B = 0$. Then $A \neq 0$ as otherwise the equation is $C = 0$ which either defines the plane or the empty set, in either case not a line. The line in question is the vertical one $Ax + C = 0$, i.e., $x = -C/A$. The distance from $\langle \alpha, \beta \rangle$ to L is measured horizontally:

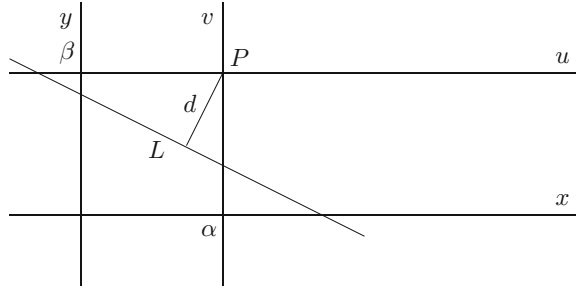
$$d_L(\alpha, \beta) = \left| x - \frac{-C}{A} \right| = \frac{|Ax + C|}{|A|} = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}},$$

since $B = 0$.

Case 2. $A = 0$. Similar.

Case 3. $A, B \neq 0$. Let L denote the line given by the equation $Ax + By + C = 0$ and let P denote the point $\langle \alpha, \beta \rangle$. A simple translation of axes will not change L and P , just their representations. Write

Fig. 2.41 Distance from P to L



$$x = u + \alpha, y = v + \beta \text{ i.e., } u = x - \alpha, v = y - \beta.$$

P acquires the uv -coordinates $\langle 0, 0 \rangle$ and the equation of L is transformed into

$$L : A(u + \alpha) + B(v + \beta) + C = 0,$$

i.e.,

$$Au + Bv + (A\alpha + B\beta + C) = 0.$$

Write D for $A\alpha + B\beta + C$.

In terms of u, v , the distance from P to L is the distance from the origin to L , which is the distance d from P to the intersection of L and the line perpendicular to L connecting L to the origin. (See Fig. 2.41.)

The equation of this perpendicular is thus $Bu - Av = 0$. The intersection point is the simultaneous solution to the pair of equations

$$\begin{aligned} Bu - Av &= 0 \\ Au + Bv &= -D, \end{aligned}$$

and simple algebra tells us this is

$$u_0 = \frac{-AD}{A^2 + B^2}, \quad v_0 = \frac{-BD}{A^2 + B^2}.$$

The distance from $\langle u_0, v_0 \rangle$ to the uv -origin P is the square root of

$$u_0^2 + v_0^2 = \frac{A^2 D^2}{(A^2 + B^2)^2} + \frac{B^2 D^2}{(A^2 + B^2)^2} = \frac{(A^2 + B^2) D^2}{(A^2 + B^2)^2} = \frac{D^2}{A^2 + B^2}.$$

Taking the square root,

$$d_L(\alpha, \beta) = d = \frac{|D|}{\sqrt{A^2 + B^2}} = \frac{|A\alpha + B\beta + C|}{\sqrt{A^2 + B^2}}.$$

□

Recall the definition

$$d_\gamma(t) = d_L(\gamma(t)) = d_L(x(t), y(t))$$

for a given line L with equation $Ax + By + C = 0$ and a parametrisation $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ of a curve, $\gamma(t) = \langle x(t), y(t) \rangle$. If each of $x(t)$ and $y(t)$ is differentiable at some point t , then d_γ^2 is also differentiable at t :

$$f(t) = d_\gamma^2(t) = \frac{(Ax(t) + By(t) + C)^2}{A^2 + B^2},$$

whence

$$f'(t) = \frac{2(Ax(t) + By(t) + C)}{A^2 + B^2} (Ax'(t) + By'(t)).$$

2.3.27 Corollary *Let $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$ be a continuous parametrisation of a curve that is not a straight line, with $\gamma(a) \neq \gamma(b)$, and suppose the component functions $x(t)$ and $y(t)$ are differentiable on (a, b) . Let L be the line*

$$(y(b) - y(a))(x - x(a)) = (x(b) - x(a))(y - y(a)) \quad (2.37)$$

connecting the points $\gamma(a)$ and $\gamma(b)$. There is a $c \in (a, b)$ such that

