Where Are Limits Needed in Calculus?

R. Michael Range

Abstract. A method introduced in the 17th century by Descartes and van Schooten for finding tangents to classical curves is combined with the point-slope form of a line in order to develop the differential calculus of all functions considered in the 17th and 18th centuries based on simple purely algebraic techniques. This elementary approach avoids infinitesimals, differentials, and similar vague concepts, and it does not require any limits. Furthermore, it is shown how this method naturally generalizes to the modern definition of differentiability.

1. INTRODUCTION. Have you ever wondered why we burden our students with limits when teaching about tangent lines and differentiation of rational, root, and similar algebraic functions? Of course, as experienced mathematicians, we know that limits ultimately cannot be avoided. However, the early emphasis on limits in the context of differentiation of numerous algebraic examples may cause quite a bit of confusion. After all, the calculation of such derivatives relies primarily on algebraic techniques to rewrite the difference quotient in such a way that one can cancel the troublesome $h = \Delta x \neq 0$ from the denominator. The final answer then follows by what appears to be just *plugging in* $h = 0$. Surely most students, when first shown the derivative of $y = x^2$, are hard pressed to understand the subtlety of the statement $\lim_{h\to 0}(2x + h) = 2x$, a conceptual leap that took mathematicians close to two centuries to fully understand and to formulate correctly. Yes, we try to teach our students that we do need to take the *limit* as $h \to 0$, rather than just *evaluate* at $h = 0$. On the other hand, evaluating at $h = 0$ is eventually justified by invoking the *continuity* of the relevant functions. No wonder today's students in a standard first calculus course typically retain little about limits. They have grown up with graphing calculators, and continuity—at least its intuitive geometric interpretation—looks obvious to them. Consequently it is difficult for them to grasp the need for limits as long as one considers only algebraic functions.

Thinking about these difficulties in the teaching of elementary calculus, I was somewhat surprised to discover that there is a very simple and natural algebraic approach to differentiation of algebraic functions that avoids limits altogether and justifies the students' "easy" calculation of derivatives by "plugging in." More surprising was the realization that this approach had not been used systematically in the early days of calculus, when mathematicians struggled unsuccessfully for nearly two centuries to resolve the inconsistencies and mysteries regarding infinitely small quantities, infinitesimals, and differentials that are zero or nonzero depending on what suits the purpose.[1] While the basic idea already appears in the work of René Descartes (1596–1650) and Frans van Schooten (1615–1660), their method remained on the sidelines because—in the words of Howard Eves—"Here we have a general process which tells us exactly what to do to solve our problem, but it must be confessed that in the more complicated cases the required algebra may be quite forbidding" [**6**, pp.

doi:10.4169/amer.math.monthly.118.05.404

[1]Nearly 100 years after the beginnings of calculus, Leonhard Euler (1707–1783) still had not yet found a convincing explanation. For example, he wrote in his 1755 *Institutiones Calculi Differentialis,* the definitive calculus textbook of that time: "There is no doubt that any quantity can be diminished until it all but vanishes and then goes to nothing. But an *infinitely small quantity* is nothing but a vanishing quantity, and so it *is really equal to 0*" (emphasis added) [**5**].

284–285]. It turns out that the algebraic complications encountered in the 17th century vanish if one introduces the *point-slope form* of lines rather than using a second distant point, such as the point of intersection of the tangent with a coordinate axis, as was typically done in those days. This basic form—familiar to every high school student today—was apparently not known explicitly until the late 18th century, when Gaspard Monge (1746–1818) introduced it in a paper published in 1784 (see [**2**, pp. 205–206]). All this raises some interesting historical questions that I will discuss in a separate article. Here I just point out that this simple approach to derivatives is not limited to algebraic functions, but applies just as easily to formal power series, that is, to the most general type of function accepted in the 17th and 18th centuries. Consequently, all the *differential* calculus of the 17th century could have been done without infinitesimals and differentials, and without any of the inconsistencies that came with them. Stated differently, the real need for limits begins with integrals as limits of approximating sums, or—from today's perspective, which requires careful attention to the *convergence* of power series—with the differential calculus of the elementary transcendental functions.

In more recent times the crux of Descartes and van Schooten's approach to tangents has resurfaced in algebraic geometry, where the tangent at a regular point of an algebraic curve is defined to be that line which intersects the curve with multiplicity greater than 1. (See [**13**, p. 73], for example.) However, the focus there is different, and I do not know any text that develops derivatives in a systematic and elementary way from that perspective.

In this article I will describe this algebraic process to define derivatives and use it to verify the standard rules of differentiation without using any limits. Furthermore, I will briefly discuss how this approach naturally leads to the modern definition of differentiability, in an elegant formulation that apparently was introduced by Constantin Carathéodory (1873–1950) in the first half of the 20th century, and which unfortunately is not widely known. Perhaps some readers will be sufficiently intrigued to consider taking a fresh look at how we think about and teach calculus.

2. DERIVATIVES. The basic problem in (differential) calculus is to define the instantaneous velocity of a particle at a particular moment, given its position $s = f(t)$ as a function of time t, and to find a practical method to calculate it. The equivalent geometric version involves the tangent to the graph of a function at a given point. Students are quite familiar with this latter version in the context of tangents to circles. In fact, this classical example clearly illustrates the critical difference between a *secant*, which intersects the circle at two distinct points, and a *tangent*, which intersects the circle at just one point. Of course, the crux is the fact that in the latter case the one point of intersection is a "double point." Multiplicities matter! This is evident when a tangent is slightly perturbed: suddenly the point of tangency splits into two points.

