Calculus Revisited
Part 1

A Self-Study Course

Supplementary Notes

Center for Advanced
Engineering Study

Herbert I. Gross

Massachusetts
Institute of Technology

Cat. No.-26-4001

CALCULUS REVISITED
PART 1

A Self-Study Course

SUPPLEMENTARY NOTES

Herbert I. Gross

Center for Advanced Engineering Study
Massachusetts Institute of
Technology

PREFACE

One might expect that, armed with good lectures and lecture notes, an excellent textbook, and exercises worked out in meticulous detail, the student would have no need for any additional components in a self-study course. Yet, there are many reasons why supplementary notes are necessary.

For one thing, topics that are a prerequisite for calculus today were not part of the curriculum several years ago. Thus, for the student who has been away from school for any appreciable length of time, it turns out that he may not ever have studied the prerequisite material. For example, much of the "new" mathematics is involved with the study of sets. In particular, the textbook assumes that the reader is at least somewhat familiar with the concept. Since this might not be so, we have taken the precaution of making sure that the fundamentals of set theory are available in the supplementary notes.

Other topics in the text are presented adequately but without sufficient motivation or generality. That is, they are introduced to solve a particular problem, but they are not developed in their own right or in greater overview. In such cases, the supplementary notes are used to shed more light on the topics. For example, our treatment of mathematical induction is from this point of view.

Still other topics are, to put it bluntly, just plain difficult. For these topics, it often happens that even with the textbook, lectures, and exercises, the student has great difficulty. Thus, another use of the supplementary notes is to supply other points of view and additional "rehash" of these difficult concepts in the hopes that the student will better grasp the concept if he has a choice of approaches to it. Our treatment of infinitesimals and the definite integral are examples of such topics in this volume of notes.

Finally, there are those topics which traditionally are beyond the scope of a beginning course but which are important and within the grasp of the student who is taking a refresher course. The supplementary notes try to treat these topics adequately and in a self-contained way. For example, our notes on uniform convergence fall into this category.

Thus, the supplementary notes are another part of our overall package. In many places the material is self-contained and can be learned and appreciated quite apart from the actual course. In other places, no attempt is made to make the material self-contained. Rather, it is our intent only to supplement the treatment given in other parts of the package. In these places, the student who tries to read the notes as a self-contained entity will be somewhat at a disadvantage.

In all cases, however, the notes are meant to help you obtain mastery of the subject and as a result they should be read carefully as they are assigned.

Whenever possible the notes are written in an informal style and proofs are included only so that interested students will not be left "up in the air." That is, since the notes do not in any way duplicate the text, virtually every proof in the notes is given simply because it does not appear anywhere else in the package. When important abstract proofs are supplied in the text, the supplementary notes try only to motivate how and why the proof was developed, and the actual proof is left for the interested student to glean from the text.

Cambridge, Massachusetts
June 1970

Herbert I. Gross

TABLE OF CONTENTS

I. An Introduction to the Theory of Sets

A.	Introduction	1
B.	Set Notation	4
C.	Two Special Sets	11
D.	Set-Builder vs. Roster Method	16

II. The Arithmetic of Sets

A.	Unions, Intersections, and Complements	1
B.	Contrast Between Addition and Union	9

III. An Introduction to Functions and Graphs

A.	Introduction	1
B.	Functions and Sets	3
C.	Application of Functions to Real Numbers	13
D.	Graphs	21
E.	Inverse Functions	28

IV. Limits

A.	Introduction	1
B.	The Problem of Zero Divided by Zero (0/0)	3
C.	A Semi-Rigorous Mathematical Approach to Limits	7
D.	The Limit Concept in Terms of Graphs	15
E.	Limits - A Rigorous Approach	20