$$(y(b) - y(a))x'(c) = (x(b) - x(a))y'(c). \quad (2.38)$$

If we consider only the case where $x(b) \neq x(a)$, we can rewrite (2.37) in the more familiar form,

$$\frac{y - y(a)}{x - x(a)} = \frac{y(b) - y(a)}{x(b) - x(a)},$$

and note that (2.38) can be rewritten

$$\frac{y'(c)}{x'(c)} = \frac{y(b) - y(a)}{x(b) - x(a)}.$$

Thus, since

$$\frac{dy}{dx}(t) = \frac{y'(t)}{x'(t)},$$

this tells us that the slope of the tangent line to the curve passing through $\gamma(c)$ equals the slope of the secant line connecting $\gamma(a)$ and $\gamma(b)$. Or, rather, it will tell us that as soon as we have formally defined “smooth” and “tangent”, an easy enough but slightly subtle matter.

Proof of Corollary 2.3.27. By Lemma 2.2.31 there is a point $c \in (a, b)$ at which $d_\gamma(c)$ is maximum. For such c ,

$$f'(c) = \frac{2(Ax(c) + By(c) + C)}{A^2 + B^2} (Ax'(c) + By'(c)) = 0. \quad (2.39)$$

Because the curve is not a straight line, $d_\gamma(c) > 0$ and the fraction in (2.39) is not 0. Thus we have

$$Ax'(c) + By'(c) = 0, \text{ i.e., } Ax'(c) = -By'(c).$$

But if we expand (2.37) we find that

$$A = y(b) - y(a), \quad B = -(x(b) - x(a)).$$

Thus

$$(y(b) - y(a))x'(c) = (x(b) - x(a))y'(c).$$

□

Ignoring the geometric interpretation, Corollary 2.3.27 is already quite strong, encompassing the classroom versions of the Mean Value Theorem and an often poorly motivated generalisation called the Cauchy Mean Value Theorem:

2.3.28 Corollary (Classroom Mean Value Theorem)²⁰⁶ Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b) . There is a $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a},$$

or, if one prefers,

$$f(b) = f(a) + f'(c)(b - a).$$

Proof Apply Corollary 2.3.27 to $\gamma(t) = \langle t, f(t) \rangle$

□

2.3.29 Corollary (Cauchy Mean Value Theorem) Let $f, g : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b) . Assume $g(a) \neq g(b)$ and $g'(x) \neq 0$ for all $x \in (a, b)$. There is some $c \in (a, b)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Proof Apply Corollary 2.3.27 to $\gamma(t) = \langle f(t), g(t) \rangle$.

□

2.3.30 Remark The assumption $g(a) \neq g(b)$ is redundant. It follows by Corollary 2.3.28 from the assumption that $g'(x) \neq 0$ for any $x \in (a, b)$. The latter assumption

²⁰⁶The use of the word ‘‘Classroom’’ here is a local one. The reader will not find it elsewhere in the literature and I introduce it merely to distinguish the theorem as stated from the myriad of forms of the Mean Value Theorem as the one familiar from the first year Calculus course. When the distinction is unimportant, I drop the adjective.

is not redundant, as evidenced by the functions $f(x) = x^2$, $g(x) = x^3$ on the interval $[-1, 1]$. for which the only value of c satisfying (2.38),

$$(f(1) - f(-1))g'(c) = (g(1) - g(-1))f'(c),$$

i.e.,

$$(1 - 1)3c^2 = (1 - (-1))2c,$$

i.e., $0 = 4c$, is $c = 0$. But $g'(0) = 0$, whence the division yielding

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}$$

cannot be performed. (Exercise: Examine the same pair of functions on $[-1, 2]$.)

Corollary 2.3.28 permits simpler proofs of Corollaries 2.3.21 and 2.3.24.