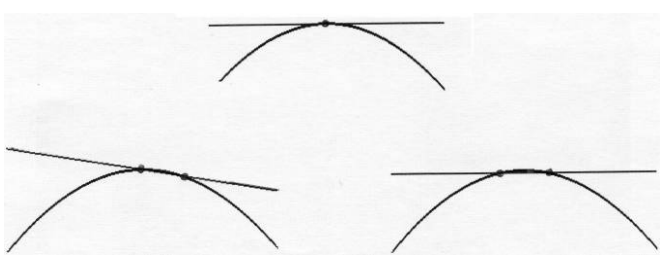

Figure 1. Perturbation of a tangent reveals *two* points of intersection.

Algebraically, finding the points of intersection of a line with a circle leads to a quadratic equation which has either no (real) solution, one single solution of multiplicity two (a double zero), or two distinct solutions. Elementary algebra easily allows us to distinguish these cases. That's all there is to it. No limits, no infinitesimals, no contradictions with first assuming $h \neq 0$ and then plugging in $h = 0$ after all.

The rest involves applying this simple purely algebraic process systematically. As I will show, this works quite easily for all algebraic functions. To illustrate the simplicity of the method, let us consider first the tangent to $f(x) = x^2$ at the point (a, a^2). The equation of a generic line through that point is given by $y = a^2 + m(x - a)$, where m is the slope. Its points of intersection with the graph of f are found by solving the equation

$$x^2 - a^2 - m(x - a) = 0,$$

which factors to

$$[(x + a) - m](x - a) = 0.$$

The solutions a and $m - a$ coincide, i.e., we have a double point of intersection, precisely when $m = 2a$.

Next we apply this procedure to the case of a polynomial $P(x)$ of degree $r \geq 2$. The equation of a generic line through the point $(a, P(a))$ is $y = P(a) + m(x - a)$. The points of intersection of this line with the graph of P are found by solving the equation

$$P(x) - P(a) - m(x - a) = 0.$$

We need to find the slope m so that this equation has a double (or higher) zero at $x = a$. As we shall see, this condition determines the number m uniquely. In order to find the desired m, we proceed as follows. Since $x = a$ is a zero of $P(x) - P(a)$, by standard algebra one can factor

$$P(x) - P(a) = q(x)(x - a),$$

where q is a polynomial of degree $r - 1$ which can be readily computed by the polynomial division algorithm. Now $q(x) - q(a)$ clearly has a zero at a as well, so, by the same argument, $q(x) - q(a) = k(x)(x - a)$ for some polynomial k of degree $r - 2$. Then

$$\begin{aligned} P(x) - P(a) - m(x - a) &= (q(x) - m)(x - a) \\ &= (q(a) - m)(x - a) + [q(x) - q(a)](x - a) \\ &= (q(a) - m)(x - a) + k(x)(x - a)^2. \end{aligned}$$

This representation shows that $P(x) - P(a) - m(x - a) = 0$ has a zero of multiplicity greater than one at $x = a$ if and only if $m = q(a)$, i.e., the number $q(a)$ is the desired slope of the tangent line to the graph of P at $(a, P(a))$. In particular, one sees that there exists a well defined tangent at every point of the graph of a polynomial.

We thus are justified in making the following definition.

Definition. The tangent line to the graph of a polynomial $P(x)$ at the point $(a, P(a))$ is the (unique) line through $(a, P(a))$ which intersects the graph at that point with

multiplicity at least 2. The slope of the tangent line is given by $q(a)$, where q is the polynomial factor in the representation $P(x) - P(a) = q(x)(x - a)$. The slope of the tangent is called the derivative of P at a, and it is denoted by $P'(a)$.

This definition also applies if the graph of P is a line L, i.e., if P has degree ≤ 1. Note that in this case another line can intersect L with multiplicity greater than one only if the two lines coincide.

Example. For a positive integer n the derivative of $f(x) = x^n$ is obtained as follows. Fix a and factor $x^n - a^n = q(x)(x - a)$, where

$$q(x) = \sum_{j=0}^{n-1} x^{n-1-j} a^j.$$

Then $f'(a) = q(a) = na^{n-1}$. A comparison with the standard calculation of the derivative via limits reveals that the required algebra hasn't changed at all. In fact, in order to evaluate

$$\lim_{x \to a} \frac{x^n - a^n}{x - a},$$

one factors the numerator and cancels the *nonzero* factor $(x - a)$, thereby obtaining

$$\frac{x^n - a^n}{x - a} = \frac{q(x)(x - a)}{x - a} = q(x) \text{ for } x \neq a.$$

Note how avoiding quotients and shifting to double points completely removes the difficulty with relating $q(x)$ for $x \neq a$ to $q(a)$.

It is clear that for every polynomial P its derivative defines a new function $y = P'(x)$. By standard rules (see below), P' is a polynomial as well.

