HOLDEN, Lyman Sanford, 1926-  
MOTIVATION FOR CERTAIN THEOREMS OF THE  
CALCULUS.  
The Ohio State University, Ph.D., 1966  
Mathematics  

University Microfilms, Inc., Ann Arbor, Michigan
MOTIVATION FOR CERTAIN THEOREMS OF THE CALCULUS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the
Degree Doctor of Philosophy in the Graduate School of
The Ohio State University

By

Lyman Sanford Holden, A.B., M.A.

* * * * * *

The Ohio State University
1966

Approved by

[Signature]
Adviser
School of Education
ACKNOWLEDGMENT

The author is indebted to Dr. Harold C. Trimble for the encouragement and guidance shown during the writing of this dissertation. By his example Dr. Trimble has aroused continuing interest in mathematical education on the part of the author.
VITA

September 23, 1926 .......... Born, Oberlin, Ohio

1944-1946 ................. U. S. Naval Reserve

1950 ..................... A.B., Oberlin College, Oberlin, Ohio

1953-1956 ................. Teacher, Parma Public Schools,
                        Parma, Ohio

1956-1958 ................. Graduate Assistant, Department of
                        Mathematics, The Ohio State University

1958 ..................... M.A., Mathematics, The Ohio State
                        University

1958-Present .............. Instructor, Southern Illinois University,
                        Edwardsville Campus, Edwardsville, Illinois
## CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. THE PROBLEM</td>
<td>1</td>
</tr>
<tr>
<td>II. SCOPE AND METHOD</td>
<td>5</td>
</tr>
<tr>
<td>III. L'HOSPITAL'S RULE</td>
<td>9</td>
</tr>
<tr>
<td>IV. THE CHAIN RULE</td>
<td>31</td>
</tr>
<tr>
<td>V. CAUCHY'S FORMULA</td>
<td>52</td>
</tr>
<tr>
<td>VI. CURVATURE</td>
<td>70</td>
</tr>
<tr>
<td>VII. SUMMARY AND SUGGESTIONS FOR FURTHER STUDY</td>
<td>111</td>
</tr>
</tbody>
</table>

Bibliography | 118
I. THE PROBLEM

A basic problem in mathematical education is the problem of giving meaning to mathematical ideas as they are presented. It is widely held that methods of presentation are to be preferred which encourage insights on the part of the student into the nature of a result, and which lead him to see that the result should obtain. Current recommendations urge teachers to avoid vacuous manipulation of symbols in the mathematics classroom. The evidence seems to indicate that when the ideas are meaningful the student takes greater pleasure in his work, is usually more successful in problem solving, and is better able to generalize the ideas and apply them to new and different situations. The authors of the Twenty-Fourth Yearbook of The National Council of Teachers of Mathematics were cognizant of this problem, and included the following "axiom" of mathematics education on the first page of this book.

Axiom 1. The best learning is that in which the learned facts, concepts, and processes are meaningful to and understood by the learner.¹

In recent years considerable attention has been given to problems of providing motivation in the teaching of mathematics at the elementary and secondary levels. Improved instructional techniques are characteristic of a modern mathematics program. For example, teachers at these levels are encouraged to use the discovery method of teaching, the feeling being that heightened interest on the part of the student will result. Attesting to interest in this problem are many articles in professional journals that suggest methods for making mathematics meaningful to students. Some of the newer mathematics programs incorporate discovery methods and motivating techniques into their textbooks.

As yet, however, insufficient attention has been given to this problem as it relates to the teaching of college level mathematics. It is true that the college student has in general greater maturity than the student in the grades, and that in order to be in college he has necessarily shown scholastic ability. Nevertheless, it is the broad thesis of this study that at the college level too the best pedagogy will be characterized by an attempt to make the mathematical ideas meaningful to the student.

College teachers sometimes present mathematical ideas that are devoid of intuitive meaning for their students. It happens that a correct statement of a theorem is given, followed by a proof that is logically sound, while few clues are supplied which might help in

---

2 The terms "providing motivation" and "well-motivated" are used repeatedly throughout this study. The author's intention is that the word "motivated" shall be synonymous with the word "meaningful."
anticipating the result. Over the years many clever, short proofs for basic theorems of mathematics have been discovered. To discover such a proof is a source of great satisfaction for a mathematician. Yet the very brevity of the deduction may inhibit understanding on the part of the student.

Sometimes an auxiliary function is introduced at an early stage in a proof which, from the point of view of the student, is apparently pulled out of the air and bears no discernible relationship to the situation under consideration. Yet the auxiliary function is chosen so cleverly that the desired result follows easily. The use of such proofs compounds the learning problem for the student. He must first attempt to follow the sequence of steps set forth, and then must try to relate the logical argument to the theorem being proved. Too often the latter relationship is completely obscured by the brevity of the proof. The student may feel that although the proof appears to be logically valid, he is not convinced of the reasonableness of the result. Neither can he see how such a proof might have been discovered. The beginning calculus student finds it difficult enough to follow many of the usual proofs without unnecessarily increasing his frustration by the use of artificially introduced auxiliary functions.

This study will give suggestions for increasing the intuitive appeal of certain important theorems which are today a part of the first course in the calculus. While problems associated with the meaningful presentation of ideas should be considered in almost all mathematics
teaching, such problems are brought to the teacher's attention with
great force at several points in the calculus course. We are, therefore,
addressing here the question of motivating mathematics at the level
of the calculus.
II. SCOPE AND METHOD

This study will address problems of motivation in connection with four important theorems from elementary calculus: l'Hospital's rule, Cauchy's formula (the extended mean value theorem), the chain rule, and the formula for the curvature of a plane curve. These theorems belong to that part of the calculus known as the differential calculus. The theorems were selected because the writer believes that in each case the usual classroom presentation of the result is accomplished without good motivation.

For each of the four theorems we will present techniques whereby the results incorporated in these theorems can be made plausible. In each case a geometric representation for the theorem will be suggested. The pictorial representation is believed to be quite important for the student. It can be used to give meaning to the symbols that he is working with. It will greatly facilitate recalling the statement of the theorem, and will often supply clues to a proof or to an understanding of a proof. In order to treat l'Hospital's rule and the chain rule pictorially it will be necessary to make additional assumptions, to impose additional conditions, on the usual statement of the result. We will thereby lose some generality but will make gains with respect to motivating the result.

The methods to be suggested might be useful in classroom
presentations to students in the first year of their study of the calculus and in textbooks for the beginning course. They are designed to assist the student during his first encounter with the theorems by giving meaning to the symbols he is manipulating and by enabling him to see that the result should obtain.

The initial impetus for the study was received from a recent article by Christilles in which he calls for more creative textbooks. He says in this article:

A definite aid in the teaching of mathematics would be an increase in the number of available textbooks that demonstrate the actual manner in which the subject matter was developed. We have too long been satisfied with the traditional development standardized by the majority of our textbook writers.

If we are to produce creative, productive mathematicians we must teach them in a manner that clearly illustrates the means by which mathematics is discovered. Unfortunately, it is easier to write a well-organized textbook emphasizing the so-called natural development of the subject matter. This natural development rarely follows the actual order in which the mathematics was developed and, more significantly, usually de-emphasizes the initial inductive means used in obtaining the proofs of the theorems contained therein.\(^3\)

It is clear that a double benefit accrues from a well-motivated presentation of a result, a presentation that exposes the inductive methods used to discover the result. First, the result makes more sense to the student; the symbols have greater meaning, and he uses the result with confidence. Secondly, the student is being shown how a creative mathematician works. He sees an example of the development of a mathematical idea, and is brought to realize that a theorem does not spring full blown from the mathematician's head but is developed by hunches and guesses and trial proofs. Thus the student may realize that creativity in mathematics is not a super-human quality, but is instead one that he might be able to develop.

The basic methodology used in this study was searching out and examining older calculus textbooks. An attempt was made to find, through the study of older books, an inductive or heuristic basis for each of the four theorems cited earlier. In three cases it is felt that the historical search was successful.

In particular, the study includes a discussion of the Marquis de l'Hospital's proof of the rule that bears his name, as found in a book called Analyse des Infiniment Petits. This book, written by l'Hospital, has been called the first calculus textbook and was first published in 1696. From the examination of l'Hospital's proof we shall see how the first formula of the rule can be discovered, and under certain simplifying assumptions a very easy proof developed.

Through a study of ways to deduce the first mean value theorem, we shall obtain ideas for motivating the use of each of two auxiliary functions that are commonly employed in the proof of Cauchy's formula.
The formula itself, and the two auxiliary functions will each be given a geometric interpretation.

The formulas relating to the curvature of a plane curve receive extensive treatment in Chapter VI. The source of motivating ideas in this case is Sir Isaac Newton's book *The Method of Fluxions and Infinite Series*. We will see in this chapter how far removed the usual textbook proof of today is from the inductive methods used by this early writer.

In the case of the fourth theorem included here, the chain rule, the historical search failed to disclose how this theorem might have been discovered. The presentations of this theorem and the proofs of it are remarkably uniform over the years. A discussion of the chain rule will nevertheless be included, since an article recently published in *The Mathematics Teacher* has been of great assistance in obtaining a geometric interpretation of this rule.

The pages that follow contain the results of this investigation. As one might expect, the motivational method to be suggested changes from theorem to theorem. It is hoped that these results will serve two purposes: to show how the introduction of the particular theorems can be accomplished in a more meaningful way, and also to illustrate a research technique that is perhaps widely applicable. It appears that a search into the historical development of a mathematical idea may disclose an inductive basis for the idea.
III. L'HOSPITAL'S RULE

L'Hospital's rule is a cleverly conceived scheme for evaluating certain kinds of indeterminate forms which occur in the calculus. In this chapter we compare a modern formulation of L'Hospital's rule with the rule as it appeared for the first time in a calculus textbook, a textbook written by L'Hospital himself. Following the historical study of the rule, we state and prove a theorem which, the writer suggests, can be thought of as an analogue of the rule as given by L'Hospital. The theorem to be presented is not the most general one possible, but it does have two characteristics which recommend it for use in the classroom. First, the theorem incorporates the essential ideas of L'Hospital's rule. Secondly, both the theorem and its proof admit easy geometrical interpretations. The theorem can be used to familiarize a student with the basic techniques needed to apply the rule, and the proof is of a kind to convince him that the rule should obtain.

We will be primarily concerned in this chapter with the first formula of the rule. By the first formula we mean the evaluation of the indeterminate form $0/0$ as the quotient of two derivatives, each evaluated at a point $x = a$.

If $f(x)$ and $g(x)$ are continuous functions on an interval containing $x = a$, and if $f(a) = g(a) = 0$, then $\frac{f(x)}{g(x)}$ is said to assume the indeterminate form $0/0$ at $x = a$. In other words, $\frac{f(x)}{g(x)}$ is indeterminate.
at $x = a$ and of the form $0/0$. Under these conditions $\lim_{x \to a} \frac{f(x)}{g(x)}$ cannot be evaluated as $\lim_{x \to a} \frac{f(x)}{g(x)}$, since the theorem which states that $\lim_{x \to a} g(x)$ requires that $\lim_{x \to a} g(x)$ be different from zero. Nevertheless, the limit may exist. The fact that a limit may exist is illustrated with the basic formula for the derivative of a function $f(x)$:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Here both the numerator and the denominator of the difference quotient approach zero, yet the derivative which is the limit of that quotient may exist. L'Hospital's rule is a method which is often helpful in finding a value for an indeterminate form $0/0$, and it can also be used to evaluate other kinds of indeterminate forms.

The student's first introduction to L'Hospital's rule is often a bewildering experience. The bewilderment is caused by two characteristics of our modern presentation. First, L'Hospital's rule as given in a modern textbook is not a single rule but a sequence of several theorems. The rule is so widely applicable that the statement becomes very cumbersome. Textbook authors usually try to include most of the special cases in which the rule can be used, and the result is a complicated series of theorems. Secondly, the usual proof is poorly motivated. This proof uses Cauchy's formula (also called the extended mean value theorem) to establish in just a few lines the first formula of L'Hospital's rule. But Cauchy's formula is itself a rather complicated expression, and, for
the student, is not clearly related to the situation under consideration. For these reasons the student usually experiences some difficulty with l'Hospital's rule.

To illustrate these remarks, and for reference purposes, we include at this point a modern statement of the rule. These theorems are thought to be representative, and are taken from a calculus book by Goodman. The roles of $f(x)$ and $g(x)$ have been interchanged to gain consistency with the discussion to follow.

Theorem 2 (l'Hospital's rule). Suppose that $f(a) = g(a) = 0$, $f(x)$ and $g(x)$ are each continuous on $a < x < b$, and differentiable in $a < x < b$, and $g'(x) \neq 0$ in $a < x < b$. Then

$$\lim_{x \to a^+} \frac{f(x)}{g(x)} = \lim_{x \to a^+} \frac{f'(x)}{g'(x)}$$

whenever the latter limit exists. If the limit on the right side of equation (9) is $+\infty$, or $-\infty$, then the limit on the left side is $+\infty$ or $-\infty$ respectively.

Proof. We apply Theorem 1 [the generalized mean value theorem, Cauchy's formula] but instead of the interval $a \leq x \leq b$, we use $t$ as the variable and $x$ as the righthand end point. Thus we consider the interval $a < t < x$. The generalized mean value theorem states that there is a \( \xi \) with $a < \xi < x$ such that

$$\frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$  

---

But \( g(a) = f(a) = 0 \), so (10) simplifies to

\[
\frac{f(x)}{g(x)} = \frac{f'(\xi)}{g'(\xi)}, \quad a < \xi < x. \tag{11}
\]

Now let \( x \to a^+ \). Since \( a < \xi < x \), \( \xi \) must also approach \( a^+ \). Hence if

\[
\lim_{\xi \to a^+} \frac{f'(\xi)}{g'(\xi)} = L
\]

then by (11) the left side of (9) has the same limit \( L \). But this is just the statement of Theorem 2.

Q.E.D.

L'Hospital's rule is exactly the same in case the independent variable \( x \) is tending to \( \infty \) instead of some finite number \( a \). Precisely stated we have

Theorem 3. Suppose that \( \lim_{x \to \infty} f(x) = 0 \) and \( \lim_{x \to \infty} g(x) = 0 \), \( f(x) \) and \( g(x) \) are each differentiable in \( M < x < \infty \), and \( g'(x) \neq 0 \) in \( M < x < \infty \). Then

\[
\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)} \tag{14}
\]

whenever the latter limit exists.

One of the attractive features of l'Hospital's rule is that it works for the indeterminate form \( \infty/\infty \) just as it works for \( 0/0 \).
Theorem 4. Let \( \lim_{x \to a} f(x) = \infty \) and \( \lim_{x \to a} g(x) = \infty \), and suppose that \( g'(x) \neq 0 \) in \( a < x < b \). Then

\[
\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}
\]

(15)

whenever the latter limit exists.

Thus we can see the complexity of this rule as it is known today, and have an indication of the variety of special cases to which it is applied.

It is our intention to compare these theorems with the statement of l'Hospital's rule as it appeared for the first time in a calculus textbook. This rule was included in the book *Analyse des Infiniment Petits*, written by the Marquis de l'Hospital, a book that has been called the first calculus textbook. This book, first published in 1696, enjoyed great popularity throughout much of the eighteenth century as is evidenced by the fact that new editions appeared in the years 1715, 1730, 1768, and 1781. This writer was able to examine one of the 1781 editions, which contains, in addition to the original text, commentary and expository notes by a M. Le Fevre. The use of different type sizes

---


in this edition makes it easy to distinguish M. Le Fevre's remarks from the original writing of l'Hospital. We continue with l'Hospital's presentation of the rule that bears his name as found in this book under the unenlightening designation "Solution de quelques Problèmes qui dépendent des Méthodes précédentes."

Problem

Fig. 130. 163. Suppose we are given a curve AMD (AP = x, PM = y, AB = a) such that the value of the ordinate y is expressed as a fraction, where the numerator and the denominator each become zero when x = a, that is, when the point P coincides with the given point B. We ask, what then should be the value of the ordinate BD?

Let there be given two curves ANB, COB, which have the line AB as their common axis and such that the ordinate PN represents the numerator, and the ordinate PO the denominator of the fraction which everywhere gives PM: so that

\[ PM = \frac{AB \times PN}{PO} \] . It is clear that these two curves meet at the point B, by the supposition that PN and PO each become zero when the point P coincides with the point B. If one imagines an ordinate bd infinitely close to BD, which intersects the curves ANB, COB at points f, g, then it follows that

*Art. 2. \[ bd = \frac{AB \times bf}{bg} \] , which does not differ from BD. Therefore we have only to find the ratio of bg to bf. Now it is clear that the abscissa AP becomes AB, the ordinates PN and PO become zero, and as AP becomes Ab they become bf and bg. From this it follows that these ordinates bf, bg are the differences of the ordinates at B and b with reference to the curves ANB, COB; and consequently that if one takes the difference of the
numerator and divides it by the difference of the denominator, after setting $x = a = Ab$ or $AB$, one has the value that we seek for the ordinate $bd$ or $BD$. That which was to be shown.

Our Figure 1 above shows the drawing that accompanies l'Hospital's discussion of this rule.

\[\text{Figure 1}\]

\[\text{Ibid., pp. 186, 187.}\]
Immediately following the presentation of the rule, l'Hospital applies it to the following example:

Example 1.