Simple proof of Corollary 2.3.21. Assume $f'(x) > 0$ everywhere in the interior of I and let $x, y \in I$ with $x < y$. By Corollary 2.3.28 there is some c with $x < c < y$ such that

$$f'(c) = \frac{f(y) - f(x)}{y - x}.$$

Thus $f(y) - f(x) = f'(c)(y - x) > 0$ since $y > x$ and $f'(c) > 0$. Thus $f(y) > f(x)$. \square

The corresponding proof of Corollary 2.3.24 is similar, as are proofs of the weaker variants given in Exercise 2.3.25. Of these I only single out part iii for demonstration here:

2.3.31 Corollary (Constant Function Theorem) *Let $f : I \rightarrow \mathbb{R}$ be differentiable on the interval in question and suppose $f'(x) = 0$ for all $x \in I$. Then: f is constant on I .*

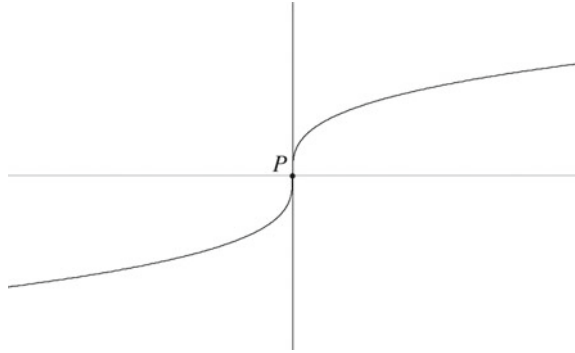
Proof Fix $a \in I$ and let $x \neq a$ be any other element of I . By Corollary 2.3.28 there is an element c between a and x such that

$$f(x) - f(a) = f'(c)(x - a) = 0 \cdot (x - a) = 0,$$

whence $f(x) = f(a)$. \square

The classroom version of the Mean Value Theorem is the ultimate interest in the present work and we will begin its study in earnest in the next chapter. Up till now, however, we have been concerned with the general geometric form as introduced in the Preface, and we still have a few loose ends to clear up. These concern the meaning to be assigned to the words “smooth” and “tangent”.

In the traditional Calculus course, one mainly treats curves of the form $y = f(x)$, i.e., the graphs of functions. Here the definition of the tangent to the curve at a point is fairly simple:

Fig. 2.42 An inverse cubic

2.3.32 Definition Let the function $y = f(x)$ be defined on the interval $[a, b]$ and let $c \in [a, b]$ and assume f is differentiable at c . The *tangent to the curve* $y = f(x)$ at $\langle c, f(c) \rangle$ (or, simply, at c) is the line with equation

$$y = f(c) + f'(c)(x - c). \quad (2.40)$$

This definition is not entirely satisfactory. Consider the inverse cubic function of Fig. 2.42. According to Definition 2.3.32 it will have no tangent at P because, at $c = 0$, $f'(c)$ is infinite. Yet as a curve, it is not much different from the graph of the cubic (see Fig. 2.30 on page 80, above), which has tangents everywhere. The present curve is the reflexion of a cubic across the line $y = x$ and the reflexion of the tangents to the cubic ought to be considered as tangents of the cube root. To do so, we need to allow infinite slopes, i.e., vertical lines, as tangents.

In doing this, some care is necessary. The inverse cubic is not much of a problem. For $y = x^{1/3}$, we have

$$f'(0) = \lim_{h \rightarrow 0} \frac{h^{1/3} - 0^{1/3}}{h} = \lim_{h \rightarrow 0} \frac{1}{h^{2/3}} = +\infty.$$

For a curve like the cycloid, which might also be considered to have a vertical tangent, we have, for example, distinct left- and right-sided limits at the cusps:

$$\begin{aligned} \lim_{h \rightarrow 0+} \frac{f(h) - f(0)}{h} &= \lim_{t \rightarrow 0+} \frac{(1 - \cos t) - 0}{t - \sin t} = +\infty \\ \lim_{h \rightarrow 0-} \frac{f(h) - f(0)}{h} &= \lim_{t \rightarrow 0-} \frac{(1 - \cos t) - 0}{t - \sin t} = -\infty. \end{aligned}$$

Here, by $\lim_{h \rightarrow 0+}$ we mean to imply that the variable h is restricted to lying to the right of 0 in the definition:

$$\lim_{h \rightarrow 0+} g(h) = L$$

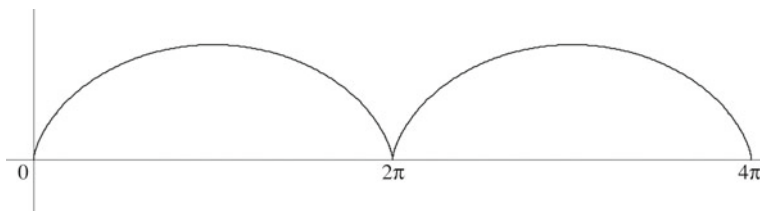


Fig. 2.43 Only one non-tangent

means that for any $\epsilon > 0$ there is some $\delta > 0$ such that

$$0 < h < \delta \Rightarrow |g(h) - L| < \epsilon.$$

Similarly, $\lim_{h \rightarrow 0^-}$ requires the conclusion of the implication to hold only for $-\delta < h < 0$.

