A New Look at Calculus:

How an Old Idea Naturally Leads from Simple Algebra to the Heart of Analysis

R. Michael Range
State University of New York at Albany
and Park City, Utah

Or

CALCULUS:

Have We Been Teaching it Wrong?

In the typical introductory calculus course the tangent problem is motivated by some familiar geometric examples, e.g., tangents to circles,

but then quickly moves on to introduce the idea that the tangent to a curve at a point P arises as

the "limit" of lines through P and a

second distinct point $Q \neq P$ on the curve

as $Q \to P$.

Even as simple a case as $y = x^2$ requires the student to cope with the puzzling statement

$$\lim_{h \to 0}(2a + h) = 2a$$

While this seems "obvious" to students, instructors typically warn that one cannot just "evaluate" $h = 0$, but that something more complicated is going on.

The tangent problem is thus tied from the very beginning to the novel concept of "limit", something much more complicated than the student has encountered before.

Moreover, the definition of derivative through the limit of difference quotients that formally end up with the meaningless expression

$$\frac{0}{0}$$

makes matters even more complicated and mysterious for the student.

Typically, the further discussion of tangents and derivatives is then placed on hold and students are guided through a lengthy—and often quite technical—discussion of limits, continuity, and so on.

More often then not, this turns out to be quite boring for the student, or worse, the student finds it difficult to understand, gets lost, and looses interest.

But is this really necessary?

Please:

Try to forget everything you know about calculus!

Just remember your high school days,

and a bit about the quadratic equation, especially the special case of a double zero (discriminant = 0), and a bit about polynomials, such as the fact that if a polynomial f has a zero at a, then

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

where q is just another polynomial.

Thank you!

Let us now discuss a different approach.

It has its roots in the basic ancient question:

WHAT IS A TANGENT?

Over 2,300 years ago the Greek geometers Euclid and Apollonius formulated the essential idea as follows:

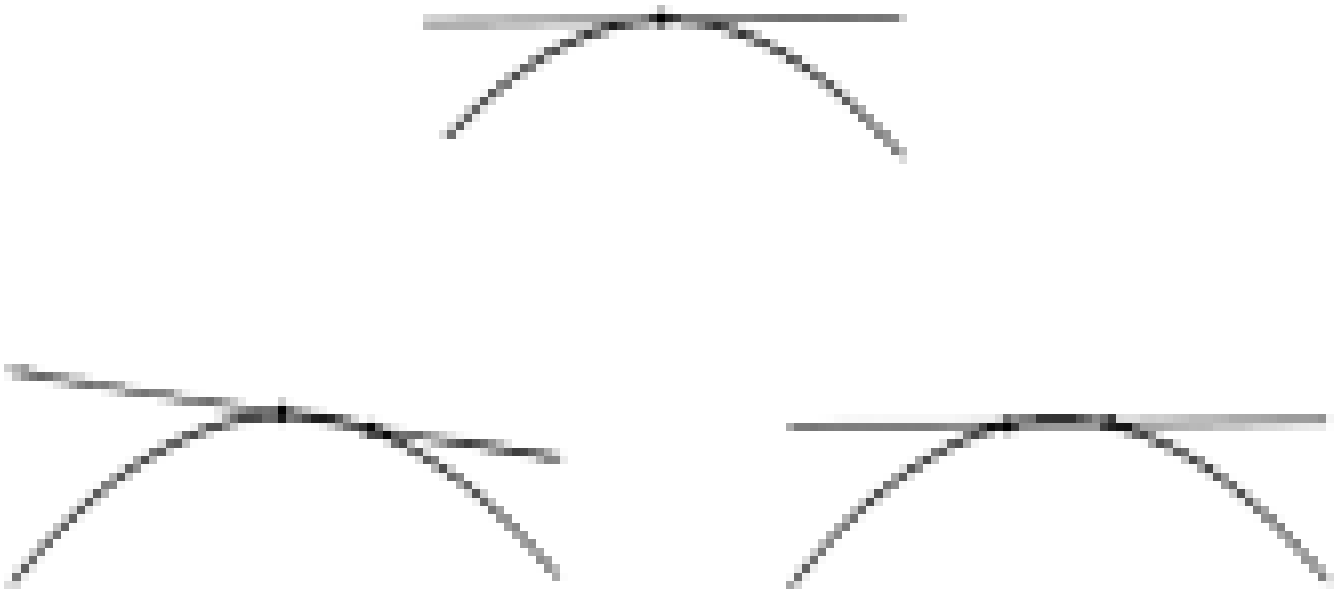
A tangent to a curve is a line
which touches the curve
but does not cut it!