Let \[ y = \frac{\sqrt{2a^3x - x^4} - a^{3/2}a x}{a - \sqrt[3]{a} x^3} \].

Clearly, when \( x = a \), the numerator and the denominator of the fraction each equal zero. Therefore we take the difference

\[ \frac{a^3 \frac{dx}{x} - 2x^3 \frac{dx}{x^4}}{\sqrt{2a^3x - x^4}} - \frac{a a \frac{dx}{x}}{3 \sqrt[3]{a} x^3} \]

of the numerator, and divide it by the difference

\[ - \frac{3a \frac{dx}{x}}{4 \sqrt[4]{a^3x}} \]

of the denominator, after setting \( x = a \); i.e., we divide

\[ - \frac{4}{3} a \frac{dx}{x} \]

by \(- \frac{3}{4} \frac{dx}{x}\). This gives \( \frac{16}{9} a \) as the value of BD which we seek.⁹

Not unexpectedly the mathematics presented here appears somewhat obscure to a modern reader. By present standards, the language is imprecise and the symbolism is inadequate. But the inadequacies are to be attributed to the mathematics of 1696, not to l'Hospital. The calculus had but recently been invented. It was to be 125 years until Cauchy, in the 1820's, published the fundamental notions of the calculus with a rigor approaching that of our own today. To Cauchy goes much of the credit for the adoption of the theory of limits as the true metaphysics of the calculus. And it was still later, around 1870, that

---

⁹Ibid., p. 187.
Weierstrass and his school accomplished the arithmetization of analysis. Consequently the imprecision is not surprising.

Let us consider some of the more obvious ways in which l'Hospital's writing and thinking, as exemplified in this rule, differs from ours today. In so doing we will interpret his ideas in modern notation and this result will appear more like the rule as we know it. We will then go on to present some ideas for motivating the first formula of the rule, ideas which have been developed after careful consideration of l'Hospital's problem.

L'Hospital does not use the Cartesian system of axes with which we are so familiar. He refers to the line AB as a common axis of the two curves, and implies (with the equation $AB = a$) that $A$ is the zero point on this axis. But we seek in vain for any reference to another axis perpendicular to this one. The temptation is strong to superimpose a coordinate axis system onto Figure 1 with the origin at $A$ and the $x$-axis along $AB$. If one does this, however, the figure becomes incorrect, for $PN$ would be positive, $PO$ negative, and the quotient of these numbers would be negative. Thus $M$ should be below the line $AB$, not above it. When we subsequently present a modern analogue of l'Hospital's figure we must be careful to represent the quotient function as above the $x$-axis when both $N$ and $O$ are on the same side of the axis, and below this axis when $N$ and $O$ are on opposite sides.

We also remark the absence of functional notation. L'Hospital talks about "... the curve $AMB$ ..." as if the collection of points itself defined a relationship, rather than about a functional relationship between two variables, $y = f(x)$, which has this curve as its graph.
Had the concept of a function and an adequate notation been developed at this time, l'Hospital could have expressed his ideas more succinctly. ANB would be the graph of a function $f(x)$, COB the graph of a function $g(x)$, and AMD the graph of the quotient function $f(x)/g(x)$. Furthermore, one of the conditions becomes $f(a) = g(a) = 0$. The number that is sought, the value of the ordinate BD at $a$, has the modern name $\lim_{x \to a} [f(x)/g(x)]$. It is also necessary to ask what l'Hospital meant by the word "difference." On the first page of Analyse he defines a variable as a quantity which continuously increases or diminishes; and the infinitely small amount by which a variable increases or diminishes in certain situations is called the "difference" of that variable. Thus it appears that differences in variables, $x$ and $y$ say, would correspond to the quantities $\Delta x$ and $\Delta y$. Later in the text, however, it becomes clear that l'Hospital's difference is what we would now call a differential. Thus if $y = f(x)$, then $dy = f'(x) \, dx$, and this $dy$, presently called a differential, is l'Hospital's difference. We know that at a point, for a given $dx$, $dy$ is an approximation to $\Delta y$. It is doubtful that l'Hospital had in mind any distinction between $\Delta y$ and $dy$. The belief that his difference is our differential is reinforced when we read, in the example, that

the difference of $\sqrt{2a^3x - x^4} - a^3 \sqrt{a}x$ is $\frac{a^3 \, dx - 2x^3 \, dx}{\sqrt{2a^3x - x^4}} - \frac{a \, dx}{3 \sqrt{a}x}$, clearly a differential.

---

10 Ibid., p. 1.
The reference to Article 2 is interesting. In Article 2 we find this statement, taken as a supposition: "We agree that one can use indifferently one for the other of two quantities which differ only by an infinitely small amount." 11 It is this supposition that allows l'Hospital to say that bd does not differ from BD. Such shenanigans were used in place of the limiting processes that we now employ. This supposition is one of two 12 which form the basis for his work and which for l'Hospital appeared to be "... so self-evident as not to leave the least doubt about their truth and certainty on the mind of an attentive reader." 13 He adds that if he were so inclined he could prove them in the manner of the ancients. Many generations of mathematics students must have accepted statements of this kind as axiomatic, since they were proposed by such authoritative mathematicians, the while being a bit mystified by such reasoning. The history of infinitesimals is a fascinating story to a mathematics teacher, and serves to illuminate the process of creation in mathematics. 14

11 Ibid., p. 2.

12 The other supposition, found on page 3 of the same work, is this: "We assume that a curved line can be considered a collection of an infinite number of straight lines, each infinitely short." To define a tangent line to a curve at a point l'Hospital simply extends one of these little line segments.

13 Ibid., Preface, p. xix.

We give now a theorem which can be thought of as a modern analogue of l'Hospital's problem. The theorem is a simplified version of the first formula of the rule, yet incorporates the essential ideas of this rule. This theorem is recommended to us by its simplicity, and by the fact that it lends itself to pictorial representation. The statement and proof of the theorem will be followed by a discussion of the geometric interpretation.

Theorem A. Let \( f(x) \) and \( g(x) \) be functions with the following properties:

1. \( f(x) \) and \( g(x) \) are continuous on an interval containing \( x = a \) as an interior point,
2. \( f(a) = g(a) = 0 \),
3. \( f'(a) \) and \( g'(a) \) exist, with \( g'(a) \neq 0 \),
4. \( g(x) \neq 0 \) for \( x \) in the interval, but \( x \neq a \).

Then

\[
\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}.
\]

Proof. Since \( f(a) = g(a) = 0 \), it follows from the continuity of \( f(x) \) and \( g(x) \) that \( \lim_{x \to a} f(x) = f(a) = 0 \), and \( \lim_{x \to a} g(x) = g(a) = 0 \). Thus \( \lim_{x \to a} \frac{f(x)}{g(x)} \) cannot be evaluated as the quotient of two limits. But since \( f(a) = g(a) = 0 \), we have for \( x \) near \( a \)

\[
\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{[f(x) - f(a)]/[x - a]}{[g(x) - g(a)]/[x - a]}.
\]
But \( \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = f'(a) \), and \( \lim_{x \to a} \frac{g(x) - g(a)}{x - a} = g'(a) \neq 0 \). Consequently

\[
\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \cdot \lim_{x \to a} \frac{g(x) - g(a)}{x - a} = \frac{f'(a)}{g'(a)}
\]

The writer suggests that an immediate improvement can be made in the heuristic appeal of l'Hospital's rule by using Theorem A accompanied by a diagram showing the graphs of the functions involved. In no textbook that the writer has examined has this been done. We are basically concerned here with two functions, \( f(x) \) and \( g(x) \), which satisfy the conditions \( f(a) = g(a) = 0 \), and with the function \( \frac{f(x)}{g(x)} \) which is the quotient of these two functions. A figure similar to Figure 2 below allows the student to see the functions \( f(x) \) and \( g(x) \), to see that these functions have zeros at \( a \), and to visualize the quotient function \( \frac{f(x)}{g(x)} \). The fact the \( \frac{f(x)}{g(x)} \) is undefined at \( x = a \) (since \( g(a) = 0 \)) can also be represented graphically by indicating a missing point in the graph of this function.

![Figure 2](image-url)
In Figure 2 the graph of \( \frac{f(x)}{g(x)} \) is drawn as a continuous curve, since this function is continuous as long as \( g(x) \neq 0 \). The dotted line represents \( \lim_{x \to a} \frac{f(x)}{g(x)} \), the number that we seek. Thus the basic entities involved in the theorem are all pictured.

There are problems associated with obtaining the graph of the quotient function \( \frac{f(x)}{g(x)} \). In theory, the definition of this function is quite simple: \( \frac{f(x)}{g(x)} \), provided \( g(x) \neq 0 \). In practice, graphing the quotient function can be very tedious if one proceeds by computing values at isolated points. Fortunately, a technique for constructing with straight edge and compass the graph of the quotient of two functions from the graphs of the individual functions has recently been published by Haddock and Hight.\(^{15}\) The method they suggest can be used to advantage here, and was used to obtain the graph of the quotient function in the final example of this chapter.

The appeal of this theorem is increased by the fact that the essential equations in its proof can also be given graphic meaning. In fact, a method of proof suggests itself when one makes a certain key observation, which is described in the next paragraph. This observation is readily accesible to the student, and it seems likely that this is the relationship that l'Hospital saw in his Figure 130. The symbol "\( \equiv \)" is here used to mean "identically equal to."

Because \( f(a) = g(a) = 0 \), we have \( f(x) = f(x) - f(a) \) and 
\( g(x) = g(x) - g(a) \). Consequently

\[
\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)}
\]

for all \( x \) near \( a \) but different from \( a \). Note that \( g(x) - g(a) \neq 0 \) here, since \( g(x) \neq 0 \) for \( x \neq a \). Thus the ordinate of the quotient function \( \frac{f(x)}{g(x)} \) is identically equal to an expression that begins to suggest the quotient of two derivatives. These quantities can be seen in Figure 3.

The suggestion becomes stronger upon dividing numerator and denominator by \( [x - a] \), to obtain

\[
\frac{f(x)}{g(x)} = \frac{[f(x) - f(a)]/[x - a]}{[g(x) - g(a)]/[x - a]},
\]

which expresses the ordinate of the quotient function as the quotient of two difference quotients. The remaining steps should be easy for the student to follow.

Since \( \lim_{x \to a} [f(x) - f(a)]/[x - a] = f'(a) \), and \( \lim_{x \to a} [g(x) - g(a)]/[x - a] = g'(a) \neq 0 \), we have

\[
\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{[f(x) - f(a)]/[x - a]}{[g(x) - g(a)]/[x - a]} = \lim_{x \to a} \frac{[f(x) - f(a)]/[x - a]}{g'(a) \neq 0} = \lim_{x \to a} \frac{[g(x) - g(a)]/[x - a]}{[x - a]} = \frac{f'(a)}{g'(a)}.
\]

It should come as no surprise to the student that \( g'(a) \) need not be zero when \( g(a) \) is zero, but it can nevertheless be pointed out in the figure that the slope of the curve \( g(x) \) is in general not zero at a zero of the function. This means, of course, that \( \frac{f'(a)}{g'(a)} \) is determinate even though \( \frac{f(x)}{g(x)} \) is indeterminate, which is the essential kernel of l'Hospital's rule.
It is usual, when teaching l'Hospital's rule, to caution the student against the common mistake of differentiating \( \frac{f(x)}{g(x)} \) as a quotient. It is the quotient of two derivatives that we seek, not the derivative of a quotient. A study of Theorem A should help eliminate this common mistake. The student can see a derivative being formed in the numerator and in the denominator, and the usual admonition should be unnecessary.
We next investigate the limitations of Theorem A. How restrictive is this result, and what generality has been lost? In the first place, the theorem treats only the indeterminate form 0/0 which a quotient assumes as \( x \to a \). A glance at the statement of l'Hospital's rule from Goodman's book shows the variety of special cases in which the rule can be used, only one of which is treated in our theorem.

In the second place, the simplified theorem does not permit successive applications of the rule to the same indeterminate form. A typical statement of the rule imposes certain conditions on the functions \( f(x) \) and \( g(x) \), and then gives a conclusion in this form:

\[
\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)},
\]

provided the latter limit exists. Now it may happen that \( \lim_{x \to a} \frac{f'(x)}{g'(x)} \) is again of the form 0/0. With the conclusion as stated here, when \( f'(x) \) and \( g'(x) \) satisfy the conditions one can apply the rule again and conclude without further proof that

\[
\lim_{x \to a} \frac{f'(x)}{g'(x)} = \lim_{x \to a} \frac{f''(x)}{g''(x)}.
\]

This process continues until one arrives at a determinate limit,

\[
\lim_{x \to a} \frac{f^{(n)}(x)}{g^{(n)}(x)},
\]

where \( f^{(n)}(a) \) and \( g^{(n)}(a) \) both exist and \( g^{(n)}(a) \) \( \neq 0 \).

The assertion is then made that

\[
\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{f^{(n)}(a)}{g^{(n)}(a)}.
\]

For example,

\[
\lim_{x \to 0} \frac{x - \sin x}{x^3} = \lim_{x \to 0} \frac{1 - \cos x}{3x^2} = \lim_{x \to 0} \frac{\sin x}{6x} = \lim_{x \to 0} \frac{\cos x}{6} = \frac{1}{6}.
\]
Now one of the assumptions of Theorem A is that one obtains a
determinate limit after just one differentiation of each function.
Clearly the conclusion we have used, that \( \lim_{x \to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)} \), lacks
generality in that it does not allow this kind of repeated application
of the rule. But even in the case where repeated differentiations are
needed, the student eventually performs an evaluation in the manner
suggested by our simplified theorem. The only difference is that he is
evaluating n-th order derivatives rather than 1st-order derivatives.
Thus it is felt that the lack of generality in this respect does not
materially distract from the value of the theorem as a method for
motivating l'Hospital's rule.

We suggest that Theorem A can be used, in the classroom and in
textbooks, to increase the intuitive appeal of l'Hospital's rule. The
student sees that at least the first formula of the rule is a plausible
result. Since Theorem A incorporates many of the basic ideas and tech­
niques connected with l'Hospital's rule, through its use the student
can learn these techniques in a meaningful way.

We select the problem of finding

\[
\lim_{x \to 1} \frac{\ln x}{x - 1}
\]

to illustrate these remarks. Note that \( \frac{\ln x}{x - 1} \) assumes the form 0/0 as
x goes to 1. Referring to the notation of Theorem A, \( f(x) = \ln x \),
\( g(x) = x - 1 \), \( \frac{f(x)}{g(x)} = \frac{\ln x}{x - 1} \), and \( a = 1 \). We check to see that these
functions possess the properties required by the theorem.

(1) \( g(x) \) is continuous everywhere, and \( f(x) = \ln x \) is continuous for
x > 0, hence both functions are continuous at x = 1.
(2) \( f(1) = \ln 1 = 0 \), and \( g(1) = 1 - 1 = 0 \).

(3) \( f'(x) = \frac{1}{x} \) exists for \( x > 0 \), and \( g'(x) \equiv 1 \); thus \( g'(1) = 1 \neq 0 \).

(4) \( g(x) \) has just one zero, at \( x = 1 \), hence \( g(x) \neq 0 \) for \( x \neq 1 \).

Thus from Theorem A and the facts that \( f'(1) = 1/1 = 1 \) and \( g'(1) = 1 \), we obtain

\[
\lim_{x \to 1} \frac{\ln x}{x - 1} = \frac{1/1}{1} = 1.
\]

The process of graphing the functions involved here can be a profitable exercise for a student. The graphs will show the properties possessed by the functions, and will illustrate the meaning of the limit number, 1. Figure 4 shows graphs of \( f(x) = \ln x \), \( g(x) = x - 1 \), and \( f(x)/g(x) = [\ln x]/[x - 1] \).

Note that this figure allows the student to see that \( x = 1 \) is a common zero of the functions \( \ln x \) and \( x - 1 \), that \( \ln x \) and \( x - 1 \) are continuous at \( x = 1 \), and that these functions possess derivatives there. Of greater importance is the fact that the student can see, not only the quotient function \([\ln x]/[x - 1]\), but also that the ordinates of this function are close to 1 for \( x \) near 1 on the \( x \)-axis. Thus the dotted line in Figure 4 gives pictorial meaning to the equation

\[
\lim_{x \to 1} \frac{\ln x}{x - 1} = 1.
\]

The graph of \([\ln x]/[x - 1]\) can be obtained using straight edge and compass in the manner suggested by Haddock and Hight, or by point-wise computation of \([\ln x]/[x - 1]\) using a table of natural logarithms. The use of coordinate paper is highly recommended.
\[
\frac{f(x)}{g(x)} = \frac{\ln x}{x-1}
\]
The problem of constructing the graph of \( \frac{\ln x}{x - 1} \) is a bit richer in interesting side results than one might at first expect. Since the student is interested in the behavior of this function near \( x = 1 \), he will first obtain some points on the graph corresponding to \( x \)'s near 1. With careful work, the points on \( \frac{\ln x}{x - 1} \) will be above the line \( y = 1 \) for \( x \)'s to the left of \( x = 1 \), and below this line for \( x \)'s to the right of \( x = 1 \). But this question naturally arises: What is the behavior of the quotient function as \( x \to 0^+ \), and as \( x \to +\infty \)? Answers can be found in the following way.