For the inverse cubic, we can say that the derivative exists and is $+\infty$, while for the cycloid we would have to say the derivative does not exist because a single two-sided limit does not exist.

2.3.33 Remark Actually, the situation is even more subtly complicated than this. Definition 2.3.32 allowed one to consider the derivative at the endpoint of an interval, which derivative, of course, depends on the limit of the difference quotient at the endpoint. This will necessarily be a one-sided limit, as our definition of limit required $f(x+h)$ to approach the candidate for the limit only for $x+h$ in the interval under consideration. Thus, if we define the curve C by,

$$\begin{aligned} x(t) &= t - \sin t \\ y(t) &= 1 - \cos t, \quad t \in [0, 4\pi], \end{aligned}$$

as in Fig. 2.43, we have $f'(0) = +\infty$, $f'(4\pi) = -\infty$, but $f'(2\pi)$ is undefined. This truncated cycloid does not have cusps at 0 and 4π , whence we can take the vertical lines there to be tangents (using the equations $x = 0$ and $x = 4\pi$ in place of (2.40)). We could extend the cycloid beyond the endpoints, thus restoring the cusps and losing the tangents, but we could also extend the curve differently, for example by adding a copy of C reflected across the x -axis, yielding “true” tangents at 0 and 4π , as in Fig. 2.44. Whether or not there is a tangent at 2π in this new figure will depend on how we define “tangent” for parametrically defined curves and how we parametrise the curve.

So, how do we define the notion of tangent for a parametrically definable curve? The classroom definition for the graph of a function $y = f(x)$ is unambiguous. We may express f in numerous ways, but when it comes to numerical values they all agree. For example, for $f(x) = \sin^2 x$, it makes no difference if we represent f by the expression $\sin^2 x$ or $1 - \cos^2 x$: Not only are $y = \sin^2 x$ and $y = 1 - \cos^2 x$ the same curve, but each value of x in the domain of f determines the same point

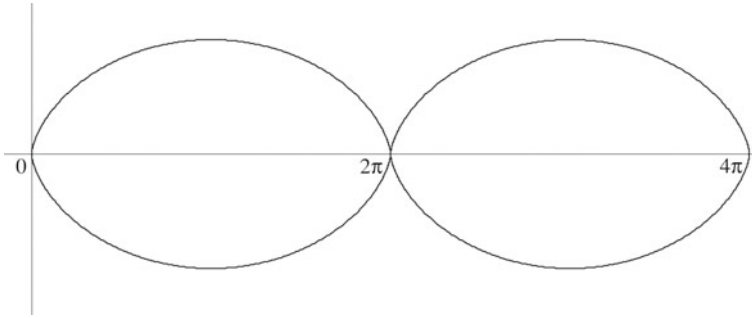


Fig. 2.44 Is there a tangent at 2π ?

$\langle x, y \rangle$ on the graph. General curves have many different parametrisations that behave differently; the graph of a function is, essentially, a single canonical parametrisation of a curve:

$$\begin{aligned} x(t) &= t \\ y(t) &= f(t), \quad t \in \text{domain}(f). \end{aligned}$$

2.3.34 Exercise We can give two essentially different parametrisations of the curve of Fig. 2.44 as follows. First, let

$$f(t) = t - \sin t, \quad g(t) = 1 - \cos t$$

be the functions used in the usual parametrisation of the cycloid, and define γ_1, γ_2 as follows.

$$\begin{aligned} \gamma_1 : \quad x_1(t) &= \begin{cases} f(t), & 0 \leq t \leq 4\pi \\ 4\pi - f(t - 4\pi), & 4\pi < t \leq 8\pi, \end{cases} \\ y_1(t) &= \begin{cases} g(t), & 0 \leq t \leq 4\pi \\ -g(t), & 4\pi < t \leq 8\pi. \end{cases} \\ \gamma_2 : \quad x_2(t) &= \begin{cases} f(t), & 0 \leq t \leq 4\pi \\ 4\pi - f(t - 4\pi), & 4\pi < t \leq 8\pi, \end{cases} \\ y_2(t) &= \begin{cases} g(t), & 0 \leq t \leq 2\pi \text{ or } 4\pi < t \leq 6\pi \\ -g(t), & 2\pi < t \leq 4\pi \text{ or } 6\pi < t \leq 8\pi. \end{cases} \end{aligned}$$