Remark. The definition of derivative and the procedure to calculate it readily generalize to rational functions $R(x) = P(x)/Q(x)$ at any point a where R is defined, i.e., where $Q(a) \neq 0$. Note that if $R(a) = 0$, then $P(a) = 0$ as well, and hence $P(x) = q_P(x)(x - a)$ for some polynomial $q_P(x)$. Consequently $R(x) = q(x)(x - a)$, where $q(x) = q_P(x)/Q(x)$ is a rational function defined at a. If $R(a) \neq 0$, one obtains a corresponding factorization

$$R(x) - R(a) = q(x)(x - a)$$

with another rational function q defined at a. In analogy to polynomials we say that a rational function $S(x)$ has a zero at $x = a$ of multiplicity $\geq n$, where n is a positive integer, if $S(x) = k(x)(x - a)^n$ for some rational function $k(x)$ defined at a.[2] As in the case of polynomials, it follows that a rational function R is differentiable at every point a where it is defined, i.e., its graph has a (unique) tangent line at the point $(a, R(a))$ defined by the property that it intersects the graph of R at $(a, R(a))$ with multiplicity at least two. The slope of the tangent (i.e., the derivative) equals the value $q(a)$, just as in case of polynomials.

[2] The analogous notion will later be applied to more general algebraic functions S, the critical feature being that the factor k will be a function given by an algebraic expression defined on some open interval containing the point a.

3. BASIC RULES. The standard rules for differentiation follow readily. For example, to prove the product rule, write $f_j(x) = f_j(a) + q_j(x)(x-a)$, $j = 1, 2$, and note that

$$f_1(x)f_2(x) = f_1(a)f_2(a) + f_2(a)q_1(x)(x-a) + f_1(a)q_2(x)(x-a) + q_1(x)q_2(x)(x-a)^2.$$

It follows that

$$(f_1f_2)(x) - (f_1f_2)(a) = [f_2(a)q_1(x) + f_1(a)q_2(x) + q_1(x)q_2(x)(x-a)](x-a) = q(x)(x-a),$$

so that $(f_1f_2)'(a) = q(a) = f_2(a)q_1(a) + f_1(a)q_2(a) = f_2(a)f_1'(a) + f_1(a)f_2'(a)$.

A similar argument proves the quotient rule. These rules imply that the derivative of any polynomial (rational) function is again a polynomial (rational) function.

The chain rule is even easier. Suppose f and g are two (rational) functions, with g defined at a and f defined at $b = g(a)$, so that the composition $f \circ g$ is defined at a. If $f(y) - f(b) = q_f(y)(y-b)$ and $g(x) - g(a) = q_g(x)(x-a)$, where q_f and q_g are rational, substitute $y = g(x)$ and $b = g(a)$ to obtain

$$f(g(x)) - f(g(a)) = q_f(g(x))(g(x) - g(a)) = q_f(g(x))q_g(x)(x-a).$$

Since $q_f(g(x))q_g(x)$ is rational, one obtains

$$(f \circ g)'(a) = q_f(g(a))q_g(a) = f'(g(a))g'(a).$$

Note that it is not necessary to prove the differentiability of the composition by a separate argument, since it is known that the composition of rational functions is again rational, and hence—as observed earlier—it is differentiable where defined.

Inverse functions require just a little bit more care, but no really new idea is needed. First of all, zeroes of polynomials are isolated, and therefore, if $R(x) = P(x)/Q(x)$ is defined at a and $R(a) \neq 0$ (i.e., $P(a) \neq 0$ and $Q(a) \neq 0$), there exists an interval I containing a such that $R(x) \neq 0$ for $x \in I$; furthermore, $R(x)$ does not change sign on I. We shall use the following lemma.

Lemma. *Suppose R is a rational function defined at a and $R'(a) \neq 0$. Then there exists an open interval I containing a such that R is one-to-one on I, and hence $R|_I$ has an inverse function S defined on the set $J = R(I)$.*

I omit the algebraic proof of this fact, which can be handled without using the mean value theorem.

Let R be a rational function defined at a, and suppose $R'(a) \neq 0$. Choose the interval I such that the inverse S of R is defined on the set $J = R(I)$. Consider the factorization

$$R(x) - R(a) = q(x)(x-a).$$

Since $q(a) = R'(a) \neq 0$, by shrinking I we can also assume that $q(x) \neq 0$ for all $x \in I$. For $x \in I$ substitute $y = R(x)$, $x = S(y)$ and $b = R(a)$, $a = S(b)$ in the factoriza-

tion above, and rearrange to obtain

$$S(y) - S(b) = \frac{1}{q(S(y))}(y - b) \text{ for } y \in J.$$

Since S and $q \circ S$ are not rational in general, we do need to verify that S has a "derivative" at $y = b$, i.e., that there exists a unique line $z = S(b) + m(y - b)$ which intersects the graph of S at a double point. The preceding formula obviously suggests that

$$m = \frac{1}{q(S(b))} = \frac{1}{q(a)} = \frac{1}{R'(a)}$$

will be the slope of the desired tangent line. In order to verify this, note that

$$S(y) - [S(b) + m(y - b)] = \left[\frac{1}{q(S(y))} - m\right](y - b).$$

Since q is rational, so is $1/q$, and hence there is a factorization

$$\frac{1}{q(x)} - \frac{1}{q(a)} = k(x)(x - a),$$

where k is rational as well. After substituting $x = S(y)$ and $a = S(b)$ it follows that

$$\begin{aligned}\frac{1}{q(S(y))} &= \frac{1}{q(a)} + k(S(y))(S(y) - S(b)) \\ &= \frac{1}{q(a)} + k(S(y))\frac{1}{q(S(y))}(y - b) \text{ for } y \in J.\end{aligned}$$

Combining these results one obtains

$$S(y) - [S(b) + m(y - b)] = \left(\frac{1}{q(a)} - m\right)(y - b) + \left[k(S(y))\frac{1}{q(S(y))}\right](y - b)^2,$$

which shows that the expression on the left has a zero of multiplicity at least 2 at $y = b$ (see footnote 2) if and only if $m = 1/q(a)$. So S has indeed an appropriate tangent line at $(b, S(b))$ whose slope is $S'(b) = 1/q(a) = 1/R'(S(b))$, as expected.