For \( 0 < x < 1 \), both \( \ln x \) and \( x - 1 \) are negative, so \(|\ln x| = -\ln x\), \(|x - 1| = -(x - 1)\), and

\[
\frac{|\ln x|}{|x - 1|} = \frac{-\ln x}{-(x - 1)} = \frac{\ln x}{x - 1}.
\]

Furthermore, in this interval \(|\ln x| > |x - 1|\), or \(|\ln x|/|x - 1| = \ln x/(x - 1) > 1\). Moreover, \(|\ln x| \to +\infty\) as \( x \to 0^+ \) and \(|x - 1| \to 1\) as \( x \to 0^+ \). Thus

\[
\frac{|\ln x|}{|x - 1|} = \frac{\ln x}{x - 1} \to +\infty \quad \text{as} \quad x \to 0^+.
\]

For \( 1 < x < +\infty \), \( \ln x \) and \( x - 1 \) are both positive, and \( \ln x < x - 1 \). Consequently, in this interval \( \frac{\ln x}{x - 1} < 1 \). Now investigation of the limiting behavior of this quotient as \( x \to +\infty \) leads naturally into an extension of l'Hopital's rule to another indeterminate form. As \( x \to +\infty \), \( \ln x \to +\infty \) and also \( x - 1 \to +\infty \), so that \( \frac{\ln x}{x - 1} \) assumes the indeterminate form \( \infty/\infty \). These observations yield no information for the student about \( \lim_{x \to +\infty} \frac{\ln x}{x - 1} \).

It can be suggested at this point that the trick of using the quotient
of the derivatives of the two functions might again prove helpful. Since
the derivatives cannot, strictly speaking, be evaluated at the "point"
$+\infty$ , we use instead the limit as $x \to +\infty$ of the quotient of deriva-
tives to obtain

$$\lim_{x \to +\infty} \frac{\ln x}{x - 1} = \lim_{x \to +\infty} \frac{1/x}{1} = 0/1 = 0.$$ 

This not only tells us how the quotient function behaves for large $x$, but
in addition suggests that l'Hospital's rule might be useful in the eval-
uation of quotients which assume the form $\infty/\infty$ as $x \to +\infty$.

Finally, one other little problem can be presented which
relates to these graphs. One can ask : What is the point of intersection
of the quotient function $[\ln x]/[x - 1]$ with the function $f(x) = \ln x$ ?
The answer to this question follows, and knowledge of the answer will
increase the accuracy of the graphs.

We seek a value for $x$ such that

$$\frac{\ln x}{x - 1} = \ln x,$$

or

$$x - 1 = \frac{\ln x}{\ln x} = 1.$$ 

The solution is $x = 2$.

Consequently the graphs of $\frac{\ln x}{x - 1}$ and $\ln x$ intersect at the point
$(2, \ln 2)$. 

IV. THE CHAIN RULE

This chapter is devoted to pointing out a relationship between two general formulas from the differential calculus and two techniques for graphing functions. It is felt that appropriate graphs can be used to give greater meaning to the differentiation formulas, and can even be used to help the student "discover" the formulas for himself. The formulas to which reference is made are the formulas for the derivative of the sum of two functions and for the derivative of the composition of two functions (the chain rule). We will interpret these formulas in relation to the graphs of the sum of two functions and the composition of two functions. Using the geometric interpretation of the derivative as the slope of a curve, it will be possible to provide improved motivation for the formulas.

Relating the formula for the derivative of the sum of two functions to the graph of the sum function is very easily accomplished, yet this relationship is seldom used today. A prerequisite for understanding the relationship is a knowledge of the addition of ordinates method for graphing the sum of two functions. Modern textbooks stress the concept of function and often include material on the formation of new functions from given functions, so it is likely that the student will be familiar with the addition of ordinates technique.

To interpret the chain rule, we will employ a technique for graphing the function which is the composition of two given functions.
Techniques for graphing composite functions have been known for some time, but to the writer's knowledge all of these construct the graph as a curve in three-space. Included in a recent article in The Mathematics Teacher is a technique which gives the graph of the composite function as a curve in the plane, and this is particularly well suited to our purpose here.

We first discuss the geometric techniques for constructing the graphs that we need, and follow this with the interpretation of the differentiation formulas in connection with these graphs.

Let \( f(x) \) and \( g(x) \) be real valued functions of the real variable \( x \), and let \( D_f \) and \( D_g \) denote the respective domains of these functions. One can define a new function \( (f+g)(x) \), called the sum of these two functions, having domain \( D_{f+g} \), in the following way:

\[
(f+g)(x) = f(x) + g(x), \quad \text{where} \quad D_{f+g} = D_f \cap D_g.
\]

Thus the function \( (f+g)(x) \) maps the real number \( x \) into the unique real number \( f(x) + g(x) \).

Given the graphs of \( f(x) \) and \( g(x) \) in the Cartesian plane, it is a simple matter to construct the graph of \( (f+g)(x) \). Reference is now made to Figure 5. Given a point \( x_0 \) in \( D_{f+g} \) we seek the point \( P(x_0,(f+g)(x_0)) \), where \( (f+g)(x_0) \) is the algebraic sum of the two

---

16 Methods for representing composite functions as curves in three dimensions can be found in the following references:


ordinates \( f(x_0) \) and \( g(x_0) \). This point can be located in the plane in the following way. On the line \( x = x_0 \) we set a compass to measure \( |g(x_0)| \). Using this radius, we put one point of the compass on the point \( R(x_0, f(x_0)) \) and strike the line \( x = x_0 \) above \( R \) if \( g(x_0) > 0 \) and below \( R \) if \( g(x_0) < 0 \). The point so determined will be \( P(x_0, (f+g)(x_0)) \). If \( g(x_0) = 0 \), then \( P \) coincides with the point \( R(x_0, f(x_0)) \). As many points as desired on \( (f+g)(x) \) can be determined in this way, by geometrically adding ordinates. For the sake of the student, it is important to stress the fact that the ordinate to any point on \( (f+g)(x) \) is the algebraic sum of the ordinates to \( f(x) \) and to \( g(x) \).

![Figure 5](image-url)
We next turn our attention to the function which is the composition of two given functions. Suppose \( y = f(u) \) and \( u = g(x) \), where each of \( y, u, \) and \( x \) is a real variable. We can define a new function \( y = F(x) \) called the composition of \( f \) with \( g \) in this way:

\[
y = F(x) = f[g(x)], \text{ where } D_F =
\]

Some authors define the composition of two functions without introducing an intermediate variable \( u \). The notation suggested here which employs a variable \( u \) that is at the same time a dependent and an independent variable has been adopted advisedly. From experience in the classroom, it appears that this notation is the one best suited for later reconciliation with the statement of the chain rule as found in a typical calculus textbook.

Haddock and Hight have recently published an article which includes a method for constructing the graph in the plane of the composite of two given functions.\(^{17}\) It is this method that will be used to give pictorial meaning to the chain rule. Notice that a straight edge only is required to obtain the desired graph. The procedure outlined below is essentially that of Haddock and Hight, but the notation has been slightly modified.

Suppose we are given two functions, \( y = f(u) \) and \( u = g(x) \), and seek a pictorial representation of \( y = F(x) = f[g(x)] \) (see Figure 6). It is convenient to think of both the horizontal and the vertical axis of reals as playing a double role. Thus we graph \( y = f(u) \), thinking

\(^{17}\)Haddock and Hight, op. cit., p. 4.
of the vertical axis as the $y$-axis and the horizontal axis as the $u$-axis, and then graph $u = g(x)$ with the vertical axis as the $u$-axis and the $x$-axis playing its customary role. The line $y = x$ is important in this technique, and it is shown in Figure 6. Now, given a point $x = x_0$ on the $x$-axis, $x_0 \in D_p$, we wish to locate the point $P(x_0, f[g(x_0)])$. This is accomplished as follows. From the point $Q(x_0, g(x_0))$ we draw a horizontal line to intersect the graph of $y = x$ at the point $R(g(x_0), g(x_0))$. A vertical line through $R$ will intersect $y = f(u)$ in the point $S(g(x_0), f[g(x_0)])$, and a horizontal line through $S$ will intersect the line $x = x_0$ in the desired point $P(x_0, f[g(x_0)])$ which is on the graph of $F(x) = f[g(x)]$. The arrows in Figure 6 indicate the path to be followed. In a sense, the graph of $g(x)$ and the line $y = x$ serve to map real numbers on the horizontal axis back into that axis. The procedure carries $x_0$ to the curve $g(x)$, to the line $y = x$, and then back to the horizontal axis at the point $u_0 = g(x_0)$. $u_0$ is then carried to the point $y_0 = f(u_0)$ on the vertical axis, and the last horizontal line marks off this ordinate on the line $x = x_0$ at the point $P$. 
Figure 6
The writer recently tried this technique with a class of first quarter calculus students, and no great difficulty was encountered. Using coordinate paper, these students were able to graph with very good accuracy the function \( y = \cos x^2 \) from the graphs of \( y = \cos u \) and \( u = x^2 \). This procedure is recommended by the fact that the student gains insight into the nature of composite functions, and by the fact that he sees \( y = F(x) = f[g(x)] \) as a new function, distinct from but related to the two given functions.

Now the theorem from the differential calculus which we intend to relate to the addition of ordinates method of graphing the sum of two functions is this:

**Theorem.** If \( f \) and \( g \) are differentiable functions of \( x \), then

\[
\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx}.
\]

That is, the derivative of the sum of two differentiable functions of \( x \) is the sum of the two derivatives.

The usual proof runs as follows:

Let us work at the point \( x \) in \( D_{f+g} \). It is assumed that both \( f \) and \( g \) possess derivatives at \( x \). Then

1. \( (f+g)(x) = f(x) + g(x) \),
2. \( (f+g)(x+\Delta x) = f(x+\Delta x) + g(x+\Delta x) \),
3. \( \Delta(f+g) = (f+g)(x+\Delta x) - (f+g)(x) = f(x+\Delta x) - f(x) + g(x+\Delta x) - g(x) \), or
4. \( \Delta(f+g) = \Delta f + \Delta g \). Then
5. \( \frac{\Delta(f+g)}{\Delta x} = \frac{\Delta f}{\Delta x} + \frac{\Delta g}{\Delta x} \).
Since the limit of the sum of two functions is the sum of their limits, it follows that

\[
\lim_{\Delta x \to 0} \frac{\Delta (f+g)}{\Delta x} = \lim_{\Delta x \to 0} \left( \frac{\Delta f}{\Delta x} + \frac{\Delta g}{\Delta x} \right) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} + \lim_{\Delta x \to 0} \frac{\Delta g}{\Delta x}.
\]

Thus,

\[
\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx}.
\]

This is by no means a difficult theorem for a mature mathematician. The proof is a perfectly straightforward application of the definition of the derivative to the function (f+g) and of the definition of the function (f+g). The beginning student is usually quite willing to accept this theorem as giving a reasonable kind of result, and he goes happily along differentiating polynomials, for example, term by term and obtaining correct derivatives. This writer conjectures, however, that the average student does not get much out of the proof. Since the theorem comes quite early in the first course in the calculus, he has had but limited experience with the formation of derivatives and with the \( \Delta \) symbol. Thus he may miss completely the point that this theorem is a statement about a point-wise property of the sum function (f+g), namely that (f+g) possesses a derivative at a point \( x_0 \) whenever \( f'(x_0) \) and \( g'(x_0) \) exist, and that \( (f+g)'(x_0) = f'(x_0) + g'(x_0) \).

By employing the geometric interpretation of the derivative as the slope of a curve and using graphs of the functions \( f, g, \) and \( f+g \), each of the steps (1) to (4) in the proof can be given pictorial meaning. To the writer's knowledge, the association of this theorem with
the graph of the sum function \((f+g)\) does not appear in any textbook on
the elementary calculus. This association can be so easily established
as to make it possible for the student to discover the theorem itself
along with a proof. The pictures can be used, in the manner indicated
below, to establish the identity which is basic to the proof, namely the
identity \(\Delta(f+g) = \Delta f + \Delta g\).

The basic clue that serves to give direction to the investi-
gation, both here and in the case of the chain rule, is this: that the
derivative of the new function is expressible in terms of the deriva-
tives of the functions of which it is made up. Thus in the present case
we seek an expression for \(\frac{d(f+g)}{dx}\) in terms of some simple combination
of \(\frac{df}{dx}\) and \(\frac{dg}{dx}\). Furthermore, we work at a point \(x_o\), say, in \(D_{f+g}\).

We now present a sequence of ideas, relating to Figure 7,
which lead to the theorem we are concerned with, and which constitute
a proof of the theorem. It is understood that the student is well
versed in the addition of ordinates method of graphing the sum function,
and that he fully understands that at any point \(x\) the ordinate to the
curve \((f+g)\) is the algebraic sum of the ordinates to \(f\) and to \(g\) at \(x\).

We give \(x_o\) an increment \(\Delta x\), and this induces a change in \(f\), \(g\),
and \((f+g)\). Now
\[
\Delta g = g(x_o + \Delta x) - g(x_o) \quad \text{and} \quad \Delta f = f(x_o + \Delta x) - f(x_o).
\]

These increments are labelled in Figure 7. Also
\[
(i) \begin{cases}
(f+g)(x_o) = f(x_o) + g(x_o) \\
(f+g)(x_o + \Delta x) = f(x_o + \Delta x) + g(x_o + \Delta x).
\end{cases}
\]
The equations (i) involve nothing more than the fact that the ordinates to the curve \((f+g)\) corresponding to the points \(x_o\) and \(x_o+\Delta x\) are the sum of the ordinates to \(f\) and \(g\) corresponding to these points. When visualized in the figure, these equations seem much less ominous than when they are looked at as just a collection of marks.

To obtain the derivative of \((f+g)\) at \(x_o\) we need an expression for \(\Delta(f+g)\), and the same reasoning as is used to form \(\Delta f\) and \(\Delta g\) leads to the equation

\[
\Delta(f+g) = (f+g)(x_o+\Delta x) - (f+g)(x_o).
\]

Note that \(\Delta(f+g)\) is also labelled in Figure 7. Next we seek a relationship between \(\Delta(f+g)\), \(\Delta f\), and \(\Delta g\). The desired relationship becomes clear when we examine the expression for \(\Delta(f+g)\) and the equations (i) above:

\[
\Delta(f+g) = (f+g)(x_o+\Delta x) - (f+g)(x_o)
= f(x_o+\Delta x) + g(x_o+\Delta x) - f(x_o) - g(x_o)
= f(x_o+\Delta x) - f(x_o) + g(x_o+\Delta x) - g(x_o).
\]

Thus

\[
(ii) \quad \Delta(f+g) = \Delta f + \Delta g.
\]

This is the key identity which we wished to establish — that the change in \((f+g)\) corresponding to any increment \(\Delta x \neq 0\) is the sum of the change in \(f\) and the change in \(g\). In Figure 7 \(\Delta g > 0\), \(\Delta f < 0\), but \(\Delta g > |\Delta f|\) and consequently \(\Delta(f+g) = \Delta f + \Delta g > 0\).

One can also bring home the meaning of the identity (ii) by asking the following questions. Suppose \(f(x)\) was a constant function between \(x_o\) and \(x_o+\Delta x\) (assume \(\Delta x > 0\)), and suppose \(g(x)\) increased on this interval by an amount \(\Delta g\). What then would be the behavior of \((f+g)\)?
Clearly \((f+g)\) would increase by the amount \(\Delta g\). Suppose \(g(x)\) was constant, and corresponding to some \(\Delta x > 0\) \(f(x)\) decreased by an amount \(\Delta f\). What would happen to \((f+g)\)? It is clear from our graphic interpretation that \((f+g)\) would drop by the same amount, \(\Delta f\). Now suppose \(g(x)\) increased a bit and \(f(x)\) decreased a bit - what would happen to \((f+g)\)? Well, \((f+g)\) would change by an amount corresponding to the algebraic sum of these increments. Familiarity with the construction of the sum function \((f+g)\) should facilitate understanding of these ideas.

The remaining steps should cause no trouble, the direction to take being established by the definition of a derivative. Both sides of the identity (ii) are divided by \(\Delta x\) to give

\[
\frac{\Delta (f+g)}{\Delta x} = \frac{\Delta f}{\Delta x} + \frac{\Delta g}{\Delta x}.
\]

One must then appeal to the theorem which states that the limit of a sum of two functions is the sum of the two limits to obtain:

\[
\lim_{\Delta x \to 0} \frac{\Delta (f+g)}{\Delta x} = \lim_{\Delta x \to 0} \left( \frac{\Delta f}{\Delta x} + \frac{\Delta g}{\Delta x} \right) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} + \lim_{\Delta x \to 0} \frac{\Delta g}{\Delta x},
\]

or finally

\[
\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx}.
\]
Figure 7
The Chain Rule.

Given functions \( y = f(u) \) and \( u = g(x) \) we can construct a new function \( y = F(x) \) by the operation of composition:

\[ y = F(x) = f[g(x)]. \]

Now the derivative of this new function, \( F'(x) \), can be computed in terms of the derivatives of its component functions according to the following rule.