V. Mathematical Induction

A.	Introduction	1
B.	An Application to Limits	7
C.	Limitations of Induction	10

VI. Infinitesimals and Differentials

A. Introduction 1

B. Infinitesimals 6

C. The Chain Rule 10

D. A Note about Inverse Functions 18

E. Differentials 24

VII. The Trigonometric Functions

A. Introduction 1

B. Trigonometry without Triangles 13

C. Derivatives of the Trigonometric Functions 23

D. Circular and Hyperbolic Functions 29

E. The Inverse Circular Functions 36

VIII. Integral Calculus

A. Introduction 1

B. A Discussion of Areas of Non-Rectilinear Figures 2

C. Area as a Differential Equation 26

D. The Fundamental Theorems of Integral Calculus 33

IX. Logarithms Revisited

A. Logarithms without Exponents 1

B. A Note on the Connection Between $\ln x$ and $\log_e x$ 14

X. Sequences and Series

A. Introduction 1

B. (Infinite) Sequences 4

C. Cauchy Sequences 10

D. Series 16

E. Absolute Convergence and Rearrangements 25

F. Sequences of Functions 37

G. Uniform Convergence 44

H. Uniform Convergence of Power Series 57

Chapter I
AN INTRODUCTION TO THE THEORY OF SETS

A. Introduction.

It is a toss-up, at least at the elementary level, whether the number line or sets receives the most publicity in treatments of "modern" mathematics. Yet, the number line can be traced back to 600 B.C., while the "newer" concept of sets, as a self-contained study, can be traced back to about 1850 A.D. Among other things, this should serve as adequate evidence that "modern" is used, not so much as a synonym for "new," but rather, as a synonym for "meaningful" or "useful".

Why is the study of sets so meaningful? A precise answer to this question would result in a multi-volume text, but for our immediate purposes it might suffice as a rather crude approximation to say that, in the same way that numbers are the building blocks of arithmetic, sets are the building blocks of all mathematics. Thus, we find that sets can be used to help us examine every mathematical system, they can be used to help us better understand the basic ideas of probability theory including the topics of permutations and combinations, they can be used in the study of logic (Boolean Algebra) including the designing of computers, they can be used to enhance our ability to study quantitative relationships (known in the literature as the theory of functions), and they can be used to help us gain an objective insight into the concept of infinity; in fact, the study of sets allows us to study the entire concept of counting in an extremely beautiful way.

With all of this ballyhoo about sets and how profound they are, it will probably seem surprising when we now confess that, when all is said and done, a set is nothing more than a

COLLECTION. Thus, while the study of sets may have had its formal origin in 1850, sets themselves must have been known from an intuitive point of view since the "dawn of consciousness". That is, whether we use the word "set" when we refer to a set of dishes or a set of books, or whether we use a synonym for "set" when we refer to a flock (a collection) of sheep or a herd (collection) of cattle, or even more indirectly when we talk about the world series (a set of baseball games) or the Boston Redsox (a set of baseball players), the fact remains that we are dealing with the basic, simple, "self-evident" concept of a collection. In short, the present emphasis in the curriculum on the concept of sets should make us feel a little like the high school freshman who was amazed when he found out that he had been speaking prose his entire life!

In all fairness, however, there is one refinement concerning the mathematical definition of a set which should be stressed. Since, from a practical point of view, mathematics is the hand-maiden of science and technology, and since, whenever possible, the scientist desires quantitative (or, at least, objective) measurements rather than subjective ones (for example, in a precise experiment, he would feel more comfortable knowing that he wanted water heated to 80°F than to be told to use "lukewarm" water), it should not be hard to understand that the mathematician might want the idea of objectivity carried over to his concept of a set.

For example, our mathematician might balk at studying the set of all beautiful paintings because the membership rule for this set is quite subjective. Indeed, the winner of any beauty contest depends on the panel of judges; with a different panel, we might well have a different winner. This is exactly what we mean by the cliche "Beauty is in the eye of the beholder."

In summary, for scientific purposes at least, we wish to study only those sets which are more than just random collections, subjectively accumulated. Somehow or other we want a bit more precision than that of the first grade youngster (this is a true story, and sets are really being studied this early in the curriculum) who summarized what he had learned about sets by saying: "A set is a bunch of dogs." Instead, we would like to limit our study of sets to those for which the members can be tested objectively - that is, in a precise way which does not depend on the "judge".

To get a more concrete idea of what we are talking about, consider the set of all people now living who were born on April 2, 1929. If we wish to decide whether a person belongs to this set, we need only ascertain the date of his birth. Such a task does not depend on the person who is chosen to serve as judge (to be sure, the efficiency in ascertaining the date of birth may depend on who is the judge, but the test for membership involves only the information, not how it was obtained). Thus the set of all living people born on April 2, 1929, has an objective test for membership.

With the above as background, let us try to introduce the concept of a well-defined set. A SET IS SAID TO BE WELL-DEFINED IF, GIVEN ANY OBJECT, THERE IS AN OBJECTIVE RULE WHICH ALLOWS US TO DETERMINE WHETHER THE OBJECT BELONGS TO THE SET, OR NOT. ONE OR THE OTHER MUST HAPPEN, BUT NOT BOTH.