Remark. Note that the preceding discussion can be completely contained within the rational numbers $\mathbb{Q}$. Since no limits are ever taken, there is no need to invoke or use the completeness of $\mathbb{R}$. This is clear as long as one stays within the class of rational functions with rational coefficients. Of course, the values of inverse functions on $\mathbb{Q}$ will typically not lie in $\mathbb{Q}$, so the domains need to be restricted further. For example, for $f(x) = \sqrt{x}$, one considers the domain $D_f = \{x^2 : x \in \mathbb{Q} \text{ and } x \neq 0\}$. The relevant feature is that this restricted domain is still dense in the interval $\{x \in \mathbb{R} : x > 0\}$, i.e., there are no *visible* gaps in that part of the graph of f which consists of the points with *rational* coordinates.

4. ALGEBRAIC FUNCTIONS. Starting with the constant functions and $f(x) = x$, we generate a collection of functions $\mathcal{A}$ by applying the standard algebraic operations, compositions, and taking inverses a finite number of times, where the relevant functions are restricted to appropriate domains consisting of finite unions of open intervals, so that applicable quotients, compositions, and inverses are defined and differentiable.

The sum of two functions $f_1, f_2 \in \mathcal{A}$ with domains D_1 and D_2 is defined on the domain $D = D_1 \cap D_2$ provided D is not empty. Similar conventions need to be applied when one considers other algebraic operations involving functions in $\mathcal{A}$. By suitably generalizing the arguments used in Section 3 to functions $f \in \mathcal{A}$, one verifies the following statement.

Factorization Lemma. *If $f \in \mathcal{A}$ and a is in the domain of f, then there exists $q \in \mathcal{A}$ defined on the domain of f such that*

$$f(x) - f(a) = q(x)(x - a).$$

Corollary. *Given $f \in \mathcal{A}$ and the above factorization, then*

$$f(x) - [f(a) + m(x - a)] = (q(a) - m)(x - a) + k(x)(x - a)^2$$

for some other $k \in \mathcal{A}$ which is defined on the domain of f, and hence at a in particular.

Geometrically, this means that the line described by the linear function $y = f(a) + m(x - a)$ intersects the graph of $y = f(x)$ at $(a, f(a))$ with multiplicity at least two (see footnote 2) if and only if $m = q(a)$, and consequently the line given by $y = f(a) + q(a)(x - a)$ is the tangent to the graph of f at $(a, f(a))$, i.e., the function is differentiable at a with derivative $f'(a) = q(a)$. Furthermore, the usual rules of differentiation apply to functions $f \in \mathcal{A}$, and hence $f' \in \mathcal{A}$.

Example. Let $f(y) = \sqrt{y}$ be the inverse of $y = x^2$ on $x > 0$. Then f is differentiable on $I = (0, \infty)$ and

$$f'(y) = \frac{1}{2x} = \frac{1}{2\sqrt{y}}.$$

Let $g(x) = x^2 - 3x$. Since $g(x) > 0$ on $D = (-\infty, 0) \cup (3, \infty)$, the composition $(f \circ g)(x) = \sqrt{x^2 - 3x}$ is defined on D, is in $\mathcal{A}$, and is differentiable on its domain D, with

$$(f \circ g)'(x) = \frac{1}{2\sqrt{x^2 - 3x}}(2x - 3) \text{ for } x \in D.$$

It is natural to ask at this point how the double point method applies to the fundamental theorem of calculus. Of course integrals—even of simple rational functions such as $f(x) = 1/(1 + x^2)$—transcend the realm of algebra. Some version of limit, perhaps hidden behind suitable estimations, is needed to introduce definite integrals. Similarly, the definition of double point will need to be extended in order to accommodate nonalgebraic functions defined by integrals. In order to get to the heart of the matter quickly, at an elementary introductory level the basic properties can be introduced by appealing to the intuitive concept of area under the graph of a function, and that will suffice for our purposes. Specifically, only the following properties will be used in the proof of the version of the fundamental theorem of calculus stated below:

(i) linearity in the integrand,
(ii) additivity in the interval of integration,
(iii) the fact that the integral of a constant c over $[a, b]$ equals $c(b - a)$, and
(iv) the basic estimate: if $|f| \leq M$ on $[a, b]$, then $\left|\int_a^b f(x)dx\right| \leq M(b - a)$.

Theorem. *Suppose $f \in \mathcal{A}$ is defined on the open interval I and let c be a point in I. Define the function F on I by*

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

For any $a \in I$ the line given by $l(x) = F(a) + f(a)(x-a)$ intersects the graph of F at $(a, F(a))$ with an appropriately defined multiplicity greater than or equal to 2, i.e., the function F is differentiable at a with $F'(a) = f(a)$.