Theorem (The Chain Rule).

Let \( y = f(u) \) be a differentiable function of \( u \), and \( u = g(x) \) be a differentiable function of \( x \). Then the composite function \( y = F(x) = f[g(x)] \) is a differentiable function of \( x \), and the derivative of \( y \) with respect to \( x \) is the product of the derivative of \( y \) with respect to \( u \) and the derivative of \( u \) with respect to \( x \); i.e.,

\[ F'(x) = f'(u) \cdot g'(x) \]

or

\[ \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}. \]

The proof of this theorem is complicated by the fact that for certain functions \( u = g(x) \) and certain points \( x \) in \( D_u \), when we give \( x \) an increment \( \Delta x \neq 0 \) the corresponding increment in \( u \), \( \Delta u \), may be zero. We therefore make the assumption that for each \( x \) in \( D_u \) there corresponds a number \( \sigma > 0 \) such that for all \( \Delta x \) satisfying \( |\Delta x| < \sigma \) the corresponding values of \( \Delta u \) are never equal to zero.\(^{18}\) Under this assumption

\(^{18}\) If it is the case that there are arbitrarily small values of \( \Delta x \) distinct from zero for which \( \Delta u = 0 \), then the statement of the theorem is still correct, but the proof is invalid.
assumption we can write the identity
\[ \frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x} , \quad |\Delta x| < \sigma . \]

Taking limits, and noting that when \( \Delta x \to 0 \), then \( \Delta u \to 0 \), we have:
\[
\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \left( \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x} \right) \\
= \left( \lim_{\Delta u \to 0} \frac{\Delta y}{\Delta u} \right) \cdot \left( \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \right),
\]
or
\[ \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} . \]

David Gans has written a perceptive article in the American Mathematical Monthly which bears directly on the matter of assuming that \( \Delta u \neq 0 \) when \( \Delta x \neq 0 \).\(^\text{19}\) He classifies textbook proofs of the chain rule into three categories. A category A proof consists of writing the identity \( \frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x} \), taking limits to obtain the result, and no mention is made of the complicating fact that \( \Delta u \) may be zero when \( \Delta x \) is not zero. A category B proof is like the above, but includes a brief remark that \( \Delta u \) must not be zero. Gans says: "...this proof [category B] can be regarded as an improvement over the preceding [category A] only if it is granted that it is better to leave a student mystified by a difficulty than ignorant of it."\(^\text{20}\) A category C proof is rigorous and complete and uses an \( \varepsilon \)-procedure of some sort. Of category C proofs


\(^{20}\)Ibid., p. 115.
Gans says this: "However, it is hard to see how these comparatively sophisticated proofs can be understood by any but a very small minority of students when it is realized that the latter are usually in contact with the calculus for only about a month when the chain rule is presented to them."

The last paragraph of this article contains Gan's recommendations for giving a proof that is both rigorous and understandable. He says:

If \( x_1 \) is the value of \( x \) under consideration, and \( u_1 \) the corresponding value of \( u \), all that is now necessary [for a rigorous, understandable proof] beyond what is presented in textbooks is to agree to consider only functions \( u(x) \) for which \( \Delta u \) is not zero if \( x \) is sufficiently close to \( x_1 \) but unequal to it. That is, we state and prove the chain rule only for such functions. Thus we only exclude functions taking on the value \( u_1 \) infinitely many times in the neighborhood of \( x_1 \). Except for the trivial case when \( u(x) \) is constant in this neighborhood, such excluded functions never occur in first year work, and but rarely in the following years. The student should be told that the chain rule also applies to these functions but that the proof of this fact is best delayed until such time as he works with them.\(^{21}\)

Thus Gans advocates a proof in category D, say, which retains the simplicity of a category A proof by excluding certain unusual functions, and which avoids the possibly excessive rigor of a category C proof. A category D proof would include an expanded explanation of the difficulty with \( \Delta u \) being zero. It would include examples of

\(^{21}\)Ibid., p. 116.
elementary functions \( u(x) \) (\( \sin x \), for example) where \( \Delta u \) can be zero when \( \Delta x \) is not zero, but where \( \Delta u \) cannot be zero if \( \Delta x \) is sufficiently small.

It is felt that the chain rule as stated above is consonant with Gan's recommendations.

By making a slight sacrifice in the generality of the theorem, we obtain a brief and simple proof. But brevity of exposition does not assure understanding on the part of the student. Experience with teaching this theorem in the classroom shows that it is a difficult result for the student to apply. Often the symbols involved are largely devoid of meaning, and the student learns to perform differentiations in a mechanical sort of way. As teachers and textbook writers, we make almost no effort to have the chain rule appear reasonable to the learner. This may be because we don't know how to make it appear reasonable, or because it is deemed unnecessary to do so.

The writer suggest that the chain rule can be made plausible by employing the geometric technique for graphing composite functions which was outlined above. We consider the following sequence of ideas, and refer to Figure 8 for pictorial assistance. The language used refers to graphs as shown in Figure 8. For a different orientation the words up, down, left, and right would be modified to conform to the arrangement of the graphs. It is hoped that by following this sequence of ideas the student will benefit in two ways: the symbols he is working with will have increased meaning, and the result itself will appear plausible.
Figure 8 shows first of all the graphs of three functions: 
\[ u = g(x), \quad y = f(u), \quad \text{and} \quad y = F(x). \]
We seek an expression for \( F'(x) \) in terms of the derivatives \( f'(u) \) and \( g'(x) \) of its component functions.

We work at a point \( x_0 \) in \( D_F \), and thereby emphasize the fact that the chain rule is a statement about a point-wise property of \( F(x) \), a fact that is easily lost sight of by the student.

A change in \( x \) by the amount \( \Delta x \neq 0 \) induces a chain of related changes. It generates a change in \( u \), shown in Figure 8. If 
\[ u_0 = g(x_0), \]
and 
\[ u_0 + \Delta u = g(x_0 + \Delta x), \]
then 
\[ \Delta u = g(x_0 + \Delta x) - g(x_0). \]

In view of the way the geometric technique for graphing works, 
\( x_0 \) maps up to \( g(x) \), across to the line \( y = x \) where it is reflected toward the point \( u_0 = g(x_0) \) on the \( x \)-axis. Similarly, \( x_0 + \Delta x \) maps up, across, and down to the point \( u_0 + \Delta u = g(x_0 + \Delta x) \). One can imagine the line segment above \( x_0 + \Delta x \) which represents \( \Delta u \) as being carried left to the line \( y = x \), rotated \( 90^\circ \) clockwise, and subsequently appearing as an interval on the horizontal axis. Now, referring to the graph of \( y = f(u) \) we can visualize the change in \( y \) which is generated by this \( \Delta u \). \( u_0 \) maps up and across into \( y_0 = f(u_0) \) and \( u_0 + \Delta u \) maps into \( y_0 + \Delta y = f(u_0 + \Delta u) \).

Thus \( \Delta u \) has brought about a change \( \Delta y = f(u_0 + \Delta u) - f(u_0) \) in \( y \).

It is apparent in Figure 8 that the increment \( \Delta x \) has also occasioned a change \( \Delta y \) in the composite function \( y = F(x) \). Now what justification is there for writing the same symbol, \( \Delta y \), for the change \( \Delta f \) in \( f \) and also for the change \( \Delta F \) in \( F \)? Well, we can see in two ways
that $\Delta f = \Delta F$. Graphically, the two horizontal lines through $y_0$ and $y_0 + \Delta y$ on the vertical axis which measure $\Delta f$ proceed parallel to each other to the right to also measure $\Delta F$. Also, using analytic symbols, we write:

$$\Delta y = f(u + \Delta u) - f(u)$$

$$= f[g(x_0 + \Delta x)] - f[g(x_0)].$$

But from the definition of $F(x)$,

$$\Delta F = f[g(x_0 + \Delta x)] - f[g(x_0)],$$

and equality obtains.

With the aid of the figure we can visualize three difference quotients, $\frac{\Delta u}{\Delta x}$, $\frac{\Delta y}{\Delta u}$, and $\frac{\Delta F}{\Delta x} = \frac{\Delta y}{\Delta x}$. Working with the hint that $\frac{\Delta y}{\Delta x}$ can be expressed as some simple algebraic combination of $\frac{\Delta y}{\Delta u}$ and $\frac{\Delta u}{\Delta x}$, the student should experience little difficulty in seeing that it is the product of these two which gives the desired identity:

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x}.$$

The approach here is different from, in fact the reverse of, the usual one. Usually upon starting the proof one writes

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x},$$

and the student is encouraged to think that the right member here is obtained by a decomposition of the left member, after multiplying the left member by one renamed $\Delta u/\Delta u$, $\Delta u \neq 0$. The graphical approach is the reverse in the sense that we first give meaning to the symbols $\frac{\Delta y}{\Delta x}$, $\frac{\Delta y}{\Delta u}$, and $\frac{\Delta u}{\Delta x}$, and then ask how they are related. Thus we compose, under multiplication, $\frac{\Delta y}{\Delta u}$ and $\frac{\Delta u}{\Delta x}$ to obtain $\frac{\Delta y}{\Delta x}$.
From the assumption that \( f \) and \( g \) are differentiable functions, it follows that \( \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} = g'(x_o) \) and \( \lim_{\Delta u \to 0} \frac{\Delta y}{\Delta u} = f'(u_o) \). The figure also allows us to see that \( \Delta u \to 0 \) as \( \Delta x \to 0 \), for the change in the function \( u = g(x) \) is apparently small for a small change in \( x \). Taking limits on both sides of the identity

\[
\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x}
\]

gives the result

\[
F'(x_o) = f'(u_o) \cdot g'(x_o).
\]

Since the point \( x_o \) was an arbitrary point, we can drop the subscript \( o \) and obtain the chain rule in this form:

\[
F'(x) = f'(u) \cdot g'(x),
\]

or

\[
\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.
\]

The utility of a pictorial representation similar to Figure 8 does not end here, however. It can be further used to explain the difficulty with \( \Delta u \) being zero when \( \Delta x \) is not zero. For example, if \( u = g(x) = k \) is a constant function then the graph of \( g(x) \) becomes the horizontal line \( u = k \), and \( \Delta u = g(x_o + \Delta x) - g(x_o) = k - k = 0 \), for all \( \Delta x \neq 0 \). Consequently, \( g'(x_o) = 0 \). Under these conditions no change is induced in \( y = F(x) = f[g(x)] \) by a change \( \Delta x \) in \( x_o \), since

\[
\Delta F = F(x_o + \Delta x) - F(x_o)
\]
\[
= f[g(x_o + \Delta x)] - f[g(x_o)]
\]
\[
= f[k] - f[k]
\]
\[
= 0.
\]
But if $\frac{\Delta F}{\Delta x} = 0$ for all $\Delta x \neq 0$, then $F'(x_0) = 0$. Since $f'(u_0)$ is assumed to exist, we have

$$F'(x_0) = 0 = f'(u_0) \cdot 0$$

and the chain rule appears to hold true in this case.

Now suppose that $u = g(x)$ is not a constant function, but does drop back toward the horizontal axis (thinking of $g(x)$ as pictured in Figure 8) and assumes again the value $u_0 = g(x_0)$ at some point $x_0 + \Delta x'$, $\Delta x' \neq 0$. Then $\Delta u = g(x_0 + \Delta x') - g(x_0) = u_0 - u_0 = 0$. Such behavior is by no means unusual for functions, and the student can see clearly how it can happen that $\Delta u = 0$ when $\Delta x \neq 0$. Now it can be pointed out that although this can happen, all but a very few of the non-constant functions which are dealt with in elementary calculus are such that when $|\Delta x|$ is small $\Delta u$ will be different from zero. For it to be otherwise, $u = g(x)$ would have to assume the value $u_0$ infinitely many times in every neighborhood of $x_0$, and a function which does this is not often encountered in the beginning calculus course.
V. CAUCHY'S FORMULA

There is a type of proof not uncommon in mathematical exposition which is characterized by the use of an artificially introduced auxiliary function. The use of such a function is justified by the fact that it serves to establish the desired result as a consequence of a previously deduced theorem. Auxiliary functions of this kind are used in the calculus, and in higher mathematics. Too often no attempt is made to motivate the use of the auxiliary function by relating it to the theorem being considered. As a consequence of this the student is mystified, and may experience great difficulty when asked to reproduce the proof because he cannot remember the particular form of the auxiliary function. Furthermore, unless it is pointed out how one might be led to consider just this function, the student cannot be expected to reproduce the form by logical thinking. He is justifiably curious about the origin of the function.

George Polya has written to this point in an article entitled "With, or Without, Motivation?" 22 In the article he gives two proofs of a certain inequality. As one might anticipate from the title, one proof is well motivated in the sense that most steps in the deduction appear natural to the reader. This proof is quite long. The other

---

proof is very brief and employs an auxiliary function the use of which
is justified only by the fact that it works. Polya clearly implies that
of the two the proof with the heuristic approach is the better, it
being more psychologically satisfying for the reader. He realizes, how­
ever, that a demonstration may become much too long if one attempts to
supply plausibility arguments for each step, and points out that such
arguments can best be given verbally rather than in writing.

Polya uses the colorful phrase "deus ex machina" to describe
the introduction into a proof of a contrived auxiliary function. The
term "deus ex machina," meaning a god from a machine, traces back to
early Greek theatre. When the characters in a play found themselves in
apparently insoluble difficulties, a god would be introduced onto the
stage, quite literally brought on by a machine of some sort, in order
to extricate the players from their difficulties. From this early use,
the term has come to describe any person or thing that is introduced
suddenly or unexpectedly and that provides an artificial or contrived
solution to a problem. The aptness of the term as used in the article
is apparent.

Polya's point of view is expressed in these words:
A step of a mathematical argument is appropriate if it is
essentially connected with the purpose, if it brings us nearer
to the goal. It is not enough, however, that a step is
appropriate: it should appear so to the reader. If the step
is simple, just a trivial, routine step, the reader can
easily imagine how it could be connected with the aim of
the argument. If the order of presentation is very carefully
planned, the context may suggest the connection of the step
with the aim. If, however, the step is visibly important, but
its connection with the aim is not visible at all, it appears as a "deus ex machina" and the intelligent reader or listener is understandably disappointed.23

Polya concludes the article with these remarks which stress the importance of intuition and inductive thinking in mathematics:

I cannot omit a final remark on logic. Some authors distinguish two branches of logic, deductive logic and inductive logic. Yet these two branches differ widely. Deductive logic is a firmly established branch of science, and became in its latest development, as symbolic logic, practically a branch of mathematics. Inductive logic is an interesting subject of philosophical discussion, but can scarcely be regarded as an established science. Deductive logic is concerned with the validity of proofs. Inductive logic which I would prefer to call heuristic logic, in order to emphasize its wider scope, is concerned with plausible inference only. That deductive logic is closely connected with mathematics, is widely recognized; some modern authors think that its proper object is the analysis of the deductive structure of mathematical theories. Now I come to my point: I think that also heuristic logic is closely connected with mathematics, but not with mathematical theories and their deductive structures, rather with mathematical problems and the invention of their solution.24

23Ibid., p. 685.

24Ibid., pp. 690, 691.
The deduction of a certain theorem known as Cauchy's formula furnishes a striking example of the use of inadequately motivated auxiliary functions. A cleverly conceived function is pulled out of the hat, so to speak, and the desired result follows in just a few lines. Most calculus textbooks include little or no justification for the use of the auxiliary function. Two distinct types of auxiliary functions are commonly used in deducing Cauchy's formula. One of these resembles the equation of a line, and the other is a function defined by a determinant. In each case a perfectly reasonable explanation for the use of the function is possible. To accomplish this explanation in geometrical terms is the purpose of this chapter.

It is apparently correct to attribute the theorem that we are about to examine to Cauchy. The theorem appears in an Addition to his work Résumé des Lecons donees a L'Ecole Royale Polytechnique sur Le Calcul Infinitésimal published in 1823. The words used by Cauchy to describe the result lead one to believe that it is his own work. Cauchy's proof of Cauchy's formula is lengthy and entirely analytical, being accomplished without the assistance of a single diagram. The deduction of this result has been greatly shortened and improved over the years. The formula under consideration is also sometimes called The Extended Mean Value Theorem, or The Extended Law of the Mean.

---

Let us turn now to the theorem itself as it appears in a modern textbook. To illustrate both types of auxiliary functions, two statements of the theorem are included. The first is from Smail's *Analytic Geometry and Calculus*, the second from Goodman's book *Analytic Geometry and the Calculus*. Some of the notation has been changed slightly in the quotation from Goodman's book in order that our presentation here have consistent symbolism.

Smail states and proves the theorem in this way:

If $f(x)$ and $g(x)$ are continuous in the interval $a < x < b$, and if the derivatives $f'(x)$ and $g'(x)$ exist at every point of the interval $a < x < b$, and if $g'(x)$ is not zero at any point within the interval $(a,b)$, except possibly at $a$ or $b$, then there is at least one value $\xi$ of $x$ between $a$ and $b$ such that

\begin{equation}
\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}, \quad (a < \xi < b).
\end{equation}

This theorem may be proved in a manner similar to that used in §150 for the previous mean value theorem. We construct the auxiliary function:

(a) \quad F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{g(b) - g(a)} [g(x) - g(a)].