As to the decision of whether we have a "plain" set or a well-defined set, let us simply observe that any set we elect to study in this course is very likely a well-defined set. For example, most mathematics books refer to the SET of rational numbers. When we look up the definition of a rational number we find that a rational number is any number which is the quotient of two integers. Observe that this definition supplies us with a well-defined, objective test for membership. Assuming that our judge can perform the operation known as division (and if he can't, he shouldn't be studying rational numbers anyway), he has an objective test for determining whether a given number is rational. To be sure, some tests may be more simple to apply than others but this does not affect the test for membership. (For example, in terms of decimals, the rational numbers are precisely those decimals which either terminate or repeat the same cycle endlessly.)

Our point is that in virtually every situation wherein one would want to use sets, it turns out that the set is a well-defined set. By the same token, by our insisting that we deal only with well-defined sets, we introduce the necessary machinery for removing from our conversation those sets for which membership testing is subjective.

In all that follows in this text, we shall assume, unless specifically stated to the contrary, that all our sets are well-defined sets.

B. Set Notation

With the idea of a set firmly entrenched in our minds, let us now turn to the "nitty-gritty" that is a part of every game - the basic nonemclature. For better or for worse, exciting or boring, like it or not, we must learn basic vocabulary before we can do anything else; so we might as well get started!

To begin with, unless it is otherwise specified, the convention is that we denote sets by upper case letters, and we use lower case letters to denote members (often called

ELEMENTS) of a set. Thus, sets would be denoted by A, B, C, etc., and elements by a, b, c, etc.

Then, as is the case in all studies, we invent a convenient shorthand. Namely if we wish to indicate that b is an element of B, a member of B, or, more colloquially, that b belongs to B, we write:

$$b \in B$$

where the symbol "ε" stands for "is an element of."

If, on the other hand, we wish to indicate that b is not an element of B, i.e. it is false that b is an element of B, we use the standard mathematical gimmick of placing a "slash mark" through the symbol that denotes the relation. In other words

$$b \notin B$$

By way of illustration, let W denote the set of whole numbers. Then one would write $3 \in W$, since 3 is a whole number, but we would write $1/2 \notin W$, since 1/2 is not a whole number.

At this point it is important to make sure that we realize that ε is a relation between an ELEMENT and a SET. It is not a relation between two SETS. For example, consider the set of states known as the United States. Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut are elements of this set; but the set that these six states make up (namely, The New England States) is not an element of the United States (that is, the New England States are not a state)*,

*This should not be confused with the fact that it is possible that the elements of a set are themselves sets. For example, consider the state of Maine which is an element of the United States. It is itself a set of people. (Notice in this case that an individual person is not a member of the set of states.)

On the other hand, it is clear that the New England States are a part of the United States. In still other words, every state which is a member of the New England States is a member of the United States.

In more abstract terms, if A and B are sets, it is possible that all elements of A are elements of B (or more concisely, all A's are B's). In terms of our new language, this says that if $x \varepsilon A$ then $x \varepsilon B$.

At any rate, this motivates our next item of vocabulary. We invent a new bit of shorthand and write

$$A \subset B$$

as an abbreviation for

All A's are B's.

Moreover, when we write $A \subset B$, the accompanying prose is to say that A is a SUBSET of B, or if we wish to place the emphasis on B, we say that B is a SUPERSET of A.

By the way, it should be noted that when we say

All A's are B's

we cannot be sure about the truth of

All B's are A's

If we wish to emphasize that A is a subset of B but that there is at least one member of B which is not a member of A, we often write $A \subsetneqq B$.

By way of illustration, let, as before, W denote the set of whole numbers, and let R denote the set of rational numbers. Now, clearly every whole number is a rational number. (A rational number is any number which is the quotient of two integers; thus if n is a whole number, we have $n = n \div 1$, and n is the quotient of two integers.) On the other hand, not every rational number is a whole number. For instance, 1/2 is not a whole number since there is no whole number whose double is 1. The point is that this whole paragraph, if we use our new language, can be elegantly abbreviated into:

$$W \subsetneq R!$$

Also by way of review notice that, according to our definitions, we would never write $W \varepsilon R$ since ε is reserved for relating an element to a set, and both R and W denote sets.