Proof. By the factorization lemma, $f(t) = f(a) + q(t)(t-a)$, where $q \in \mathcal{A}$. The properties of integrals listed above imply that

$$\begin{aligned} F(x) - F(a) &= \int_a^x f(t)\,dt = \int_a^x [f(a) + q(t)(t-a)]\,dt \\ &= f(a)(x-a) + Q(x), \end{aligned}$$

where $Q(x) = \int_a^x q(t)(t-a)dt$. Given a bounded closed interval $J \subset I$ such that a is in its interior, there exists $M < \infty$ such that $|q(t)| \le M$ for $t \in J$, and therefore $|q(t)(t-a)| \le M\,|x-a|$ for $x \in J$ and t with $|t-a| \le |x-a|$. By the basic estimate one thus obtains $|Q(x)| \le M\,|x-a|^2$ for $x \in J$. This last estimate, which would surely hold if Q were a function in $\mathcal{A}$ with a zero at a of multiplicity at least 2, is taken as the appropriate generalization of multiplicity greater than or equal to 2.[3] ∎

Note that one cannot conclude that Q, and consequently F, are functions in the class $\mathcal{A}$. In fact, as is well known, for $f(x) = 1/x$ on $(0, \infty)$ and $c > 0$ the function F is *not* in $\mathcal{A}$.

Finally we observe that the class $\mathcal{A}$ of functions is more restricted than the class of algebraic functions according to the following standard definition.

A function f defined on an interval I is *algebraic* if there exists a polynomial $P(x, y)$ such that $P(x, f(x)) = 0$ for all $x \in I$.

Starting with a polynomial $P(x, y)$, rather than focusing on an implicitly defined function f, whose *existence* typically requires tools outside of algebra, we consider the algebraic curve which is the graph of the equation $P(x, y) = 0$ and determine the tangent at a *regular*[4] point (a, b) on the curve by the double point method. Assuming degree $P = r \ge 2$ and $P(a, b) = 0$, by standard algebra there exist polynomials $A(x, y)$ and $B(x, y)$ such that

$$P(x, y) = A(x, y)(x-a) + B(x, y)(y-b).$$

While A and B are not determined uniquely by P and (a, b), their values at (a, b) are unique. For example, it follows from $P(x, b) = A(x, b)(x-a)$ that $A(a, b) =$

[3]The estimation can easily be improved by replacing $q(t)$ by $q(a) + k(t)(t-a)$ in the integral above, where k is an appropriate function in $\mathcal{A}$. It then follows that

$$F(x) - [F(a) + f(a)(x-a)] = \left[\frac{q(a)}{2} + r(x)\right](x-a)^2$$

for a function r which satisfies $|r(x)| \le M\,|x-a|$ for some constant M.

[4]Recall that a point (a, b) is a *regular* point of the curve defined by $P(x, y) = 0$ if at least one of the partial derivatives of P at (a, b) is nonzero. It is necessary to restrict the double point method to this special case, since at a nonregular point a (unique) tangent may not exist. For example, the graph of $xy = 0$ does not have a tangent at the point $(0, 0)$ in any obvious way.

$P_x(a, b)$, and similarly $B(a, b) = P_y(a, b)$, where we have used one of the standard notations for partial derivatives.

By introducing analogous factorizations of $A(x, y) - A(a, b)$ and $B(x, y) - B(a, b)$, and after substituting and rearranging, it follows that

$$P(x, y) = A(a, b)(x - a) + B(a, b)(y - b) + Q(x, y),$$

where $Q = (x - a)^2 Q_1 + (x - a)(y - b)Q_2 + (y - b)^2 Q_3$ for some polynomials Q_v, $v = 1, 2, 3$, of degree $\leq r - 2$. Since (a, b) is a regular point, let us assume first that $B(a, b) \neq 0$ and consider a line $y = b + m(x - a)$ through (a, b). The points of intersection of this line with the graph of $P(x, y) = 0$ are the solutions of $P(x, b + m(x - a)) = 0$. From the preceding representation of P one obtains the equation

$$\begin{aligned} P(x, b + m(x - a)) &= A(a, b)(x - a) + B(a, b)m(x - a) + Q^{\#}(x, y)(x - a)^2 \\ &= [A(a, b) + B(a, b)m](x - a) + Q^{\#}(x, y)(x - a)^2 \end{aligned}$$

for some polynomial $Q^{\#}(x, y)$ and $y = b + m(x - a)$. This shows that the multiplicity of the zero at $x = a$ is greater than one precisely when $A(a, b) + B(a, b)m = 0$. Since we assumed that $B(a, b) \neq 0$, this equation has the unique solution $m = -A(a, b)/B(a, b)$.[5] The line with that slope, whose equation we can rewrite as $A(a, b)(x - a) + B(a, b)(y - b) = 0$, is thus the unique line through (a, b) which intersects the curve in a point of multiplicity at least two. An analogous argument gives the same conclusion if $A(a, b) \neq 0$, which covers the case when $B(a, b) = 0$. Altogether, one obtains:

If $P(x, y) = A(x, y)(x - a) + B(x, y)(y - b)$ and either $A(a, b) \neq 0$ or $B(a, b) \neq 0$ (i.e., if (a, b) is a regular point of the curve $P(x, y) = 0$), then the line defined by

$$A(a, b)(x - a) + B(a, b)(y - b) = 0$$

is the tangent line to the curve at (a, b).

Example. Consider a point (a, b) on the circle $x^2 + y^2 = 1$. If $P(x, y) = x^2 + y^2 - 1$, then

$$P(x, y) = x^2 + y^2 - (a^2 + b^2) = (x + a)(x - a) + (y + b)(y - b).$$

The tangent line at (a, b) is thus given by

$$2a(x - a) + 2b(y - b) = 0,$$

where the factor 2 can of course be cancelled. Furthermore, if $b \neq 0$ the slope is $-a/b$, and if $b = 0$ (and hence $a = \pm 1$) the tangent line is given by $x = a$.