---


Since \( g'(x) \neq 0 \) at all points within the interval \((a, b)\), it follows from Rolle's theorem (§149) that \( g(a) \neq g(b) \). From the form of \( F(x) \) and the hypotheses on \( f(x) \) and \( g(x) \), it follows that \( F(x) \) is continuous in the interval \( a < x < b \), and it has a derivative at all points where \( a < x < b \). Evidently \( F(a) = F(b) = 0 \). We have

\[
\text{(b) } F'(x) = f'(x) - \frac{f(b) - f(a)}{g(b) - g(a)} g'(x) .
\]

By Rolle's theorem, \( F'(x) \) must vanish for some value \( x = \xi \) between \( a \) and \( b \). Hence,

\[
\text{(c) } F'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{g(b) - g(a)} g'(\xi) = 0 ,
\]

from which formula (1) follows at once since \( g'(\xi) \neq 0 \).

It should be noted that the mean value theorem of §150 is a special case of this theorem, for the case when \( g(x) = x \).

Goodman's statement of Cauchy's formula follows:

**Theorem 1. (The Generalized Mean Value Theorem).**

Let \( f(t) \) and \( g(t) \) be two functions, each continuous in the closed interval \( a \leq t \leq b \), and each differentiable in the open interval \( a < t < b \). Suppose further that \( g(b) \neq g(a) \), and that the derivatives \( f'(t) \) and \( g'(t) \) do not vanish simultaneously in \( a < t < b \). Then there is some point \( \xi \), with \( a < \xi < b \), such that

\[
\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)} .\tag{3}
\]

**Proof of Theorem 1.** For the reader who is familiar with
determinants let $F(t)$ be defined by the determinant

$$F(t) = \begin{vmatrix} g(t) & f(t) & 1 \\ g(a) & f(a) & 1 \\ g(b) & f(b) & 1 \end{vmatrix} \tag{4}$$

When $t = a$ the first and second rows are identical and hence $F(a) = 0$. When $t = b$ the first and third rows are identical and so $F(b) = 0$. It is easy to see that $F(t)$ satisfies the conditions of Rolle's theorem. Hence there is some suitable $\xi$ for which $F'(\xi) = 0$. Expanding the determinant (4) by minors of the first row gives

$$F(t) = g(t)[f(a) - f(b)] - f(t)[g(a) - g(b)] + g(a)f(b) - f(a)g(b). \tag{5}$$

For the reader who is not familiar with determinants, the proof can start with the function $F(t)$ defined by equation (5). A brief computation from (5) will show that $F(a) = F(b) = 0$, and hence Rolle's theorem can be applied to $F(t)$.

If we differentiate $F(t)$, we have from (5) that

$$F'(t) = g'(t)[f(a) - f(b)] - f'(t)[g(a) - g(b)]. \tag{6}$$

By Rolle's theorem there is a $\xi$ such that $F'(\xi) = 0$. For this $\xi$ equation (6) gives

$$0 = g'(\xi)[f(a) - f(b)] - f'(\xi)[g(a) - g(b)]. \tag{7}$$

From (7), simple algebraic manipulations give equation (3). However in doing the divisions involved we must be certain that
we do not divide by zero. But this is assured by the hypotheses that \( f(a) \neq f(b) \) and that \( f'(t) \) and \( g'(t) \) are not simultaneously zero.

Q.E.D.

(For another proof of Cauchy's formula that makes use of a function defined by a determinant the reader should see Widder's Advanced Calculus.\(^{28}\))

In each of these proofs, a "deus ex machina" is introduced. In the first case the god has the name

\[
F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{g(b) - g(a)} [g(x) - g(a)],
\]

and in the second case the name

\[
F(t) = \begin{vmatrix} g(t) & f(t) & 1 \\ g(a) & f(a) & 1 \\ g(b) & f(b) & 1 \end{vmatrix}.
\]

Now each of these expressions, as well as the basic equation of Cauchy's formula,

\[
\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'\left(\xi\right)},
\]

has a pleasant geometric interpretation. We shall obtain these interpretations through a discussion of the familiar mean value theorem. Cauchy's formula is an extension or generalization of the mean value theorem, and thus it is not surprising that a study of the latter should shed light on the former.

Here is a statement of the mean value theorem.

Mean Value Theorem.

Let \( f(x) \) be a function which is continuous at each point \( x \) of the interval \( a \leq x \leq b \), and let it have a derivative at each point \( x \) satisfying \( a < x < b \). Then there is a value \( x = \xi \), \( a < \xi < b \), such that

\[
(i) \quad \frac{f(b) - f(a)}{b - a} = f'(\xi).
\]

Most textbooks point out the geometric interpretation of the equation (i). The left member of (i) is the expression for the slope of the secant line AB (see Figure 9). \( f'(x) \) gives the slope of the tangent to the curve at the point \( P(x,f(x)) \). The mean value theorem therefore asserts that if the arc from \( A \) to \( B \) is smooth, then there is some intermediate point \( T \) on the arc at which the tangent is parallel to the secant line joining \( A \) and \( B \).
A typical proof proceeds as follows. We employ the auxiliary function

\[ F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a} (x - a). \]

Now \( F(x) \) satisfies all the conditions of Rolle's theorem; i.e., \( F(x) \) is continuous on the closed interval \([a,b]\), differentiable on the open interval \((a,b)\), and in addition \( F(a) = F(b) = 0 \). Thus we are assured of the existence of a real value for \( x, x = \xi \), such that \( F'(\xi) = 0 \). An easy computation of \( F'(x) \) gives the desired result.

Fortunately for the student, it is usual today to also point out the geometrical significance of the function \( F(x) \). Since one can easily show that

\[ y = f(a) + \frac{f(b) - f(a)}{b - a} (x - a) \]
is the equation of the chord joining A and B, \( F(x) \) represents, except for sign, the length of the vertical line segment RP (Figure 9) between chord and curve.

The textbook author usually drops the matter of an interpretation of \( F(x) \) at this point. But one can make the result even more plausible by pursuing the matter a bit further. From the fact that \( F(x) \) satisfies the conditions of Rolle's theorem, we know that the value \( \xi \) of the mean value theorem corresponds to a maximum (or minimum) value, \( F(\xi) \), of the function \( F(x) \) on \([a,b]\). Now the question is, how can we convince the student that the tangent to the curve should be parallel to the chord precisely at the point \( \xi \) where the vertical displacement from chord to curve achieves a maximum? One answer will now be given.

Let the equation of the curve be \( y = f(x) \). Let \( P \) be the point \( (x,f(x)) \), and let the line segment \( PQ \) measure the perpendicular distance from curve to chord (Figure 10).

![Figure 10](image)
Clearly the length PQ changes with x, and is zero at a and at b. Incidentally, a function representing this perpendicular distance might serve as an auxiliary function in the proof of the mean value theorem; but as we shall see, this function would be only a constant multiple of the usual function F(x). As x changes from a to b the displacement PQ goes from zero to at least one maximum value, and back to zero. Reinforced with a knowledge of Rolle's theorem, the student will agree that the tangent to the curve will be parallel to the chord at a point T where PQ achieves a maximum value (one can imagine a change in the orientation of the axes of a sort to make the x-axis coincide with AB). It only remains to see why RP achieves a maximum precisely at the point x where PQ achieves a maximum.

Any value of x satisfying a < x < b will determine a triangle, PQR. For all such numbers x, the angle at Q is a right angle and the angle at R is $90^\circ - \alpha$, where $\alpha$ is the inclination of the chord AB. Thus the angle at P is $\alpha$, and

$$\cos \alpha = \frac{PQ}{RP},$$

or

$$RP = (\sec \alpha) PQ.$$  

This equation states the RP is proportional to PQ, with the constant of proportionality being $\sec \alpha$. It follows that RP achieves a maximum precisely where PQ does, and that the tangent will be parallel to the chord AB at a point x where RP is a maximum.
Still another function has been used to prove the mean value theorem. This is a function defined by a determinant. We let \( K(x) \) equal the area of the triangle \( ABP \) (see Figure 10 again). As the notation implies, \( K(x) \) is a function of \( x \) and we see immediately that \( K(x) \) is zero when \( x = a \) and when \( x = b \). As a matter of fact, \( K(x) \) will achieve a maximum at a value \( x \) for which \( PQ \) is a maximum, since \( PQ \) is the altitude of the triangle \( ABP \).

A well-known formula from analytic geometry states that

\[
2K(x) = \begin{vmatrix} x & f(x) & 1 \\ a & f(a) & 1 \\ b & f(b) & 1 \end{vmatrix}.
\]

This is where the determinant appears. This determinant gives us twice the area of the triangle whose vertices are \((x,f(x)), (a,f(a)), \) and \((b,f(b))\). The number we obtain for \( K(x) \) will be positive as long as the points are entered in the determinant in counterclockwise order. When we expand the determinant by cofactors of the first row, we obtain an equivalent expression for \( 2K(x) \):

\[
2K(x) = x[f(a) - f(b)] - f(x)[a - b] + [af(b) - bf(a)].
\]

---

29Proofs of the mean value theorem which involve determinants can be found in:


What a cumbersome auxiliary function. Imagine trying to reproduce this if you didn't know where it came from!

Now \( K(x) \) is continuous on the closed interval \([a, b]\), and differentiable on the open interval \((a, b)\). Furthermore, when \( x = a \) the first and second rows of the determinant are equal and when \( x = b \) the first and third rows are equal. Thus \( K(a) = K(b) = 0 \). This function then satisfies the conditions of Rolle's theorem, and we are assured of the existence of a value \( x = \xi, \ a < \xi < b \), such that \( K'(\xi) = 0 \).

We can compute the derivative of this function either from the determinant, to get

\[
2K'(x) = \begin{vmatrix}
1 & f'(x) & 0 \\
\ a & f(a) & 1 \\
\ b & f(b) & 1 \\
\end{vmatrix},
\]

or from the expanded form,

\[
2K'(x) = [f(a) - f(b)] - f'(x)[a - b].
\]

Substituting \( \xi \) for \( x \) and equating the derivative to zero gives

\[
\frac{f(b) - f(a)}{b - a} = f'(\xi).
\]

This brief study of the mean value theorem is sufficient to accomplish our goals. We are now prepared to develop geometric meaning for Cauchy's formula, and for the auxiliary functions used in the proof of its validity. We repeat the statement of the theorem, using \( t \) as the independent variable. If one is interested in using the ideas below to motivate the formula, the use of the letter \( t \) seems advisable, since \( t \) suggests parametric equations, and this is the key idea in our interpretation.
Cauchy's Formula.

Let $g(t)$ and $f(t)$ be functions with the following properties:

1. $g(t)$ and $f(t)$ are each continuous on the closed interval $a \leq t \leq b$,
2. $g'(t)$ and $f'(t)$ exist for every $t$ in the open interval $a < t < b$,
3. $g'(t) \neq 0$ for every $t$ in the open interval $a < t < b$.

Then there is a number $\xi$, with $a < \xi < b$, such that

\[
(ii) \quad \frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.
\]

The condition $g'(t) \neq 0$ is necessary on two counts. It assures us that the quotient $f'(t)/g'(t)$ has meaning for all $t$ in $(a,b)$. It also guarantees that $g(b) \neq g(a)$; for if $g(b) = g(a)$, then by Rolle's theorem $g'(t) = 0$ for some $t$ in $(a,b)$.

Geometric meaning can be attributed to formula (ii) by reasoning as follows.

We set $x = g(t)$ and $y = f(t)$, and consider these equations to be parametric equations of some curve $C$. As $t$ changes from $a$ to $b$, the point $P(g(t), f(t))$ sweeps out a continuous curve joining $A(g(a), f(a))$ with $B(g(b), f(b))$. This curve and the points $A$ and $B$ are pictured in Figure 11 on the following page.
Under this interpretation the left side of (ii), \( \frac{f(b) - f(a)}{g(b) - g(a)} \), represents the slope of the secant line \( AB \). The expression \( \frac{f'(t)}{g'(t)} = \frac{dy}{dx} \) is the slope of the tangent to the curve \( C \) at the point \( P(g(t), f(t)) \). Consequently Cauchy's formula has an interpretation in terms of slopes analogous to the interpretation of the mean value theorem. The formula (ii) asserts that at least one point \( (g(\xi), f(\xi)) \) on \( C \) between \( A \) and \( B \) the tangent line is parallel to the secant line \( AB \) (Figure 11).

The analogy between the mean value theorem and Cauchy's formula can be extended. Following the methods that worked for the mean value theorem, we proceed to construct auxiliary functions to which Rolle's theorem may be applied. This time, however, we must keep in mind that the curve is described by parametric equations. In
this discussion the geometric meaning of the auxiliary functions used by Smail and by Goodman will become clear.

First we construct a function $F(t)$ which gives the vertical displacement from secant line $AB$ to the curve $C$. The line, of which $AB$ is a segment, has slope $\frac{f(b) - f(a)}{g(b) - g(a)}$ and passes through $A(g(a), f(a))$. The point-slope equation of this line, in terms of the parameter $t$, is:

$$\gamma(t) = f(a) + \frac{f(b) - f(a)}{g(b) - g(a)} [g(t) - g(a)].$$

That is, for any $t$ in $[a, b]$, $\gamma(t)$ is the ordinate to the line $AB$ at the point $x = g(t)$. The vertical displacement between secant and curve corresponding to a value $t$ is the difference between the ordinates to the secant and to the curve at $t$, namely, the difference between $f(t)$ and the right hand member of (1). Thus

$$F(t) = f(t) - f(a) - \frac{f(b) - f(a)}{g(b) - g(a)} [g(t) - g(a)]$$

is the auxiliary function we seek. Notice that this is exactly the function used by Smail.

We have already seen that this function satisfies the conditions of Rolle's theorem. Thus $F'(t)$ will vanish for some value $t = \xi$ between $a$ and $b$; i.e.,

$$F'(t) = f'(t) - \frac{f(b) - f(a)}{g(b) - g(a) \ g'(t)} = 0 \quad \text{when} \quad t = \xi.$$

The formula

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}$$

follows easily.
As before, an expression for the area $K(t)$ of the triangle ABP (see Figure 11) can be used as an auxiliary function. This area will again be expressed as a determinant. Referring to Figure 11 and noting the coordinates of the points A, B and P we write

$$2K(t) = \begin{vmatrix} g(t) & f(t) & 1 \\ g(a) & f(a) & 1 \\ g(b) & f(b) & 1 \end{vmatrix}.$$ 

Here is precisely the "deus ex machina" that appeared in Goodman's proof. The deduction of Cauchy's formula is accomplished as before.

Thus it has been shown that the first auxiliary function was an expression for the vertical displacement from secant line to curve, when the curve is described by parametric equations $x = g(t)$ and $y = f(t)$. The second auxiliary function can be interpreted as the area of a triangle determined by one variable point $P(g(t), f(t))$ and the fixed points A and B. These geometric interpretations are arrived at with such ease, that it seems a pity that they are not used, or at least alluded to verbally, when deducing Cauchy's formula.
VI. CURVATURE

The student must not believe that theorems have been invented or perfected by the methods in which it is afterwards most convenient to deduce them. The march of the discoverer is generally anything but on the line on which it is afterwards convenient to cut the road.

Augustus De Morgan

De Morgan could have been thinking of the mathematical concept of curvature when he wrote these words. We present this concept today, in textbook and classroom, in a manner quite at variance with the "march of the discoverer." The modern presentation is brief and complete, but as we shall see it is not the only way to present this topic. With respect to the formulas relating to curvature, we will be able to observe with considerable clarity the path followed by the discoverer. From a study of the early ideas, we shall be able to suggest approaches to curvature that are more dynamic than the modern one, approaches that capitalize to a greater extent on the inherent geometry of the idea. These approaches, as one might expect, are longer and require more algebraic manipulation, but this manipulation is given direction by clear geometric ideas.


70
In this chapter we address the problem of measuring the amount of bending or curvature exhibited by a curve in the plane. Curvature is essentially a geometric notion. From experience we know that some curves bend more than others, and of course the amount of bending is in general different at different places on the curve. If our intuitive notion of curvature is not to be violated, the measure that we devise must assign greater curvature to a segment like this

A

than to a segment like this

B

Many times in mathematics, when we set out to describe some property of a curve we are led to a description of this property at a point on the curve (for example, continuity and differentiability of functions). Such is the case with respect to curvature. It is evident from what follows that mathematicians have found it fruitful to define and measure curvature at a point. This being so, we can make a more careful statement regarding segments like those pictured above, one of which bends more than the other. We can say that the measure of curvature at a point on a segment like A should be greater than this measure at a point on B.

Of the several possible approaches to the formulas for curvature, we shall here be least concerned with that one which appears in
a typical up-to-date textbook. The basic concern of the chapter is with alternatives to this approach. But for the sake of completeness and for purposes of comparison we indicate below the method that has found the greatest favor with textbook writers.

To define curvature, we consider the angle $\alpha$ which the tangent to a curve makes with the positive direction on the $x$-axis. Let $\Delta \alpha$ denote the change in this angle corresponding to tangents at two points $P$ and $Q$ on the curve (see Figure 12).