Graph these on your graphing calculator, watching the process slowly unfold. Viewing the function as defining the trajectory of a moving particle, which would you consider as representing a smooth motion? Which has two “bounces”, i.e., cusps? What about the curve they trace out?

The point here is that functions are simpler to deal with than curves. We can easily unambiguously define differentiability for a function:

2.3.35 Definition A function $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ given by $\gamma(t) = \langle x(t), y(t) \rangle$ is *differentiable at a point* $c \in I$ if each of $x(t)$ and $y(t)$ is differentiable at c . We write

$$\gamma'(c) = \langle x'(c), y'(c) \rangle.$$

2.3.36 Exercise Which of the functions γ_1, γ_2 of Exercise 2.3.34 is differentiable at $t = 2\pi$? Find $\gamma'(2\pi)$ and $\gamma'(6\pi)$ for this function.

Defining the smoothness of a function is a matter of some delicacy, for there is not a unique notion of smoothness used in Analysis, nor two notions, nor three, ..., but an infinite number of levels of smoothness depending on how many times the function is differentiable and whether or not the last derivative is continuous. For our purposes we don't require much smoothness at all, but, as we want to conclude the existence of tangent lines from smoothness, we add an extra condition:

2.3.37 Definition A function $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ is *smooth* if for all $a, b \in I$ with $a < b$

- i. γ is continuous on $[a, b]$,
- ii. γ is differentiable on (a, b) , and
- iii. $\gamma'(c) \neq \langle 0, 0 \rangle$ for any $c \in (a, b)$.

With respect to Exercise 2.3.34, I note that the function γ_2 is smooth, but γ_1 is not.

2.3.38 Definition A curve C is *smooth* if there is an interval I and a smooth function $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ such that $C = \gamma(I) = \{\gamma(t) \mid t \in I\}$.

According to this definition, the curve of Fig. 2.44 is smooth because of the smooth parametrisation γ_2 of Exercise 2.3.34. And the cube root, $y = x^{1/3}$ is just as smooth as the cube $y = x^3$, as one can see by comparing their respective parametrisations,

$$\gamma_1(t) = \langle t^3, t \rangle \quad \gamma_2(t) = \langle t, t^3 \rangle,$$

with derivatives

$$\gamma_1'(t) = \langle 3t^2, 1 \rangle \quad \gamma_2'(t) = \langle 1, 3t^2 \rangle.$$

Conditions (i) and (ii) of the definition of smoothness are natural enough. Condition (iii) is explained by the following definition of a tangent line:

2.3.39 Definition Let $\gamma : [a, b] \rightarrow \mathbb{R} \times \mathbb{R}$ be differentiable at $c \in (a, b)$ with $\gamma'(c) \neq \langle 0, 0 \rangle$. The *line tangent to γ at c* (or: *at $\gamma(c)$*) is the line L given, according to case, by the equation,

$$\begin{aligned} x &= x(c), \text{ if } x'(c) = 0, \text{ or} \\ y &= y(c), \text{ if } y'(c) = 0, \text{ or} \\ y &= y(c) + \frac{y'(c)}{x'(c)}(x - x(c)), \text{ if } x'(c) \neq 0 \text{ and } y'(c) \neq 0. \end{aligned}$$

Were it not for the common preference for writing the equation of a line in slope-intercept form, we could consolidate these cases in the single line,

$$(y - y(c))x'(c) = y'(c)(x - x(c)),$$

or

$$y'(c)x - x'(c)y + y(c)x'(c) - x(c)y'(c) = 0.$$

And we see immediately why $\gamma'(c) = \langle 0, 0 \rangle$ is unwelcome: The resulting equation is $0x + 0y + 0 = 0$, which defines the plane and not a line. (Or, using the equations of the Definition, one has two lines $x = x(c)$ and $y = y(c)$.)