This definition does not give any precise mathematical tools to construct tangents, but it does capture the critical idea:

A tangent **"touches"** the curve in a very special way.

"Touching" means that any small change of the "touching line" will either

- no longer "touch" the curve, that is, it misses the curve, or else
- it results in a line which "cuts" the curve (i.e., intersects it) in **two points** (or more).



The point of tangency – while it looks like a ***single*** point – really covers ***two*** points (or perhaps more) that become visible as soon as the line is perturbed just a bit.

We call such a point a

DOUBLE POINT!

Simple high school algebra allows to characterize such double points: they arise as solutions of **multiplicity two (or higher)** of the relevant equation.

Certainly the case of a quadratic equation, which in certain cases has a double zero, is well known to every high school student!

So let us *define* the tangent to a curve at ***P*** as a line which intersects the curve at ***P*** in a ***double*** (or higher) ***point***, and let us apply familiar simple algebraic techniques to identify such lines.

For example, consider a parabola, that is, the graph of $f(x) = x^2$, at the point (a,a^2). The equation of a line through that point is given by

$$y = a^2 + m(x - a),$$

where m is its 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 into

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

So consider the equation $\boldsymbol{[(x + a) - m](x - a) = 0}$.

The solutions $\boldsymbol{m - a}$ and $\boldsymbol{a}$ coincide, i.e., we have a double point of intersection, precisely when

$$\boldsymbol{m = 2a.}$$

So $\boldsymbol{m = 2a}$ is the desired slope of the tangent line at the point (a,a^2).

This method easily generalizes, first to arbitrary polynomials, and then to general algebraic functions, as follows.

Take any polynomial $P(x)$ of degree $r \geq 2$.
The equation of a generic line through the point $(a, P(a))$ is

$$y = P(a) + m(x - a)$$

where m is the slope. The points of intersection of this line with the graph of $P(x)$ are found by solving the equation $P(x) = P(a) + m(x - a)$, or

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

We need to find m so that this equation has a double (or higher) zero at $x = a$.

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 $r - 1$.

Then

$$P(x) - P(a) - m(x - a) = [q(x) - m](x - a)$$

This shows that $P(x) - P(a) - m(x - a) = 0$ has a

double zero at $x = a$

if and only if the polynomial $q(x) - m$ has a zero at the point a as well. Obviously this happens precisely when

$$m = q(a).$$

The number $q(a)$ is the desired slope of the tangent line!

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 is called the derivative of *P*, and it is denoted by

P'(a) **or** ***D(P)(a)*****.**

THEOREM

The slope of the tangent line is given by

$$D(P)(a) = q(a),$$

where *q* is the polynomial factor in the representation

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

REMARK:

All this works just as well for any *rational* function *R(x)*, with the factor *q(x)* now rational as well.

Example

Find the derivative of $f(x) = x^n$ at the point (a, a^n).

We factor

$$f(x) - f(a) = x^n - a^n =$$

$$= (x^{n-1} + x^{n-2}a + x^{n-3}a^2 + \ldots + x\,a^{n-2} + a^{n-1})(x - a) =$$

$$= q(x)\,(x - a).$$

Then

$$f'(a) = q(a) = na^{n-1}.$$

The idea to use double points is nothing new.....
the technique was considered by

René Descartes (1596 – 1650)

to find normals to the ellipse (and thereby find the tangents as well), and by his expositor

Frans van Schooten (1615 – 1660)

to find tangents directly.