![Diagram of a curve with angles and tangents labeled](image)

Figure 12
Arc length \( s \) is to be measured along the curve from some fixed point on it. Of two arcs having the same length \( \Delta s \), the change in the inclination of the tangent, \( \Delta \alpha \), will be the greater on that arc which is the more bent.

Suppose that two points \( P \) and \( Q \) on the curve are separated by a length of arc \( \Delta s \). Let the inclinations of the tangents at \( P \) and at \( Q \) be \( \alpha \) and \( \alpha + \Delta \alpha \) respectively. The ratio \( \Delta \alpha / \Delta s \) is the average rate of change of \( \alpha \) with respect to the arc length \( \Delta s \). This ratio is called the **average curvature** of the arc \( PQ \) corresponding to the length \( \Delta s \).

Now let \( Q \) approach \( P \), or, what is the same thing, let \( \Delta s \to 0 \). The ratio \( \Delta \alpha / \Delta s \) will in general approach a limit, and

\[
\lim_{\Delta s \to 0} \frac{\Delta \alpha}{\Delta s} = \frac{d \alpha}{d s} = C
\]

is defined to be the curvature of the curve at \( P \). Thus curvature is defined to be the instantaneous rate of change of the angle of inclination per unit of arc length.

The formulas relating to curvature can be developed as shown below.

Let \( y = f(x) \) be the equation of the given curve. Since \( y' = \tan \alpha \) is the slope of the curve at \( P \), it follows that

\[
\alpha = \tan^{-1} y'.
\]

Then

\[
d \alpha = \frac{y''}{1 + y'^2} \, dx.
\]

Also

\[
ds = (1 + y'^2)^{1/2} \, dx.
\]
Consequently

\[ C = \frac{\frac{da}{ds}}{ds/dx} = \frac{\frac{da}{dx}}{ds/dx} = \frac{y''}{(1 + y'^2)^{3/2}}. \]

This formula gives the curvature of a curve defined by an equation \( y = f(x) \). The expression will be negative where the curve is concave down, positive where the curve is concave up, and zero at a point of inflection. Some authors write \(|y''|\) in the numerator to insure a non-negative number for the curvature.

Let \( C \) be the curvature of \( f(x) \) at \( P \). Draw a normal to the curve at \( P \) on the concave side and mark off on this normal a segment \( PS \) (see Figure 13) whose length is \( 1/C \). With \( S \) as center and radius \( 1/C \) draw the circle which is tangent to the curve at \( P \). This circle is called the circle of curvature, or osculating circle to the curve at \( P \). \( S \) is called the center of curvature, and \( PS = 1/C \) the radius of curvature.

![Figure 13](image-url)
To find the coordinates \((h,k)\) of \(S\) we use the relation \(PS = 1/C\) and the fact that the slope of \(PS\) is \(-1/y'\), where \(y'\) is evaluated at \(P\). The equation of the circle is

\[
(h - x)^2 + (k - y)^2 = PS^2 = \left(\frac{1}{C}\right)^2 = \frac{(1 + y'^2)^3}{y''^2}.
\]

Also

\[
(h - x) + y'(k - y) = 0.
\]

Eliminating \((h - x)\) between these equations gives

\[
(k - y) = \frac{1 + y'^2}{y''}.
\]

It follows that

\[
h = x - \frac{y'(1 + y'^2)}{y''},
\]

\[
k = y + \frac{(1 + y'^2)}{y''}.
\]

These are the coordinates of \(S\), the center of curvature.

This is how the basic formulas are being taught today. The student may never realize that there is any other way to develop this concept.

The notion of curvature is not a recent addition to the world of mathematics, but has been in the literature for many years. The idea has important applications in physics and engineering.

Not at all apparent from the modern treatment of curvature is the fact that there are certain primitive ideas concerning the curva-
tecture of a plane curve. These ideas are primitive in the sense that they developed long ago in the minds of mathematicians, and primitive in the sense that they seem simple and obvious to most people (at least after a little prodding). Among these primitive notions are the following.

A given circle is equally curved at all points. The circle bends in a completely regular way. The definition of curvature that is selected must assign equal curvature to all points on a circle.

Furthermore, big circles (circles with large radii) don't curve as much as little circles (circles with small radii). The larger the radius the smaller the curvature, and the smaller the radius the larger the curvature. Thus it seems natural to take as the curvature of a circle a quantity which is inversely proportional to the radius of the circle. In fact, it has turned out that we take the constant of proportionality to be 1 and write \( C = \frac{1}{R} \) where \( R \) is the radius of the circle and \( C \) is the curvature. Thus, if the radius is halved the curvature is doubled, if the radius is tripled the curvature is one-third what it was before, etc.

Next, consider an arbitrary curve and a fixed point on it. We should somehow be able to find a circle that "fits" the curve at this point better than any other circle. Some circles will be too big and flat to fit well - some too small and sharply curved to fit well (the arguments presented toward the end of this chapter prove the existence of such a circle, and give a graphic means for visualizing it). If such a circle can be found, it would seem reasonable to agree that this circle and the curve have the same curvature at this point, and therefore that the curvature of the curve should be the reciprocal of the radius of this circle.
It will prove instructive to review a bit of the history of the notion of curvature. In doing this we shall see the extent to which the "march of the discoverer" differs from the way we "cut the road" today. The historical study will also disclose two alternate methods for developing the required formulas. A modification of one of these methods seems to be particularly dynamic in the sense of points, lines, and circles moving and then coming to rest. The chapter will conclude with consideration of this last method.

The first writer to hint at a definition of curvature was Nicolas Oresme, writing in the fourteenth century. Oresme's dates are 1323-1382. He is quoted as saying: "If we have two curves touching the same line at the same point, and on the same side, the smaller curve will have the greater curvature." Oresme also states that the curvature of a circle is uniform, and there are clear indications in his writings that he thought of the curvature of a circle as being inversely proportional to the radius.

Another mathematician who did some of the early work relating to curvature was Christiaan Huygens (1629-1695). Huygens' work is found in the book Horologium Oscillatorium published in 1673.

---

31 This quotation is taken from the article cited below. A more complete account of the early history of curvature together with several references to original writings can be found here.

In Horologium Oscillatorium Huygens examines in some detail the problem of describing the involute and evolute to a curve, a problem closely related to curvature. He was able to find what amounted to the radius of curvature at points on certain special kinds of curves (the cycloid, for example), but Huygens did not have the calculus as a working tool. Consequently he did not obtain a satisfactory result in the most general case of an arbitrary function $f(x)$. The method used by Huygens, when he was successful, was to find the radius of curvature as the limiting position of the point of intersection of the normal to a curve at a point and another normal infinitely close to the first normal. In a note on pages 226, 227 of Volume 18 the editor shows how the modern formula for the radius of curvature,

$$\frac{1 + \left(\frac{dy}{dx}\right)^2}{\left(\frac{d^2y}{dx^2}\right)^{3/2}}$$

can be obtained by translating Huygens geometric argument into modern notation and then taking limits. Thus Huygens was very close to a satisfactory treatment of curvature. However, the terms "radius of curvature" and "circle of curvature" apparently do not appear in his writings.

In the "avertissement" by the editor at the beginning of Volume 18 of Huygens' Ouvre Complètes is a paragraph which summarizes
Huygens' understanding of curvature:

C. Rayon de Courbure. On a remarqué qu'en établissant la théorie des développées [involutes] Huygens ne dit rien sur la courbure. Cantor affirme: "dass Huygens an Krümmungsverhältnisse gar nicht dachte." Il est vrai qu'il n'observe pas ... que le point d'intersection de deux normales infinites voisines est le centre d'un cercle de courbure, mais la notion de courbure ne lui était nullement étrangère.\(^3\)

It was Sir Isaac Newton, no less, who first published a thorough discussion of curvature. Newton's work on this subject became available in English in 1736. In that year the book *The Method of Fluxions and Infinite Series* was published in London, being a translation of Newton's original Latin into English.\(^4\)

A photo-copy of the title page is included for its historical interest. The s's that look like f's, and the capitalization of nouns seems indicative of a strong Germanic influence on the written English of this time. Complete information about the book is given on this page: we are even told who sells the book, and where. One wonders what sort of establishments the Lamb and Temple-Bar were.

\(^3\) Ibid., pp. 41,42.

THE METHOD of FLUXIONS
AND INFINITE SERIES;
WITH ITS
Application to the Geometry of Curve-lines.

By the Inventor
Sir ISAAC NEWTON, Kz.
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.

By JOHN COLSON, M.A. and F.R.S.
Master of Sir Joseph Williamson's free Mathematical-School at Rochester.

LONDON:
Printed by HENRY WOODFALL;
And Sold by JOHN NOURSE, at the Lamb without Temple-Bar.
M.DCC.XXXVI.
Apparently, Newton was the first to use the terms "radius of curvature" and "center of curvature" in mathematical publication. At least one historian, Cantor, believed that Newton examined some of Huygens' work, and was able to develop the ideas more fully with the aid of his powerful tool, the calculus.

Newton assumes, as did Huygens, that if one considers normals to a curve at two points, one point being fixed, then the intersection of these normals will approach a definite limiting position as the second point approaches the fixed point. He called the point so determined the "center of curvature" and the distance to this point from the fixed point the "radius of curvature." This already suggests one alternate method for developing the curvature formulas.

Let us examine Newton's own words on this matter as they appear in The Method of Fluxions and Infinite Series. We shall see that ideas previously designated as primitive notions appear here. The writer believes that this material, together with several additional paragraphs to be quoted later, is the genesis of our modern conception of curvature.
PROB. V.

At any given Point of a given Curve, to find the Quantity of Curvature.

1. There are few problems concerning curves more elegant than this, or that give greater insight into their nature. In order to its resolution, I must premise these following general considerations.

2. I. The same circle has everywhere the same curvature, and in different circles it is reciprocally proportional to their diameters. If the diameter of any circle is as little again as the diameter of another, the curvature of its periphery will be as great again. If the diameter be one-third of the other, the curvature will be thrice as much, etc.

3. II. If a circle touches any curve on its concave side, in any given point, and if it be of such magnitude, that no other tangent circle can be interscribed in the angles of contact near that point, that circle will be of the same curvature as the curve is of, in that point of contact. For the circle that comes between the curve and another circle at the point of contact, varies less from the curve, and makes a nearer approach to its curvature, than that other circle does. And therefore that circle approaches nearest to its curvature, between which and the curve no other circle can intervene.

4. III. Therefore the center of curvature to any point of a curve, is the center of a circle equally curved. And thus the radius or semidiameter of curvature is part of the perpendicular to the curve, which is terminated at that center.
5. IV. And the proportion of curvature at different points will be known from the proportion of curvature of equi-curve circles, or from the reciprocal proportion of the radii of curvature.

6. Therefore the problem is reduced to this, that the radius, or center of curvature may be found.

7. Imagine therefore that at three points of the curve, σ, D, and d, perpendiculars are drawn, of which those that are at D and σ meet in H, and those that are at D and d meet in h: And the point D being in the middle, if there is a greater curvity at the part Dσ than at Dd, then DH will be less than dh. But by how much the perpendiculars σH and dh are nearer the intermediate perpendicular, so much the less will the distance be of the points H and h: And at last when the perpendiculars meet, those points will coincide. Let them coincide in the point C, then will C be the center of curvature, at the point D of the curve, on which the perpendiculars stand; which is manifest of itself.

8. But there are several symptoms or properties of this point C, which may be of use to its determination.

9. I. That it is the concourse of perpendiculars that are on each side at an infinitely little distance from DC.

10. II. That the intersections of perpendiculars, at any little finite distance on each side, are separated and divided by it; so that those which are on the more curved side Dσ sooner meet at H, and those which are on the other less curved side Dd meet more remotely at h.
11. III. If DC be conceived to move, while it insists perpendicularly on the curve, that point of it C, (if you except the motion of approaching to or receding from the point of insistence C,) will be least moved, but will be as it were the center of motion.

12. IV. If a circle be described with the center C, and the distance DC, no other circle can be described, that can lie between at the contact.

13. V. Lastly, if the center H or h of any other touching circle approaches by degrees to C the center of this, till at last it coincides with it; then any of the points in which that circle shall cut the curve, will coincide with the point of contact D.

14. And each of these properties may supply the means of solving the problem different ways: But we shall here make choice of the first, as being the most simple.

15. At any point D of the curve let DT be a tangent, DC a perpendicular, and C the center of curvature, as before. And let AB be the absciss, to which let DB be apply'd at right angles, and which DC meets in P. Draw DG parallel to AB, and CG perpendicular to it, in which take Cg of any given magnitude, and draw go perpendicular to it, which meets DC perpendicular to it, which will be Cg : go :: (TB : BD ::) the fluxion of the absciss, to the fluxion of the ordinate.

Likewise imagine the point D to move in the curve an infinitely little distance Dd, and drawing
de perpendicular to DG, and Cd perpendicular to the curve, let
Cd meet DG in F, and σg in f. Then will De be the momentum of
the absciss, de the momentum of the ordinate, and σf the
contemporaneous momentum of the right line gσ. Therefore DF
= De + \( \frac{de \times de}{De} \). Having therefore the ratio's of these moments,
or, which is the same thing, of their generating fluxions, you
will have the ratio of CG to the given line Cg, (which is the
same as that of DF to σf,) and thence the point C will be
determined.

16. Therefore let AB = x, BD = y, Cg = l, and gσ = z; then it
will be l : z :: x : y, or z = \( \frac{y}{x} \). Now let the momentum of z be \( z \times o \), (that is, the product of the velocity and of an
infinitely small quantity o,) and therefore the moments
De = \( \dot{x} \times o \), de = \( \dot{y} \times o \), and thence DF = \( \dot{x} o + \frac{\dot{y} \times o}{x} \). Therefore
'tis CG(l) : CG :: (σf : DF ::) \( \dot{z} o + \frac{\dot{y} \times o}{x} \). That is,
CG = \( \frac{\ddot{x} + \ddot{y}}{xz} \).

17. And whereas we are at liberty to ascribe whatever velocity
we please to the fluxion of the absciss \( \dot{x} \), (to which, as to an
equable fluxion, the rest may be referr'd;) make \( \dot{x} = l \), and
then \( \dot{y} = z \), and CG = \( \frac{1 + zz}{z} \). And hence DG = \( \frac{z + z^3}{z} \), and
DC = \( \frac{1 + zz \sqrt{1 + zz}}{z} \). 35

Consider now the parts of this material that relate to our central concern here. Newton has numbered these paragraphs in sequence, which makes for easy reference.

Newton lists first the ideas which he is going to take as premises, in paragraphs 2-5. There are really definitions too in these paragraphs. In 2 we recognize the ideas relating to the curvature of a circle which we have earlier called "primitive" notions.

Paragraph 3 contains Newton's definition of the circle of best fit, the definition being essentially that it is the circle that has equal curvature with the curve at the point.

In 4 he gives the definition of center of curvature and radius of curvature.

Paragraph 7 contains the justification for the method Newton is going to use to obtain the center of curvature, namely the method of intersecting normals. The ideas in 7 have heuristic appeal, and may represent a good way to argue the existence of a limiting position for the point of intersection. The normals will intersect as long as they are not parallel, and they won't be parallel in general if the curve bends at all. The normal at $\sigma$ on the more curved side of $D$ will meet the normal at $D$ sooner in the point $H$, and the normal at $d$ on the less curved side of $D$ will meet the normal at $D$ at a point $h$ such that $Dh$ is greater then $DH$. And ultimately, when the normals coincide, $H$ and $h$ also coincide.

The paragraphs 9-13 are most interesting. Newton lists here five different properties possessed by the point $C$, the center of curvature, any of which, he claims, can be used in its determination.
These properties are described for the most part in a dynamic way; i.e., one can visualize moving lines and moving circles and moving points of intersection. Of these several ways Newton chooses, in 14, the first as being the most simple. In this study, we will consider the approach suggested in 9, which is the one Newton selected first, and also a modification of the approach suggested in 13.

In paragraphs 15, 16, and 17 Newton finds the distance DC, which is the radius of curvature at the point he calls D. He assumes that C is a point on the normal to the curve at D, and finds the limiting position of C as the normal dC approaches the normal DC. He uses the single letter C to designate the point of intersection, and does not distinguish in his notation a point of intersection before the limit is taken from the limiting position of this point.

Since \( z = y = f'(x) \), and \( \dot{z} = f''(x) \), we recognize the very last equation in the quoted material as our modern formula for the radius of curvature; i.e.,

\[
DC = \frac{1 + zz}{\dot{z}} \sqrt{1 + zz}
\]

is equivalent to

\[
R = \frac{1}{C} = \frac{(1 + f'(x)^2)^{3/2}}{f''(x)}
\]

This may be the first time that this formula appeared in a mathematics textbook. Newton's discussion of curvature which we have just read is certainly the most complete to appear up to this time.
The argument used by Newton in paragraphs 15 and 16 to deduce the formula for curvature involves similar triangles, ratios, and a line $g_0$ that represents $z = f'(x)$. This argument does not seem to be well suited to use in a modern classroom, and the writer is not advocating its use at this time. However, it is not difficult to find the limiting position of the point of intersection of adjacent normals using up to date mathematical ideas and notation. We next give two modern realizations of Newton's basic ideas about intersecting normals. The first appears in a standard calculus textbook. The writer has not seen the second deduction in print.