In general, given an algebraic curve $P(x, y) = 0$, one cannot prove the existence of a (locally) defined implicit function near a regular point (a, b) on the curve just by algebra.[6] On the other hand, if $P_y(a, b) \neq 0$ and one *assumes* existence and uniqueness of an implicit function $y = f(x)$ with $f(a) = b$ in a neighborhood of a (i.e.,

[5] Note that if $A(a, b) = B(a, b) = 0$, *any* m would be a solution of this equation, so the double point method breaks down at a nonregular point.

[6] By the implicit function theorem, there exists a unique solution $y = f(x)$ of $P(x, y) = 0$ near a point (a, b) with $P(a, b) = 0$ and $P_y(a, b) \neq 0$. Since P is a polynomial, $f(x)$ is real analytic near $x = a$, i.e., $f(x)$ is given by a power series in $(x - a)$ with a positive radius of convergence.

f satisfies $P(x, f(x)) = 0$ for x in a neighborhood of a), the result we just proved shows that there exists a unique line through $(a, f(a))$ which intersects the graph of f at a double point, i.e., that line is the tangent to the graph of f at that point. While the characterization of a double point we just used in the context of a curve given by $P(x, y) = 0$ is technically different from the one we have been using for explicit functions, the two characterizations obviously agree in the special case $P(x, y) = g(x) - y$ for a given polynomial g.[7] Hence it is reasonable to use the existence of such a tangent as the definition of differentiability of the implicitly defined algebraic function f. The derivative $f'(a)$, i.e., the slope of the tangent at $(a, f(a))$, is then given by $f'(a) = -P_x(a, f(a))/P_y(a, f(a))$, that is, one recovers the familiar formula for implicit differentiation.

5. MORE GENERAL FUNCTIONS. From today's perspective, the double point method discussed here cannot be extended to more general functions without coming to grips with limits, either in the context of continuity (see the next section) or in the context of convergence of power series. In the 17th and 18th centuries, however, power series were viewed as a routine extension of polynomials, without much concern about convergence. In fact, all elementary transcendental functions, such as the natural logarithm, exponential, and trigonometric functions, had been expanded into power series by a variety of ad hoc methods, and these series representations were freely used in the development of calculus. The binomial series discovered by Newton around 1665 provided further validation for manipulations of these infinite series, at least at a formal level. According to Carl Boyer, "Infinite series were no longer to be regarded as approximating devices only, they were alternative forms of the function they represented.... [E]ncouraged by Newton, men no longer tried to avoid infinite processes, as had the Greeks, for these now were regarded as legitimate in mathematics" [**3**, pp. 432–433]. Furthermore, at that time it was simply taken for granted that *all* functions could be represented by (formal) power series and that these series could legitimately be manipulated just like polynomials.

Consequently, if we accept this historical point of view there is then no problem with extending the algebraic double point method to "arbitrary" functions, as follows. Suppose

$$f(x) = \sum_{n=0}^{\infty} b_n x^n$$

is such a function, and fix a point a. To analyze $f(x)$ near this point, substitute $x = a + h$ in the series representing f, expand each term in the sum by the binomial theorem, and rearrange, combining all terms containing the same power of h, to obtain

$$f(x) = \sum_{n=0}^{\infty} c_n h^n = \sum_{n=0}^{\infty} c_n (x - a)^n,$$

with some new coefficients c_n which depend on a and the b_n. Since $c_0 = f(a)$, it follows that

$$f(x) - [f(a) + c_1(x - a)] = \sum_{n=2}^{\infty} c_n (x - a)^n = k(x)(x - a)^2,$$

[7] More generally, by variations of the methods used here, one can show that if the function $y = f(x)$ defined implicitly by the polynomial equation $P(x, y) = 0$ is known to be in the class $\mathcal{A}$, the two characterizations of "double point" still coincide.

where $k(x)$ is just another legitimate function represented by the series

$$k(x) = \sum_{n=2}^{\infty} c_n (x-a)^{n-2} = \sum_{n=0}^{\infty} c_{n+2} (x-a)^n.$$

As in the case of algebraic functions, this shows that the line $y = f(a) + c_1(x-a)$ intersects the graph of f in a double point at $x = a$. Hence this line is the desired tangent to the graph of f, and the derivative is given by $f'(a) = c_1$. In terms of the original series $f(x) = \sum_{n=0}^{\infty} b_n x^n$ this coefficient is given by

$$c_1 = \sum_{n=1}^{\infty} b_n \binom{n}{1} a^{n-1} = \sum_{n=1}^{\infty} n b_n a^{n-1}.$$

Clearly this answer agrees with the result obtained by formally applying the known formula for the derivative to each summand, i.e., by differentiating a series as if it were a polynomial.

To summarize, the algebraic double point method provides an approach to the differential calculus of *all* functions considered in the 17th and 18th centuries, which is free of any of the apparent inconsistencies created by infinitesimally small quantities and differentials. It should be mentioned in this context that the method described here results in a formula for the derivative of a function that agrees with the one introduced by J. Lagrange in 1797 [**11**]. Lagrange's goal was to eliminate the unresolved inconsistencies inherent in differentials and related vague notions that had been in use for well over 100 years. Functions were given by (formal) power series, and he defined the derivative of f at a point a to be the coefficient of h in the series expansion of $f(a+h)$, without any reference to infinitesimals and differentials. Starting from this definition, Lagrange rigorously (aside from considerations of convergence) proved all the familiar differentiation formulas. Lagrange did not use multiplicities, but motivated his definition by using infinitesimals to relate it to the classical quotient of vanishing differentials, i.e., the *motivation* remained in the conceptual framework of Leibniz and Newton. Lagrange's definition was not adopted widely, partly because proofs were rather complicated. More significantly, this definition could no longer form the basis of the differential calculus once it was recognized that the concept of function vastly extended beyond (formal) power series.