GEOMETRIA,

à

RENATO DES CARTES

Anno 1637 Gallicè edita; postea autem
Vnà cum NOTIS

FLORIMONDI DE BEAVNE,

In Curia Blesensi Consiliarii Regii, Gallicè conscriptis in Latinam linguam versa, & Commentariis illustrata,

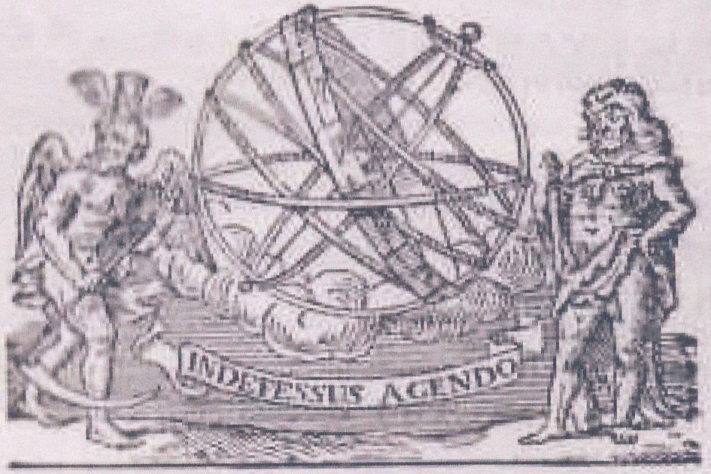
Operâ atque studio

FRANCISCI à SCHOOTEN,

in Acad. Lugd. Batava Matheseos Professoris.

Nunc demum ab eodem diligenter recognita, locupletioribus Commentariis instructa, multisque egregiis accessionibus, tam ad ulteriorem explicationem, quàm ad ampliandam hujus Geometriæ excellentiam facientibus, exornata,

Quorum omnium Catalogum pagina versa exhibet.



AMSTELODAMI,
Ex Typographia BLAVIANA, MDC LXXXIII.
Sumptibus Societatis.

The double point method seemed unsuitable for general curves, and it was eventually abandoned.

"Here we have a general process which tells us exactly what to do to solve our problem, but it must be confessed that in more complicated cases the required algebra may be quite forbidding."

(H. Eves: *An Introduction to the History of Mathematics*. 3rd ed., Holt, Rinehart & Winston, New York, NY 1969.)

Why did Descartes and van Schooten miss the elementary implementation of the

double point method

that we just discussed?

Perhaps they, as well as their contemporaries, were fixated in the Euclidean point of view that a line is defined by ***two distinct*** points.

In contrast, the ***point-slope form*** of lines was apparently unknown - or at least it was not used - in the 17^{th} and 18^{th} centuries.

Even Leonard Euler's influential classic calculus texts do not mention it.

It first appeared explicitly in 1784, in a paper by **Gaspard Monge** – well over a century after the beginnings of calculus.

Of course, by that time calculus via "differentials" and "infinitesimals" was well established and had been enormously successful, regardless of questions about the precise meaning of differentials and the foundations of calculus.

The usual rules of differentiation, including power rule, chain rule, differentiation of inverse functions at points where the derivative is nonzero, and product/quotient rules, can be proved in a straightforward manner based on the double point method and the critical factorization.

To illustrate this, let us look at the chain rule – notoriously viewed as somewhat difficult – and the details of whose proof are typically quite complicated.

As we see, the proof is very natural and completely elementary.

The **chain rule**, perhaps

the most important rule for derivatives,

is much simpler than the product rule, not to mention the quotient rule.

So why does every calculus book discuss the product and quotient rules

BEFORE

the chain rule??

By applying these techniques systematically, the double point method extends to all functions that are of algebraic type, that is, to those functions that are obtained from linear functions by applying the standard algebraic operations, compositions, and taking inverses (on suitably restricted domains) a finite number of times. Let us denote that class of functions by $\boldsymbol{A}$.

The essential result then is the following

Factorization Lemma

If $\boldsymbol{f}$ *is a function in* $\boldsymbol{A}$*, and* $\boldsymbol{a}$ *is a point in its domain, then one has*

$$\boldsymbol{f(x) - f(a) = q(x)(x - a)},$$

where $\boldsymbol{q}$ *is another function in* $\boldsymbol{A}$ *defined on the domain of* $\boldsymbol{f}$.