In the book *Calculus*, by John Randolph, we find this proof:

We consider two points $P(x, y)$ and $Q(x + \Delta x, y + \Delta y)$ on the graph of $y = f(x)$, where $y + \Delta y = f(x + \Delta x)$ (see Figure 14).
We think of \( P \) as being fixed, and assume that \( P \) is not an inflection point of \( f(x) \) [so that \( f''(x) \neq 0 \) at \( P \)]. Let \((H,K)\) be the point of intersection of normals to the graphs of \( f(x) \) at \( P \) and at \( Q \). We seek the limiting position of \((H,K)\) as \( \Delta x \to 0 \).

It is assumed that \( f \) is a function whose second derivative exists. Recall that

\[
(1) \quad f''(x) = \lim_{\Delta x \to 0} \frac{f'(x+\Delta x) - f'(x)}{\Delta x}.
\]

Furthermore, if the second derivative exists at \( x \), then the first derivative is continuous at \( x \), and

\[
(2) \quad \lim_{\Delta x \to 0} f'(x+\Delta x) = f''(x).
\]

Now the slopes of the normals at \( P \) and \( Q \) are the negative reciprocals of the slopes of the curve at these points: thus

\[
\frac{K - y}{H - x} = -\frac{1}{f'(x)}
\]

and

\[
\frac{K - y - \Delta y}{H - x - \Delta x} = -\frac{1}{f'(x+\Delta x)}.
\]

From these two equations we can obtain two expressions for \( K - y \):

\[
K - y = -\frac{1}{f'(x)} (H - x),
\]

and

\[
K - y = -\frac{1}{f'(x+\Delta x)} (H - x - \Delta x) + \Delta y.
\]
Consequently:

\[- \frac{1}{f'(x)} (H - x) = - \frac{1}{f'(x+\Delta x)} (H - x - \Delta x) + \Delta y,
\]

\[
(H - x) \left\{ - \frac{1}{f'(x+\Delta x)} - \frac{1}{f'(x)} \right\} = \frac{\Delta x}{f'(x+\Delta x)} + \Delta y,
\]

\[H - x = \frac{f'(x)}{f'(x) - f'(x+\Delta x)} (\Delta x + \Delta y f'(x+\Delta x)),
\]

and thus

\[H = x - \frac{f'(x) \left\{ 1 + \frac{\Delta y}{\Delta x} f'(x+\Delta x) \right\}}{f'(x+\Delta x) - f'(x)}.\]

Upon letting \( h = \lim_{\Delta x \to 0} H \), we therefore have (recalling (1) and (2))

\[h = x - \frac{f'(x) \left( 1 + [f'(x)]^2 \right)}{f''(x)}.\]

It is easy to check that, upon setting \( k = \lim_{\Delta x \to 0} K \), one can obtain

\[k = f(x) + \frac{1 + [f'(x)]^2}{f''(x)} .\]

The point \((h, k)\) is called the center of curvature at the point \(P\) on the graph of \(y = f(x)\), and the circle with center \((h, k)\) and passing through \(P(x, y)\) is called the circle of curvature, the radius \(r\) of this circle is called the radius of curvature, and the reciprocal of the radius of curvature is called the curvature, \(C\).
The distance formula tells us that the radius of curvature is equal to

\[
\sqrt{(h - x)^2 + (k - y)^2} = \sqrt{\left[-\frac{f'(x)(1 + [f'(x)]^2)}{f''(x)}\right]^2 + \left[1 + \frac{[f'(x)]^2}{f''(x)}\right]^2}
\]

Thus

\[
r = \frac{(1 + [f'(x)]^2)^{3/2}}{|f''(x)|},
\]

and

\[
C = \frac{1}{r} = \frac{|f''(x)|}{(1 + [f'(x)]^2)^{3/2}}.
\]

It seems that this development of the formulas is fairly well motivated, and not at all beyond the capabilities of the average calculus student. At least one other method of solution of this problem, using intersecting normals, appears in the literature. The latter method is a bit more obscure, perhaps, but it is one whose intuitive appeal might be improved by careful study. 37


or

The writer would like to present a slightly different method for developing these formulas. We make the same assumptions about $y = f(x)$, $P$, and $Q$ as in the development from Randolph's book, but we now designate the fixed abscissa of $P$ by $x_0$. The point $(H,K)$ will again be the point of intersection of normals at $P$ and at $Q$ (see Figure 15).

![Diagram](image)

Figure 15

Now as $\Delta x$ changes, the point $Q$ changes position, and the point $(H,K)$ also changes position on the normal line at $P$. Clearly $H$ and $K$ are both functions of $\Delta x$ ($P$ being fixed during the discussion). We seek an expression for $K$ as a function of $\Delta x$, and from this we will be able to find the limiting position of $K$ as $\Delta x \to 0$. 
The equation of the normal to \( f(x) \) at \( P \) is
\[(x - x_0) + f'(x_0)(y - f(x_0)) = 0,\]
and the normal line at \( Q \) has the equation
\[x - (x_0 + \Delta x) + f'(x_0 + \Delta x)(y - (f(x_0 + \Delta x)) = 0,\]
where \( x \) and \( y \) here are what have been called "running coordinates."

Now the point \((H,K)\) is on both normals, hence
\[H - x_0 + f'(x_0)(K - f(x_0)) = 0 \quad \text{and} \quad H - x_0 - \Delta x + f'(x_0 + \Delta x)(K - f(x_0 + \Delta x)) = 0.\]

From these equations we can get two expressions for \( H - x_0 \):
\[H - x_0 = -f'(x_0)(K - f(x_0)) = \Delta x - f'(x_0 + \Delta x)(K - f(x_0 + \Delta x)).\]

Equating the two expressions gives us an equation free of \( H \), and linear in \( K \):
\[-f'(x_0)(K - f(x_0)) = \Delta x - f'(x_0 + \Delta x)(K - f(x_0 + \Delta x)).\]

Solving for \( K \) gives
\[K = \frac{f'(x_0 + \Delta x)f(x_0 + \Delta x) - f'(x_0)f(x_0) + \Delta x}{f'(x_0 + \Delta x) - f'(x_0)}.\]

This is the equation for \( K \) as a function of \( \Delta x \) that we sought. But now a difficulty arises. When we take the limit of this expression as \( \Delta x \to 0 \), by the continuity of \( f'(x) \) and \( f(x) \), we arrive at the indeterminate form \( 0/0 \), and learn nothing about \( k = \lim_{\Delta x \to 0} K \).

But the fact that this expression is the quotient of two functions of \( \Delta x \), and is indeterminate at \( \Delta x = 0 \) brings to mind l'Hôpital's rule. We differentiate numerator and denominator once with respect
to $\Delta x$, remembering to treat the term $f'(x_o + \Delta x)f(x_o + \Delta x)$ as a product, and the resulting expression is no longer indeterminate. Thus

$$
\lim_{\Delta x \to 0} K = \lim_{\Delta x \to 0} \frac{f'(x_o + \Delta x)f(x_o + \Delta x) - f'(x_o)f(x_o) + \Delta x}{f'(x_o + \Delta x) - f'(x_o)}
$$

$$
= \lim_{\Delta x \to 0} \frac{f''(x_o + \Delta x)f(x_o + \Delta x) + [f'(x_o + \Delta x)]^2 + 1}{f''(x_o + \Delta x)}
$$

$$
= f(x_o) + \frac{1 + [f'(x_o)]^2}{f''(x_o)}
$$

Thus we have obtained this formula for $k$:

$$
k = f(x_o) + \frac{1 + [f'(x_o)]^2}{f''(x_o)}
$$

Since $(h,k)$ is restricted to lie on the normal at $P$, we have as before

$$
h - x_o = -f'(x_o)(k - f(x_o)) \quad \text{, where} \quad h = \lim_{\Delta x \to 0} H,
$$

from which this expression for $h$ follows:

$$
h = x_o - \frac{f'(x_o)[1 + [f'(x_o)]^2]}{f''(x_o)}
$$

From these values we can obtain the curvature, $C$, as before:

$$
C = \frac{|f''(x)|}{[1 + [f'(x_o)]^2]^{3/2}}
$$
In addition to the intersecting normal technique there is another way to develop the formulas for curvature which seems to have a great deal of intuitive appeal. There may be disadvantages to the approach that will be presented, but the writer urges consideration of this method by anyone interested in a well motivated presentation of curvature. We shall see shortly that this method too can be traced back to Isaac Newton.

We take as primitive notions, or premises, the facts that the curvature of a circle is everywhere the same, and is in fact equal to the reciprocal of the radius of the circle. The definition of curvature at a point \( P \) on a curve \( y = f(x) \) is obtained from this definition of curvature of a circle. In particular, we define in a very clear way the circle of best fit, or osculating circle, to \( f(x) \) at \( P \), and then agree that the curvature of \( f(x) \) at \( P \) equals the curvature of this circle at \( P \).

The most interesting aspect of this approach is the definition of the circle of best fit. Before giving this definition, and because a close analogy can be drawn between the slope of a curve and the curvature of a curve, let us recall the usual definition of the tangent line to a curve at a point \( P \).

Let \( P \) be a fixed point on a curve, and let \( Q \) be another point on the curve, \( Q \) being near \( P \) (see Figure 16). We consider the secant lines \( PQ \). As \( Q \) approaches \( P \) along the curve the secant lines \( PQ \) will in general approach a limiting position \( PT \). The line \( PT \) is by definition the tangent line to the curve at \( P \).
Notice that use is made of the fact that two points determine a line in the plane, and in a sense we have defined a line of best fit to the curve at P. Furthermore, we know what we mean by the slope of a line. The definition of slope of a curve is made in terms of the line of best fit at a point P. It is agreed that the slope of a curve at a point is the slope of the tangent at that point.

An analogous approach can be used with respect to curvature of an arbitrary curve at a point. We fit a circle of best fit to the curve at the point. Then, knowing what we mean by the curvature of a circle, we can agree that the curvature of the curve at the point equals the curvature of the circle at the point.

---38---

Recall l'Hospital's definition of a tangent. He thought of the curve as being made up of an infinite number of very short straight lines, and to define a tangent at a point he simply extended one of these little segments.
Consider a fixed point P on the curve, and two nearby points Q and R also on the curve. Any three distinct points in the plane determine a unique circle; hence there is one circle which passes through P, Q, and R (see Figure 17).

It is reasonable to suppose that, as both Q and R approach P along the curve, the collection of circles passing through these three points will approach one fixed circle as a limiting position. This fixed circle is the circle of best fit, or osculating circle, to the curve at P. The curvature of the curve at P is then defined to be the curvature of the circle of best fit, and is therefore equal to the reciprocal of the radius of this circle. Thus we use the curvature of a simpler figure to define the curvature of a more sophisticated figure.
This three-point approach (so called for lack of a better name) seems to give very graphic meaning to the words "circle of best fit;" it is the circle which a whole family of circles approaches as P, Q, and R come to coincide. One has the feeling that the circles are being fitted, by the restriction that they pass through the three points on the curve, and that in the limit the amount of bending of the circle and of the curve are the same.

Let's see what kind of mathematical problems we become involved in when we try to implement this idea. A search has disclosed that this three-point approach has been used in the past, and several different developments of the curvature formulas are known.

Augustus De Morgan includes this idea in a book published in 1842. His approach is appealing, but he is soon led into algebraic manipulations so complex as to be unsuitable for use in the classroom.

A German textbook by Kiepert-Stegemann, noted for its intuitive appeal, uses three points and the second difference of a function to develop the curvature formulas in quite a nice way.

One of the simpler developments of the curvature formulas using three points is to be found in a calculus book by Neelley and Tracey. Notice that Rolle's theorem plays an important role here.

---


It will be noted later that the deductive scheme used by Neelley and Tracey parallels one suggested by Newton.

109. Osculating Circle.

The osculating circle of a given curve at a point \( P \) is the limiting position of the circle through \( P \) and two other points \( P_2 \) and \( P_4 \) of the curve as \( P_2 \) and \( P_4 \) approach \( P \).

Let \( y = f(x) \) be the equation of a given curve and let the circle through the points \( P, P_2 \) and \( P_4 \) on the curve have center at \( C'(h', k') \) and radius \( R' \). Its equation is

\[
(x - h')^2 + (y - k')^2 - R'^2 = 0.
\]

Since \( P, P_2 \) and \( P_4 \) are on both the curve and the circle, if we denote by \( \phi(x) \) the left-hand member of (1) when \( y \) is replaced by \( f(x) \), then \( \phi(x) \) will vanish for the abscissas of \( P, P_2, \) and \( P_4 \).
that is,

\[
\begin{align*}
\varnothing(x) &= 0 \\
\varnothing(x_2) &= 0 \\
\varnothing(x_4) &= 0.
\end{align*}
\]

Hence by Rolle's theorem, \( \varnothing'(x) \) must vanish for at least two values of \( x \), one at \( P_1 \) for \( x_1 \) between \( x \) and \( x_2 \), the other at \( P_3 \) for \( x_3 \) between \( x_2 \) and \( x_4 \), that is

\[
\varnothing'(x_1) = 0, \quad \varnothing'(x_3) = 0.
\]

Applying the same theorem again, \( \varnothing''(x) \) must vanish for some value of the variable between \( x_1 \) and \( x_3 \), say \( x_o \), then

\[
\varnothing''(x_o) = 0.
\]

Now let \( P_2 \) and \( P_4 \) approach \( P \); then \( P_0, P_1, P_3 \) will each approach \( P \) as a limit, and \( h', k', \) and \( R' \) will approach \( h, k, \) and \( R \) of the osculating circle at \( P \). Then, corresponding to (2), (3), and (4), we have the limiting values of \( \varnothing, \varnothing' \), and \( \varnothing'' \), namely,

\[
\begin{align*}
\varnothing(x) &= (x - h)^2 + (y - k)^2 - R^2 = 0 \\
\varnothing'(x) &= 2[(x - h) + (y - k)y'] = 0 \\
\varnothing''(x) &= 2[1 + y'^2 + (y - k)y''] = 0
\end{align*}
\]

respectively.

These relations enable us to find \( h, k, \) and \( R \) in terms of \( x, y, y' \), and \( y'' \). Solving these,
we have:

\[
\begin{cases}
    h = x - \frac{y'(1 + y'^2)}{y''} \\
    k = y + \frac{1 + y'^2}{y''} \\
    R = \frac{(1 + y'^2)^{3/2}}{y''}
\end{cases}
\]

(8)

In general, the circle crosses the curve at P.\(^{41}\)

We return now to Newton's Method of Fluxions and Infinite Series, and give evidence that he too was aware of the three-point approach to curvature. We will point out that the proof we have just examined is but a modern adaptation of a technique suggested by Newton. Furthermore, the material we are about to quote will suggest a modification of the three-point idea that can be used to find the formulas for curvature and that is well motivated.

Following Newton's preliminary discussion of curvature, and his development of the formula \(DC = \frac{1 + zz}{\sqrt{1 + zz}}\), comes eight pages of examples. The formula is applied to the hyperbola, the Cissoid of Diocles, the conchoid, the cycloid, and the like. One has the feeling in reading through this that he has dismissed the theoretical treatment of curvature. But after finishing these examples, he suddenly redirects his attention to property V of the five properties possessed by the center of curvature, and writes several paragraphs

that are very interesting in view of the purpose of this study. We read:

54. And now I have finish'd the problem; but having made use of a method which is pretty different from the common ways of operation, and as the problem itself is of the number of those which are not very frequent among geometricians: For the illustration and confirmation of the solution here given, I shall not think much to give a hint of another, which is more obvious, and has a nearer relation to the usual methods of drawing tangents. Thus if from any center, and with any radius, a circle be conceived to be described, which may cut any curve in several points; if that circle be suppos'd to be contracted, or enlarged, till two of the points of intersection coincide, it will there touch the curve. And besides, if its center be suppos'd to approach towards, or recede from, the point of contact, till the third point of intersection shall meet with the former in the point of contact; then will that circle be aequicurved with the curve in that point of contact: In like manner as I insinuated before, in the last of the five properties of the center of curvature, by the help of each of which I affirm'd the problem might be solved in a different manner.

55. Therefore with center C, and radius CD, let a circle be described, that cuts the curve in the points d, D, and σ; and letting fall the perpendiculârs DB, db, σβ, and CF, to the abscess AB; call AB = x, BD = y, AF = v, FC = t, and DC = s. Then

BF = v - x, and DB + FC = y + t. The sum of the squares of these is equal to the square of DC; that is, \( v^2 - 2vx + x^2 + y^2 + 2yt + t^2 = ss \). If you would abbreviate this, make \( v^2 + t^2 - s^2 = q^2 \) (any symbol at pleasure)
and it becomes \(x^2 - 2vx + y^2 + 2ty + q^2 = 0\). After you have found \(t, v, \) and \(q\), you will have \(s = \sqrt{v^2 + t^2 - q^2}\).

56. Now let any equation be proposed for defining the curve, the quantity of whose curvature is to be found. By the help of this equation you may exterminate either of the quantities \(x\) or \(y\), and there will arise an equation, the roots of which, (\(db, DB, \sigma\beta, \) etc. if you exterminate \(x\); or \(Ab, AB, A\beta, \) etc. if you exterminate \(y\),) are at the points of intersection \(d, D, \sigma, \) etc. Wherefore since three of them become equal, the circle both touches the curve, and will also be of the same degree of curvature as the curve, in the point of contact. But they will become equal by comparing the equation with another fictitious equation of the same number of dimensions, which has three equal roots; as Des Cartes has shew'd.