A $\langle 0, 0 \rangle$ derivative at some point on a differentiable curve may accompany a geometric tangent or it may not.

2.3.40 Exercise Define γ_1, γ_2 on $[-1, 1]$ by

$$\begin{aligned}\gamma_1(t) &= \langle t^6, t^3 \rangle \\ \gamma_2(t) &= \begin{cases} \gamma_1(t), & t \leq 0 \\ -\gamma_1(t), & 0 < t. \end{cases}\end{aligned}$$

- i. Show that $\gamma_1(t)$ is continuously differentiable on $[-1, 1]$ and parametrises the parabola $x = y^2$ over the y -interval $[-1, 1]$ and the curve has a tangent at $t = 0$ although $\gamma_1' = \langle 0, 0 \rangle$.
- ii. Show that $\gamma_2(t)$ is also continuously differentiable on $[-1, 1]$ with $\gamma_2'(0) = \langle 0, 0 \rangle$, but γ_2 has a cusp at $t = 0$.

Referring once again to the functions of Exercise 2.3.34, we see that the vertical line $x = 2\pi$ is tangent to γ_2 at $t = 2\pi$ and $t = 6\pi$, while γ_1 has no tangent at these points. As for their common curve $C = \gamma_1([0, 8\pi]) = \gamma_2([0, 8\pi])$, we would say that a line is tangent to C at a point P if it is tangent there with respect to some smooth parametrisation:

2.3.41 Definition A line L is *tangent to a curve* C at a point $P \in C$ if there is some smooth parametrisation $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ of C and some point $c \in I$ such that $\gamma(c) = P$ and L is tangent to γ at c .

With all of this we can now restate Corollary 2.3.27 in the form of the Mean Value Theorem as presented to the man-in-the-street back in the Preface.

2.3.42 Theorem (Mean Value Theorem; Geometric Form) *Let C be a smooth curve with distinct points P, Q on the curve. There is a point R on the curve at which the tangent line is parallel to the segment PQ .*

You will notice that the statement of the Theorem omits the mention that R lies “between” P and Q . A curve, as we have defined it, is just a set of points. It has no orientation of its own, hence no notion of betweenness. The orientation is determined

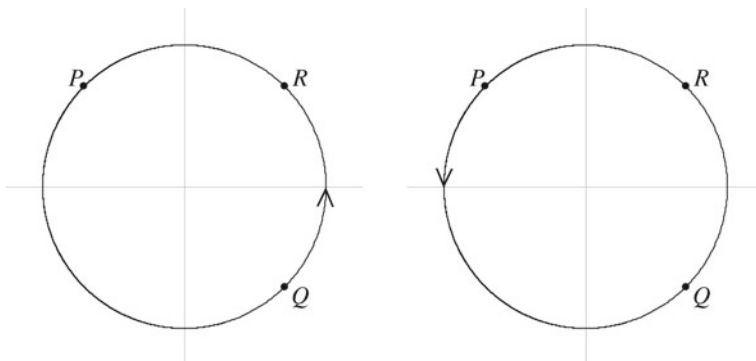


Fig. 2.45 Is R between P and Q ?

not by the curve, but by the particular parametrisation chosen to demonstrate the smoothness of the curve. For example, we can parametrise the circle by

$$\gamma_1(t) = \langle \cos t, \sin t \rangle, \quad t \in [0, 2\pi]$$

or, say, by

$$\gamma_2(t) = \langle \cos t, \sin t \rangle, \quad t \in [\pi, 3\pi].$$

Consider the points

$$P = \langle -\sqrt{2}/2, \sqrt{2}/2 \rangle, \quad Q = \langle \sqrt{2}/2, -\sqrt{2}/2 \rangle, \quad R = \langle \sqrt{2}/2, \sqrt{2}/2 \rangle,$$

as in Fig. 2.45. In the first circle, $R = \gamma_1(\pi/4)$ does not lie between $P = \gamma_1(3\pi/4)$ and $Q = \gamma_1(7\pi/4)$ because, as one traces the curve going from $t = 0$ to $t = 2\pi$, R appears before both P and Q . In the second circle $R = \gamma_2(9\pi/4)$ occurs after $Q = \gamma_2(7\pi/4)$, but before $P = \gamma_2(11\pi/4)$, as the curve is traced, whence R lies between P and Q .