The concept of multiplicities of zeroes generalizes in the obvious way: a function g of algebraic type has a *zero of multiplicity* k at the point $x = a$ if there is a factorization

$$g(x) = q_k(x)\,(x - a)^k\,, \text{ with } q_k(x) \text{ in } A \text{ and } q_k(a) \neq 0.$$

Just as in the case of polynomials, it then follows from the factorization lemma that the equation

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

has a zero at $x = a$ of multiplicity ≥ 2 if and only if $m = q(a)$, where $q(x)$ is given by the factorization

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

Thus $q(a)$ is the slope of the *tangent* line, i.e., $q(a)$ is the derivative of f at $x = a$.

Conclusion

By studying the algebraic approach first (based on double points and multiplicities), students can understand derivatives, and learn and practice all differentiation formulas

before having to learn about limits.

The obvious question that arises at this point is:

How does this purely algebraic definition of derivatives relate to the idea that the derivative is the "limit" of difference quotients?

The answer involves another simple application of the basic factorization

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

combined with an elementary estimate, as follows.

Note that it is trivial that a polynomial is bounded over any bounded interval I. Therefore, if f, and hence also q, is a polynomial, given a bounded interval I centered at a, there exists a constant K, so that

$$|f(x) - f(a)| \leq K\,|x - a| \text{ for } x \text{ in } I.$$

This estimate makes precise what is obvious to the eye as one looks at the graph of a polynomial:

$$f(x) \to f(a) \text{ as } x \to a\,.$$

We see that the algebraic factorization naturally leads us to identify a special property that is universally known as “continuity”.

Just one simple estimate proves that every polynomial is continuous!

More generally, a little bit of additional work shows that every algebraic function in the class A is locally bounded. Consequently, just as in case of polynomials, the factorization implies:

Any function of algebraic type is ***continuous*** *at every point in its domain.*

We now apply this conclusion to the factor q in the factorization

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

to obtain that

$$q(x) \to q(a) \text{ as } x \to a.$$

Thus the (algebraic) derivative $D(f)(a) = q(a)$ is approximated by

$$q(x) = [f(x) - f(a)] / (x - a) \quad (x \neq a \textit{ here}),$$

that is, by the familiar difference quotient (or average rate of change) of f !

We thus see that the algebraic approach to derivatives based on double points and factorization directly leads to

a) A rigorous precise formulation of the intuitive idea of continuity.
b) A simple and direct proof of the continuity of all algebraic functions.

c) And finally, it shows that the algebraic derivative can also be captured by a *non-algebraic* approximation process (that coincides with the traditional definition of derivative).

Most significantly, once the case of functions of algebraic type is well understood, the preceding discussion, and in particular item c) above, suggests how to proceed with more general functions, such as exponentials, trigonometric, and other transcendental functions, where definitely new concepts need to be introduced.

Let us discuss just one example to illustrate the idea. Suppose we want to find the tangent at $(0,1)$ for the function $E(x) = 2^x$. Guided by the algebraic case, we consider the factorization

$$2^x - 2^0 = q(x)(x - 0).$$

While this equation of course defines

$$q(x) = (2^x - 1)/x \text{ for any } x \neq 0,$$

in contrast to the algebraic case there is now no obvious way to define $q(0) = ???$, and so the double point method breaks down.

Motivated by item c) in the algebraic case, one recognizes that one should

$$\textit{define } \boldsymbol{q(0)} \textit{ by the "limit" of } \boldsymbol{q(x)} \textit{ as } \boldsymbol{x \to 0},$$

so that the factor $\boldsymbol{q}$ is extended to $\boldsymbol{x = 0}$ as a *continuous* function.

The new difficult problem thus is how to verify the "existence" of such a "limit", and how to determine its value.

Geometric visualization of the line through the points ***(0,1)*** and ***$(x, 2^x)$*** for ***$x \neq 0$***, whose slope is given by ***$q(x)$***, suggests that there is indeed such a limit, and numerical data generated with a computer leads to the conclusion that

$$q(0) = \lim_{x \to 0} q(x) = 0.6931471\ldots .$$

The appearance of this strange decimal expansion reveals that deeper properties of numbers, such as *completeness,* need to be introduced - at least at an intuitive level - so that one is guaranteed that the limit indeed "exists" within the real numbers.