Or more expeditiously by multiplying its terms twice by an arithmetical progression.

EXAMPLE. Let the equation be \(ax = yy\), (which is an equation to the parabola,) and exterminating \(x\), (that is, substituting its value \(\frac{yy}{a}\) in the foregoing equation,) there will arise

\[
\frac{y^4}{aa} - \frac{2y}{a} y^2 + 2ty + q^2 = 0
\]

three of whose roots \(y\) are to be made equal.  

\[
\begin{array}{cccc}
4 & 2 & 1 & 0 \\
3 & 1 & 0 & -1
\end{array}
\]

And for this purpose I multiply the terms twice by an arithmetical progression, as you see done here; and there arises \(\frac{12y^4}{aa} - \frac{4y}{a} y^2 + 2y^2 = 0\).

Or \(v = \frac{3y^2}{a} + \frac{1}{2} a\). Whence it is easily infer'd, that \(BF = 2x + \frac{1}{2} a\), as before.
58. Wherefore any point $D$ of the parabola being given, draw the perpendicular $DP$ to the curve, and in the axis take $PF = 2AB$, and erect $FC$ perpendicular to $FA$, meeting $DP$ in $C$; then will $C$ be the center of curvity desired.

59. The same may be perform'd in the ellipsis and hyperbola, but the calculation will be troublesome enough, and in other curves generally very tedious. 42

There is no doubt that we have examined an early treatment of the three-point approach to curvature. In Newton's own words, this approach, when compared to the intersecting normal approach, "... is more obvious, and has a nearer relation to the usual methods of drawing tangents." The equation $v^2 - 2vx + x^2 + y^2 + 2yt + t^2 = ss$ found in 55 is equivalent to the equation of the circle with center $C(v,t)$, passing through the point $D(x,y)$. Exactly such an equation was used in the proof from Neelley's book. In 56 we read: "By the help of this equation [the equation that defines the curve] you may exterminate either of the quantities $x$ or $y$. ..." In Neelley's proof, $y$ was replaced by $f(x)$, thereby "exterminating $y$" precisely as Newton suggests, and the result was the equation $\vartheta(x)$. But then the modern proof appealed to Rolle's theorem to obtain the result. Newton multiplied the terms of the equation twice by an arithmetical progression (which means that he is taking a second derivative) in order to obtain three equal roots. It is difficult to find in Newton's writing here sufficient justification

42Newton, op. cit., pp. 70,71.
for doing this. But up to a point the modern proof parallels exactly what Newton suggested many years ago.

Newton's solution using three points, found in 55 and 56, does not follow exactly the approach that he outlines in 54. Yet 54 suggests a modification of the three-point method that is not only appealing to the intuition, but can be used to develop the curvature formulas. The modification is that the circle is fitted to the curve in two steps, as we explain below.

Consider, in particular, these sentences from 54:

Thus if from any center, and with any radius, a circle be conceived to be described, which may cut any curve in several points; if that circle be suppos'd to be contracted, or enlarged, till two of the points of intersection coincide, it will there touch the curve. And besides, if its center be suppos'd to approach towards, or recede from, the point of contact, till the third point of intersection shall meet with the former in the point of contact; then will that circle be aequicurved with the curve in that point of contact.

The idea can be rephrased as follows. Let $P(x_o, f(x_o))$ (fixed), $Q$, and $R$ be three points on the curve $y = f(x)$. There is a unique circle passing through these three points. It is assumed that $R$ and $Q$ approach $P$ independently (Figure 18).
In step one of the fitting, we let \( R \) approach \( P \), and hold \( Q \) fixed. Clearly the secant lines \( PR \) approach a limiting position which is at the same time the tangent to the curve and the tangent to the circle at \( P \). This is what Newton means when he says: "...the circle... will there touch the curve." Thus the curve and the circle have a common tangent at \( P \), with slope given by \( f'(x_0) \).

Now after \( R \) has come to coincidence with \( P \), a unique circle through \( P \) and \( Q \) is still determined, the third condition being that the center must lie on the normal at \( P \). There is just one normal line at \( P \); since the curve and the circle share a tangent line there, they also share a normal line.

The final step in fitting the circle is accomplished by letting \( Q \) approach \( P \). As \( Q \) moves along the curve, the center of the circle also moves, but the motion of the center is restricted to lie on the normal at \( P \). This step is described by Newton with these words: "And besides, if its center be suppos'd to approach towards, or recede from, the point of contact, till the third point of intersection
shall meet with the former in the point of contact, then will that circle be aequicurved with the curve ...." 

Our final consideration in this chapter will be to see how the above approach can be carried through to develop the formulas for curvature. It is suggested that most of the steps in this deduction are well motivated by geometric considerations.

To this end, let \( y = f(x) \) be the equation of the curve, and let \( P, Q, \) and \( R \) be points on the curve as before. We suppose that the preliminaries regarding the curvature of a circle and regarding fitting a circle to a curve have been taken care of.

Lines \( CR, CP, \) and \( CQ \) are radii of the circle, hence normal to the circle at \( P, Q, \) and \( R. \) Until \( R \) coincides with \( P, \) the center \( C(H,K) \) of the circle will in general not be on the normal to the curve at \( P. \) But when \( R \to P, \) the curve and the circle share a tangent, hence also a normal direction, and \( C(H,K) \) is on this common normal.

Suppose that \( R \) has come to coincidence with \( P, \) so that the first derivative, calculated from the curve, determines the normal direction of both the curve and the circle at \( P. \) Let the coordinates of \( P \) be \( P(x_o,f(x_o)) \), while those of \( Q \) are \( Q(x_o+\Delta x,f(x_o+\Delta x)) \) (see Figure 19).
An increment in $x$, $\Delta x$, determines a unique circle. Let $C(H,K)$ be the center of this circle. The equation of the line normal to $f(x)$ at $P$ is

$$x - x_0 + f'(x_0)(y - f(x_0)) = 0.$$ 

Since $C(H,K)$ is on this normal line, $H$ and $K$ have this linear relationship:

$$H - x_0 + f'(x_0)(K - f(x_0)) = 0,$$

or

(i) $$H = x_0 - f'(x_0)(K - f(x_0)).$$

Now as $\Delta x$ changes, $C(H,K)$ will move along the normal line. Again, both $H$ and $K$ are functions of $\Delta x$, and we seek the limits, $h$ and $k$, of these coordinates as $\Delta x \to 0$. Having found these limits, we
will have the center of curvature, \( C(H,K) \), the radius of curvature, and consequently also the curvature of \( f(x) \) at \( P \). The technique to be used is similar to the one used with adjacent normals. We seek an expression for \( K \) as a function of \( \Delta x \), and take the limit of this expression as \( \Delta x \to 0 \). This time the expression for \( K \) is obtained by equating two expressions for the radius of the circle.

For any \( \Delta x \neq 0 \), since \( CP \) and \( CQ \) are radii of the circle, this equation obtains:

\[
(H - x_o)^2 + (K - f(x_o))^2 = (H - (x_o + \Delta x))^2 + (K - f(x_o + \Delta x))^2.
\]

After squaring and simplifying in a perfectly straightforward manner this equation becomes

\[
-2Kf(x_o) + f(x_o)^2 = -2\Delta x(H - x_o) + \Delta x^2 - 2Kf(x_o + \Delta x) + f(x_o + \Delta x)^2.
\]

From (i) we substitute \( x_o - f'(x_o)(K - f(x_o)) \) for \( H \), and solve for \( K \) to obtain

\[
K = \frac{f(x_o + \Delta x)^2 - f(x_o)^2 - 2\Delta xf'(x_o)f(x_o) + \Delta x^2}{2[f(x_o + \Delta x) - f(x_o) - \Delta xf'(x_o)]}
\]

When \( \Delta x \to 0 \) this expression assumes the indeterminate form \( 0/0 \). We appeal to l'Hospital's rule, and examine the quotient of two derivatives, the primes indicating differentiation with respect to \( \Delta x \). We are led to this expression:

\[
\frac{f(x_o + \Delta x)f'(x_o + \Delta x) - f'(x_o)f(x_o) + \Delta x}{f'(x_o + \Delta x) - f'(x_o)}.
\]
The reader may note that this is identically the expression that we

got for K using adjacent normals and l'Hospital's rule. This quotient

is again of the form 0/0 at Δx = 0, so we must apply l'Hospital's

rule once more to obtain

\[
\frac{f'(x_0 + \Delta x)^2 + f''(x_0 + \Delta x)f(x_0 + \Delta x) + 1}{f''(x_0 + \Delta x)}.
\]

Upon taking the limit of this expression as \(\Delta x \to 0\), we obtain exactly

as before

\[
k = \lim_{\Delta x \to 0} K = f(x_0) + \frac{1 + f'(x_0)^2}{f''(x_0)}.
\]

The remaining formulas are obtained in a manner already illustrated.

This chapter has shown a bit of the historical development of

the notion of curvature, and it is clear that several different app­

roaches to the problem can be made. For purposes of good motivation, the

writer's preference is for the three-point approach, for it has "...a

nearer relation to the usual methods of drawing tangents." Two distinct

methods for deducing the required formulas from the circle determined by

three points have been studied, one using Rolle's theorem and one using

l'Hospital's rule. In many modern calculus books, the introduction of

l'Hospital's rule follows the work on curvature. To use l'Hospital's

rule in the deduction, it would be necessary to present this topic

before taking up curvature. If one did this, i.e., presented l'Hospital's

rule first, then the development of the formulas for curvature could be

introduced as a useful application of l'Hospital's rule.
VII. SUMMARY AND SUGGESTIONS FOR FURTHER STUDY

This study has been carried out with two objectives in mind. The first objective is the rather obvious one that is implied by the title - to find methods whereby each of the four theorems can be introduced in a more meaningful way. If this objective has been achieved to any degree, then the study may have implications for classroom teaching. The second objective has been to show how an investigation into the historical development of a mathematical idea may disclose an inductive basis for the idea. If this objective has been to any degree achieved, then the study may have implications for research in mathematics education. We wish to include in this final chapter a brief discussion relative to each of these objectives.

With regard to the first objective, it is hoped that this study has brought out ideas that can be used to breathe life into the exposition of each of the four theorems. The modern textbook presentation of these theorems is unexciting, and perhaps too brief. On the other hand, the exposition as it appears in this study is perhaps too long. It is not expected that all of this material can be incorporated into a textbook. But the ideas are easily grasped, for the most part, and could be explained verbally in a short time in the classroom.

The working out of one example, similar to the one in Chapter III, should make the student familiar with the basic ideas of l'Hôpital's rule.
To develop the chain rule along the lines suggested here, the student must know how to graph composite functions. It seems worthwhile to spend time on this topic, since it serves not only as a preliminary to the chain rule itself, but also serves to reinforce the students' understanding of composite functions.

Cauchy's formula is an ominous looking expression when viewed for the first time, and the auxiliary functions used in its deduction even more so. The time spent giving graphic interpretation to the formula and to the auxiliary functions seems to be a worthwhile investment. Cauchy's formula can be made to appear as a natural extension of the mean value theorem. It is likely that if the student has studied, in connection with the first mean value theorem, both the displacement function $F(x)$ and the function that gives the area of a triangle $K(x)$, then after seeing the interpretation of Cauchy's formula in terms of parametric equations he could all but discover a proof.

The usual development of the formulas for curvature seems unusually dry and unilluminating. The student misses a wealth of historical material which supports the idea of curvature. The teacher, at least, should know something of this history, as it can be used to enliven class discussion. The analogy between two points, a line, the slope of a curve, and three points, a circle, and the curvature of a curve represents a natural sort of extension of an idea. Of course, we cannot expect to do all of mathematics simply by extending already familiar ideas, but neither should we balk at using an analogy when we find one.
In a typical calculus class the students may be separated according to ability into at least three groups: the very brightest students who will use mathematics professionally, those who lack the training and ability to succeed in the calculus course, and a middle ability group making up perhaps sixty per cent of the class. It is with this middle ability group that the motivational ideas in this study might find greatest use. This group of capable students can learn to use the calculus, but they often reserve judgment concerning a career until after their initial experience with the calculus. Not much enthusiasm for mathematics is generated in a course which appears to be composed of a collection of unrelated and poorly motivated tricks. On the other hand a well motivated introduction to the subject should give the student greater confidence in his ability to use mathematics as a natural sort of tool in scientific inquiry.

Otto Toeplitz was a man much concerned with the problem of attracting mathematically capable students. We quote from an article published by him in 1927. It is apparent that this problem, of importance in the German universities of this time, is still very much with us.

The young student who has tentatively decided to study mathematics, and is debating whether to make this decision final, would like to know how far mathematics is exciting and beautiful; whether it is worth dedicating his life to.

Let us not talk about the few chosen ones that know this, so to say, by nature - the stock mathematicians that cannot become anything except mathematicians. This five per cent of the students in our beginning course (a high estimate, by the way) learn such
techniques with ease, and hardly require a course for it; for them the problem is not worth discussing. (And it is just from this group, be it noted, that many of the university professors are recruited; and they, in planning the course, remember too easily their own mathematical development.) True, if we should teach the course on the level of this five per cent then certainly a few others would be pulled along to a height they would not attain were the course on a lower level. It has been borne out by experience, however, that certainly not everybody is pulled along; and, more important, these do not include all who have mathematical inclination. Just as the joy of playing can be taken away from a pianist at the beginning by too strenuous finger exercises, mathematics can be made unattractive, to many who are otherwise capable, by a calculus presentation that clothes the skill in apodictic rules. Not the worst among our beginners would also like to know why the things happen. \(^{43}\)

We make a few remarks now regarding the second objective of this study. It appears that an historical search into original sources can sometimes be fruitfully undertaken. Such a search may disclose the inductive means originally used to discover a result. In the case of l'Hospital's rule, we have conjectured that Theorem A embodies the ideas that made the rule seem obvious to l'Hospital. Newton's extended discussion of curvature served to suggest a modified approach to this topic. In certain cases, after an inductive basis has been found, it

---

is entirely possible that a student might be led to retrace these steps and to discover the result for himself. The traditional tell-'em, drill-'em, and test-'em approach, so often used in college teaching, could be infused with an element of discovery after an intuitively appealing approach is known.

From experience gained in collecting material for this dissertation, it appears that original writings cannot be used without some modification. When the material goes back 150 years or more, we find a development that is not very slick by modern standards, a development that we would not want to use in a modern classroom. But sometimes rewriting an idea in precise modern notation will yield a workable idea. Occasionally, of course, such a search may fail to turn up anything of use.

Several other characteristics of this study should be noted.

Everyone knows that pictures are a great aid in learning mathematics. We employ schematic representations at all levels, from pictures of kittens in the first grade to pictures of space filling curves in graduate school. Through pictures we can add a dynamic quality to our exposition. Where possible in this dissertation an attempt has been made to give meaning to results and direction to algebraic manipulation through geometric means.

For example, in discussing the chain rule, we visualized a change $\Delta x$ in $x$ as generating a whole series of changes in other variables, $\Delta u$, $\Delta f$, and $\Delta F$. If all you are looking at is an analytic expression then you can't imagine anything moving as $\Delta x \to 0$; all you notice is that
the symbol $\Delta x$ has disappeared. With the geometric representation, even
the average student can visualize the length of a line segment going
to zero, and the dependent lengths $\Delta u$, $\Delta f$, and $\Delta F$ also going to zero.

In the chapter on Cauchy's formula we gave visual meaning to
two auxiliary functions by interpreting them either as a vertical dis­
placement or as the area of a triangle. It was clear from the geometry
of the situation that both these functions were zero at the end-points
of an interval.

The dynamic quality became most apparent in the discussion of
curvature, where we fitted a circle to a curve using three points. Two
points moved to coincidence, bringing the center of the circle onto the
normal, and then the center assumed a limiting position as the third
point came to coincidence.

If these motions can be visualized using a fixed, static draw­
ing, it is conceivable that they might become dramatic when drawn for
a film. There is interest in adapting certain topics from the calculus
to films, as is evidenced by the fact that the Committee on Educational
Media (CEM) of the Mathematical Association of America is currently
making a series of films on topics in the calculus.

The type of historical research exhibited here could be
continued by any person interested in finding good motivation for
mathematical results. This study has of necessity had a very limited
focus - in fact the theorems were drawn only from the differential
calculus. One does not have to look far in the integral calculus to
find examples of results given without good motivation. The use of
unmotivated auxiliary functions is not uncommon in many parts of
mathematics. Such functions are used in algebra, number theory, and in topology. It seems likely that in almost every case there exists some good intuitive reason for the choice of the function, although this reason is not always brought to light.

It is hoped that ideas contained in this dissertation will interest persons concerned with several different types of activities in mathematics education. Textbook authors might consider modifying the usual, standardized approach to each of these theorems in view of the ideas. It is possible that the dynamic quality of the developments could be effectively adapted to use with newer educational media, such as programmed learning and films. And it is hoped that these ideas will be welcomed by those teachers who struggle to breathe life into a beginning course in calculus.
BIBLIOGRAPHY