The matter appears to be one of unnecessary subtlety and is probably best ignored, i.e., resolved by introducing the parametrisation explicitly into the statement of the Theorem.

2.3.43 Theorem (Mean Value Theorem; Algebraic-Geometric Form) *Let $\gamma : I \rightarrow \mathbb{R} \times \mathbb{R}$ parametrise a smooth curve C and let $a, b \in I$ with $a < b$ and $\gamma(a) \neq \gamma(b)$. There is some $c \in (a, b)$ such that the line tangent to γ at c is parallel to the secant connecting $\gamma(a)$ and $\gamma(b)$.*

Basically, the only difference between this statement and that of Corollary 2.3.27 is that the Theorem does not exclude the trivial case in which that portion of the curve between $\gamma(a)$ and $\gamma(b)$ coincides with the secant line.

I leave the details of the proof to the reader.

We are almost finished with this chapter. What we haven't done to (at least: my) complete satisfaction is to justify the choices of definitions of a smooth curve and a tangent to the curve. This would be done by showing that the definitions agree with the intuitively defined tangents in all classical cases and do not generalise too much, defining tangents where they shouldn't be or accepting curves that are too un-curvelike. Now, this is a matter of some depth and subtlety and, even though we have been routinely straying somewhat from our ostensible purpose of discussing the Classroom Mean Value Theorem, I should at some point get back on track. Thus, I shall, for the most part, leave the reader to convince him- or herself that these definitions are, if not written in stone, at least reasonable. I do feel compelled, however, to address the issue of space-filling curves.

2.3.44 Exercise Let $\gamma(t) = \langle \cos t^4, \sin t^2 \rangle$ on $[.0001, 45]$. Drag out your TI-83 or TI-84, put it into parametric graphing mode, enter

$$\begin{aligned} X_{1T} &= \cos(T^4) \\ Y_{1T} &= \sin(T^2), \end{aligned}$$

and set the window to

$$\begin{aligned} T_{\min} &= .0001 \\ T_{\max} &= 45 \\ T_{\text{step}} &= .05 \\ X_{\min} &= -2.35 \\ X_{\max} &= 2.35 \\ Y_{\min} &= -1.55 \\ Y_{\max} &= 1.55, \end{aligned}$$

and graph the function. What do you see? Is this a smooth function?

It should not be revealing too much to say that what one sees on one's calculator tells one more about the resolution of the calculator's display than about the nature of the graph of γ . Graphing the function on the computer at higher resolution reveals a lot of white space. If one extends γ to larger and larger intervals, the space available for the image of γ fills in more and more, and the resolution of the graph may again be overtaken by the function. Nonetheless, there are points that γ will miss. For example, for no value of $t \in (-\infty, \infty)$ does $\gamma(t) = \langle 1, 1 \rangle$. To see this, assume by way of contradiction that

$$\gamma(t) = \langle \cos t^4, \sin t^2 \rangle = \langle 1, 1 \rangle.$$

Now,

$$\begin{aligned} \cos t^4 = 1 &\Rightarrow \text{for some natural number } m, \quad t^4 = 2m\pi \\ \sin t^2 = 1 &\Rightarrow \text{for some natural number } n, \quad t^2 = \frac{4n+1}{2}\pi. \end{aligned}$$

Thus,

$$2m\pi = \left(\frac{4n+1}{2}\pi\right)^2,$$

i.e.,

$$8m\pi = (4n+1)^2\pi^2,$$

and, since $\pi \neq 0$,

$$\pi = \frac{8m}{(4n+1)^2},$$

contrary to the irrationality of π .

2.3.45 Exercise For γ as in Exercise 2.3.44,

- i. show that the points $\langle 0, 0 \rangle$, $\langle 1, -1 \rangle$, $\langle -1, 1 \rangle$, and $\langle -1, -1 \rangle$ do not lie on the curve $\gamma([.0001, 45])$; and
- ii. show that $\gamma'(t) \neq \langle 0, 0 \rangle$ for $t \in [.0001, 45]$. Conclude that γ is smooth.

One can do much better. In general, a differentiable curve misses “most” points.