6. CARATHÉODORY'S DEFINITION OF DIFFERENTIABILITY. No discussion of derivatives by algebra is complete without establishing a connection to the modern definition of derivative based on limits. As seen in Section 4, the crux of the algebraic double point method is a **factorization lemma** for the class of functions under consideration, which extends the well-known elementary result for polynomials. Given such a factorization $f(x) - f(a) = q(x)(x-a)$, the "remainder" $R(x)$, i.e., the difference between the function $f(x)$ and its tangent $f(a) + q(a)(x-a)$, which is given by

$$R(x) = f(x) - [f(a) + q(a)(x-a)] = k(x)(x-a)^2,$$

has a zero of order at least two at $x = a$, so that f is differentiable at a according to the double point method, with derivative $f'(a) = q(a)$.

To handle more general classes of functions, a weaker requirement might be the existence of a line which intersects the graph at $(a, f(a))$ with an order—no longer

restricted to integer type—*greater than one*. For example, one could require that the remainder $R(x)$ has a factorization $R(x) = g(x)(x-a)$, where $|g(x)| \leq c\,|x-a|^d$ for some c and $d > 0$.[8] This condition would essentially still be algebraic and would not require any new concepts. On the other hand, to capture the most general condition requires the introduction of limits, as follows.

The function f is differentiable at $x = a$ if the remainder $R(x)$ has a factorization $R(x) = g(x)(x-a)$, where $\lim_{x\to a} g(x) = 0$ or, equivalently, g extends continuously to a with $g(a) = 0$.[9]

This version just states the well-known approximating property of a differentiable function by a *linear* function, which is equivalent to differentiability via limits of difference quotients. Our discussion suggests that this approximating property not only is the natural extension of the algebraic double point method, but that it really is much simpler and more transparent than the classical approach via limits of difference quotients. After all, since we all know that we cannot divide by 0, shouldn't we avoid anything that comes even close to a quotient with 0 in the denominator? Furthermore, this version of differentiability is the one that readily extends to functions of several variables, whether real or complex, as well as to calculus in infinite dimensions.

Finally, by combining the linear term with the remainder, differentiability as stated above is readily seen to be equivalent to the following version.

Definition. The function f is differentiable at $x = a$ if there exists a factorization $f(x) - f(a) = q(x)(x-a)$, where the factor $q(x)$ is continuous at a. The value $q(a)$ is called the derivative $f'(a)$ of f at $x = a$.

This formulation highlights the factorization that is the crux of the algebraic double point method discussed earlier in this article. The *new* property that is critical and that needs to be carefully defined is *continuity*, which of course requires an understanding of limits. But notice that for all functions studied in the 17th and 18th centuries the existence of the factorization, with the factor q belonging to the same known class of functions, was de facto obvious, and hence—as seen above—differentiation could easily have been done rigorously already at that time without burdening the discussion with infinitesimally small quantities.

Note that translating the equation

$$f(x) - f(a) = q(x)(x-a)$$

into

$$\frac{f(x) - f(a)}{x-a} = q(x) \text{ for } x \neq a$$

provides the interpretation of the factor $q(x)$ as a rate of change for $x \neq a$. Since q is a priori known to be continuous at a, the definition $f'(a) = q(a)$ shows that the derivative is well approximated by (average) rates of change. This is the crucial fact that is so useful in most applications. Note that in the algebraic case this approximation follows trivially from $|q(x) - q(a)| = |g(x)(x-a)| \leq c\,|x-a|$.

The proofs of the basic differentiation formulas we discussed in Section 3 in the algebraic context carry over immediately to differentiability as defined here by simply

[8] Note that the case $d = 1$ was already used in the proof of the fundamental theorem of calculus in Section 4.

[9] This condition for R is equivalent to what analysts commonly state as $R(x) = o(x-a)$ (R is *little oh* of $(x-a)$).

replacing the relevant algebra results by the corresponding statements about continuous functions. The convenience of this definition becomes even more apparent when one considers its natural generalization to functions of several variables, as follows, and the corresponding proof of the chain rule—notoriously one of the more difficult theorems even in one variable—by direct substitution.