More precisely, the geometric visualization suggests that one should define

q(0) = greatest lower bound of {q(x): x > 0}.

Of course, the "existence" of such a number is guaranteed by the *completeness property* of the **real** numbers.

This is indeed the first place in the whole discussion where it is not enough to just consider the familiar rational numbers.

In other words, only when one considers the exponential function (or other *non-algebraic* functions) does the ***real*** need for new deep analysis concepts and tools become visible.

Guided by the algebraic case and the related discussion of continuity, one is now naturally led to consider the following generalization of the definition of algebraic differentiability.

DEFINITION: The function f defined near $x = a$ is ***differentiable*** at that point if there exists a factorization

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

where the factor q is continuous at $x = a$. The value

$$q(a) = \lim_{x \to a} q(x)$$

is called the ***derivative*** of f at a, and it is denoted by

$$D(f)(a) \text{ or } f'(a).$$

Of course, as we saw, every algebraic function is differentiable according to this definition at every point of its domain.

NOTE: The proofs of all the standard rules of differentiation given in the algebraic case apply directly in the more general case. One just needs to replace relevant properties of algebraic functions by the appropriate properties of *continuous* functions.

For example, the proof of the chain rule is reduced to the fact that composition and products of continuous functions are continuous.

This formulation of differentiability has been known for some time. As far as I know, it was introduced by

Constantin Carathéodory (1873 – 1950).

It appears explicitly in Carathéodory's 1950

"Funktionentheorie".

(Birkhäuser, Basel, 1950)

A translation into English was published in 1956.
(Chelsea Publ. Co., New York, 1956)

It has been used since the mid 1960s in several German texts, both in one and in several variables (real and complex).

It seems to have remained largely unknown in the English literature until fairly recently. The first occurrence known to me is in A. Browder's text, *Mathematical Analysis,* Springer, New York, 1996.

A few years later it was added in the 3rd edition of Bartle and Sherbert's *Introduction to Real Analysis,* published in the year 2000.

I believe that Carathédory's formulation offers several advantages that should make it the preferred definition in the 21st century:

- We all know that we cannot divide by zero.

  So we should avoid quotients with zeroes in the denominator as much as possible.

- It is the natural generalization

  of the elementary algebraic formulation.

- It simplifies proofs of standard rules,

  reducing technical details to basic natural properties of continuous functions.

- It is just a trivial modification of

  the fundamental idea that differentiability is equivalent to good linear approximation.

- It naturally generalizes to functions and maps

  of several variables, providing, in particular, much simpler proofs of the chain rule and of the inverse mapping theorem.

SUMMARY

• The algebraic definition of tangents and derivatives via double points provides an elementary approach WITHOUT LIMITS for an easy introduction to differential calculus.

• Students can learn all mechanical differentiation rules and many standard applications in a familiar setting without being burdened with deep new concepts involving limits.

- The algebraic approach leads naturally to the notion of continuity and limits in a concrete setting, and it paves the way for studying the calculus of non-algebraic functions, where limits become indispensable.

- The Chain Rule, i.e., the most important differentiation formula, is completely natural and elementary, and it should be discussed before the complicated product and quotient rules.

• Carathéodory's version of differentiability should be used more widely in calculus and analysis texts:

It is the natural generalization of the algebraic double point method, it provides simple proofs of standard theorems, and it allows a seamless transition from one to several variables.

References:

1. RMR: Where Are Limits Needed in Calculus?
Amer. Math. Monthly **118** (2011), 404 – 417.

2. RMR: Descartes' Double Point Method for Tangents: An Old Idea Suggests New Approaches to Calculus. Notices Amer. Math. Soc. **61** (2014), 387 – 389.

[A slightly revised translation into German has been published: *Von Descartes zu einem neuen Zugang zur Diff-erentialrechnung.* Mitteilungen der DMV **2016**, 26 – 29.]

3. RMR: *What is Calculus? From Simple Algebra to Deep Analysis*. World Scientific Publishing, Singapore, London, New Jersey, 2015.