As the reader may remember, at the end of his paper Peano cited some conditions which, when assumed in addition to continuity, prevented a curve $\gamma : [0, 1] \rightarrow [0, 1] \times [0, 1]$ from being a space-filling curve. One was that the curve be the graph of a function $y = f(x)$ and the other was that the curve be of *bounded variation*. In both cases, the reason the curve could not fill the entire unit square was that it could be fit inside a set of arbitrarily small area. I have not seen Jordan’s proof that a curve of bounded variation cannot be a space-filling curve, but can report that Robert Burckel and Caspar Goffman have published a fairly simple combinatorial proof of this result.²⁰⁷ I will not prove the result in this generality, but will present instead a simpler proof that no continuously differentiable curve is a space-filling curve.

2.3.46 Theorem *Let $\gamma : [0, 1] \rightarrow [0, 1] \times [0, 1]$ be continuously differentiable on $[0, 1]$ and let $\epsilon > 0$. There is a set $X \subseteq [0, 1] \times [0, 1]$ of area $< \epsilon$ such that $\gamma([0, 1]) \subseteq X$. In other words, the curve $C = \gamma([0, 1])$ has zero area and thus cannot equal the entire square, which has area 1.*

Proof Let $\epsilon > 0$. Writing $\gamma(t) = \langle x(t), y(t) \rangle$, we are assuming $x'(t)$ and $y'(t)$ continuous, hence bounded on $[0, 1]$. Let $B > 0$ be a common bound on $|x'(t)|$, $|y'(t)|$ for $t \in [0, 1]$. For $s, t \in [0, 1]$, the Mean Value Theorem (Corollary 2.3.28) yields

$$x(s) - x(t) = x'(t_0)(s - t), \quad y(s) - y(t) = y'(t_1)(s - t),$$

for some $t_0, t_1 \in (0, 1)$, whence

$$|x(s) - x(t)|, |y(s) - y(t)| \leq B \cdot |s - t| < \frac{B}{n}, \text{ for } |s - t| < \frac{1}{n+1}, \quad (2.41)$$

²⁰⁷R.B. Burckel and C. Goffman, “Rectifiable curves are of zero content”, *Mathematics Magazine* 44 (1971), pp. 179–180.

where $n > 1$ will be chosen shortly.

Observe

$$\begin{aligned}\gamma([0, 1]) &= \bigcup_{k=0}^{n-1} \gamma\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right) \\ &\subseteq \bigcup_{k=0}^{n-1} x\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right) \times y\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right),\end{aligned}$$

and each rectangle

$$x\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right) \times y\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right)$$

has area less than

$$\frac{B}{n} \cdot \frac{B}{n} = \frac{B^2}{n^2} = \frac{B^2}{n} \cdot \frac{1}{n}.$$

Now choose $n > B^2/\epsilon$, so that $\epsilon > B^2/n$. Then $\gamma([0, 1])$ is contained in a set of area less than

$$\sum_{k=0}^{n-1} \frac{B^2}{n} \cdot \frac{1}{n} = n \cdot \frac{B^2}{n} \cdot \frac{1}{n} = \frac{B^2}{n} < \epsilon.$$

□

Our definition of smoothness did not require continuous differentiability, and not all smooth curves are continuously differentiable, whence the above proof does not apply to them generally. It can be shown that a differentiable curve cannot be a space-filling curve, but I don't know any proof of comparable simplicity in the general case.²⁰⁸ Thus I shall simply allow Theorem 2.3.46 to stand as an indication that some sort of differentiability is the appropriate condition to add to continuity to formally capture the intuitive notion of a smooth curve. The emergence of sharp formal concepts from vague intuitive ones is an interesting study, and we have been following it throughout this chapter, but it is time now to change gears. The concepts have been formalised and we wish to consider the Mean Value Theorem itself.

²⁰⁸Burckel and Goffman prove Theorem 2.3.46 for rectifiable curves. A curve is rectifiable just in case it is of bounded variation. Using (2.41) one easily shows continuously differentiable curves to be rectifiable, whence Theorem 2.3.46 is a special case of Jordan's result. Not every differentiable function, however, is rectifiable. Gelbaum and Olmsted, *op. cit.*, pp. 140–141, cite $x^2 \sin(1/x^2)$ as an example.



<http://www.springer.com/978-3-319-52955-4>

MVT: A Most Valuable Theorem

Smorynski, C.

2017, X, 499 p. 83 illus., Softcover

ISBN: 978-3-319-52955-4