Definition. The function $f(x_1, \ldots, x_n)$ is differentiable at $x = a$ if

$$f(x_1, \ldots, x_n) - f(a_1, \ldots, a_n) = \sum_{j=1}^{n} q_j(x)(x_j - a_j),$$

where the functions q_j are continuous at $x = a$ for $j = 1, \ldots, n$.[10]

The characterization of differentiability in terms of the remainder $R(x)$ is quite standard in today's calculus texts. It seems to have been introduced explicitly for the first time by Otto Stolz in 1893 [**14**, p. 28, 132] when he used it to state and rigorously prove key results in *multivariable* analysis. On the other hand, the final definition of differentiability stated above, which hides the remainder and emphasizes the factorization—while just a trivial modification of Stolz's classical version—is apparently little known, especially in the English literature, although it also has been around for quite a while. To my knowledge it occurs first in the work of Constantin Carathéodory (1873–1950). The earliest reference I could identify is his *Funktionentheorie*, published in 1950 [**4**]. I first learned of this definition in courses in complex analysis by Hans Grauert at the University of Göttingen in Fall 1964 (one complex variable), and in Spring 1965 (several complex variables). Soon thereafter Grauert used this definition in his *real* calculus text with W. Fischer and I. Lieb [**10**], and later also in his 1974 several complex variables text with K. Fritzsche [**9**]. Subsequently, it has also been used in other German complex analysis texts (e.g., Fischer and Lieb [**7**] and R. Remmert [**12**]). None of these texts makes any reference to its origins. Most likely Grauert and Remmert had learned of Carathéodory's formulation through their teacher Heinrich Behnke in Münster, who was well acquainted with Carathéodory. Fischer and Lieb then learned about this version from H. Grauert, their doctoral advisor. Aside from the calculus text of Grauert et al., I do not know of any occurrences of Carathéodory's formulation in *real* analysis texts until it was added in the 3rd edition of Bartle and Sherbert [**1**], published in the year 2000. The latter authors credit Carathéodory without giving any specific reference. More recently Carathéodory's version was also used by Ghorpade and Limaye [**8**, pp. 107–111,138], who make reference to [**1**].

The elegant simplicity of Carathéodory's definition, and the great ease by which it allows one to prove all the standard differentiation theorems—de facto reducing everything to appropriate basic theorems about continuous functions—surely impressed me a great deal when I first learned of it in the 1960s. Perhaps the time has come to use Carathéodory's definition more widely in the teaching of calculus.

ACKNOWLEDGMENTS. I would like to thank Lindsay Childs for suggesting that I study Lagrange's approach to derivatives without infinitesimals. I also thank Ingo Lieb for helpful information about Carathéodory's definition of differentiability, and for pointing me to the work of Otto Stolz. A referee of an earlier version of this article kindly provided the reference to Gaspard Monge's paper in which the point-slope form of a line

[10] It is straightforward to verify the equivalence of this definition with the standard definition of differentiability for functions of several variables as given in most analysis texts. Furthermore, note that $q_j(a) = \partial f/\partial x_j(a)$ for $j = 1, \ldots, n$.

was introduced. Thanks also to Antonella Cupillari, Julian Fleron, Friedrich Scholz, Jing Zhang, and my wife Sandrina for helpful comments. Last but not least I thank the referee of the current version for critical remarks which led to significant improvements.

REFERENCES

1. R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 3rd ed., John Wiley, New York, 2000.
2. C. B. Boyer, *History of Analytic Geometry*, Scripta Mathematica, New York, 1956.
3. ———, *A History of Mathematics*, John Wiley, New York, 1968.
4. C. Carathéodory, *Funktionentheorie*, Birkhäuser, Basel, 1950.
5. L. Euler, *Institutiones Calculi Differentialis*, 1755. Translated by J. D. Blanton as *Foundations of Differential Calculus*, Springer-Verlag, New York, 2000.
6. H. Eves, *An Introduction to the History of Mathematics*, 3rd ed., Holt, Rinehart and Winston, New York, 1969.
7. W. Fischer and I. Lieb, *Funktionentheorie*, F. Vieweg, Braunschweig, Germany, 1980.
8. S. R. Ghorpade and B. V. Limaye, *A Course in Calculus and Real Analysis*, Springer, New York, 2006.
9. H. Grauert and K. Fritsche, *Einführung in die Funktionentheorie mehrerer Veränderlicher*, Springer-Verlag, Berlin, 1974. Translated as *Several Complex Variables*, Graduate Texts in Mathematics, vol. 38, Springer-Verlag, New York, 1976.
10. H. Grauert, I. Lieb, and W. Fischer, *Differential- und Integralrechnung,* vols. I–III, Springer-Verlag, Berlin, 1967–68.
11. J.-L. Lagrange, *Théorie des Fonctions Analytiques*, Imprimérie de la République, Paris, 1797.
12. R. Remmert, *Funktionentheorie I*, Springer-Verlag, Berlin, 1984. Translated by R. B. Burckel as *Theory of Complex Functions*, Springer-Verlag, New York, 1991.
13. I. R. Shafarevich, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, 1977.
14. O. Stolz, *Grundzüge der Differential- und Integralrechnung*, Teubner Verlag, Leipzig, 1893.

R. MICHAEL RANGE was born in Germany and raised in Milano, Italy. He earned his Diplom in Mathematik at the University of Göttingen in 1968, where lectures of Hans Grauert got him hooked on multidimensional complex analysis. A Fulbright Fellowship brought him to the United States and UCLA, where he received a Ph.D. in 1971. He has held academic positions at Yale University and at the University of Washington, as well as research positions at institutes in Bonn, Stockholm, Barcelona, and Berkeley. He has published numerous research articles and is the author of *Holomorphic Functions and Integral Representations in Several Complex Variables,* first published in 1986 by Springer. More recently he has written articles dealing with historical aspects of multidimensional complex analysis, and he has also spent much time thinking about the calculus curriculum. Range loves mountains, and is an avid downhill skier. Inspired and guided by his son, he got into ice climbing and alpine mountaineering. He also enjoys biking and traveling, and about ten years ago he earned his private pilot certificate.
Department of Mathematics and Statistics, State University of New York at Albany, Albany NY 12222
range@math.albany.edu