Another interesting point is that there seems to be some sort of a resemblance between the symbol $\subset$ and the symbol $<$ (which is used to denote "is less than" in a relation between numbers). While there is some similarity between the properties of these two symbols there is an extremely important difference. Namely, if a and b denote two unequal numbers, then it is a well-known rule of arithmetic that either $a < b$ or $b < a$. However, if A and B denote different sets, it need not be true that either $A \subset B$ or $B \subset A$. For example, let A denote the set of Frenchmen and let B denote the set of musicians. Since there is at least one musician who is not French, it is false that $B \subset A$. Similarly, since there is at least one Frenchman who is not a musician, it is also false that $A \subset B$. In fact, it should be easy to generalize the given example and observe that quite often if A and B are randomly chosen sets then both $A \subset B$ and $B \subset A$ are false. (Again, if we wish to abbreviate the statement that it is false that A is a subset of B, we utilize the "slash mark" idea and write $A \not\subset B$.)

A more interesting situation occurs when it happens that both $A \subset B$ and $B \subset A$ are true. This tells us that each A is a B and that each B is an A. When we put these two facts together, logic tells us that A and B consist of precisely the same members. In this event we prefer to say that A and B are equal and we write $A = B$. In summary:

> Given two sets A and B, we say $A = B$ if $A \subset B$ and $B \subset A$.

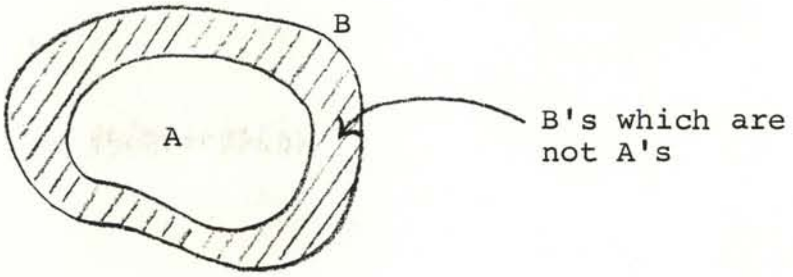
Paraphrased into plain English, $A = B$ means that A and B are two different names for the same collection (set).

At this point, the astute critic could make an interesting observation. Namely, why should one wish to give the same collection two different names? The question is well taken, and a proper answer to it will unveil an important aspect of how one finds unifying threads in the systematic study of any subject. For example, let us take an illustration from plane geometry. Suppose we let A denote the set of all triangles which have two sides of equal length, and let B denote the set of all triangles which have two angles of equal measure*.

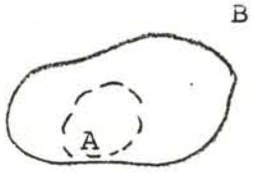
*At first glance the phrase "two angles of equal measure" may seem a stilted way of saying "two equal angles". In the "new" mathematics the point is made that it is not the angles that are equal (since they are located in different parts of space) but the measures (be the units, degrees, or radians or what have you). In a similar way it is not the two sides of an isosceles triangle that are equal (since the lines constitute two different sets of points); but the lengths of the two sides which are equal. However, now that we have paid the proper lip-service to the idea behind the language, we shall not become angry if we should forget ourselves and say things like "two equal angles". (To carry the formalism to an extreme, the old high-school geometry theorem that the sum of the angles in a triangle is 180° should read that the sum of the measures of the angles is 180°. Obviously, the sum of the angles of a triangle is 3!)

Now, there is no "inborn" reason for the geometry-unintiated to suspect that A and B are equal sets. However, in the course of time, one proves as theorems that all A's are B's and that all B's are A's; and in this way we obtain the unifying thread that the triangles which have two equal sides are PRECISELY those which have two equal angles.

As a final topic in our initial onslaught of amassing vocabulary, let us invoke the notion that a picture is worth a thousand words. One frequently uses "pictures" for "viewing" sets. These pictures are known as CIRCLE DIAGRAMS or VENN DIAGRAMS. Briefly summarized, we view a set as being contained with a closed curve (actually "closed curve," as we shall soon see, is a far better term than "circle" in the present context). Using this device we would illustrate the fact that $A \underset{\neq}{\subset} B$ by:



(By the way, if all we were given was that $A \subset B$, we would then draw the diagram as:



The dotted lines are used to warn us that we are only sure that the A-circle lies within the B-circle; and that we are not certain whether it is true that some B's are not A's.)

Earlier we had mentioned that neither the statement

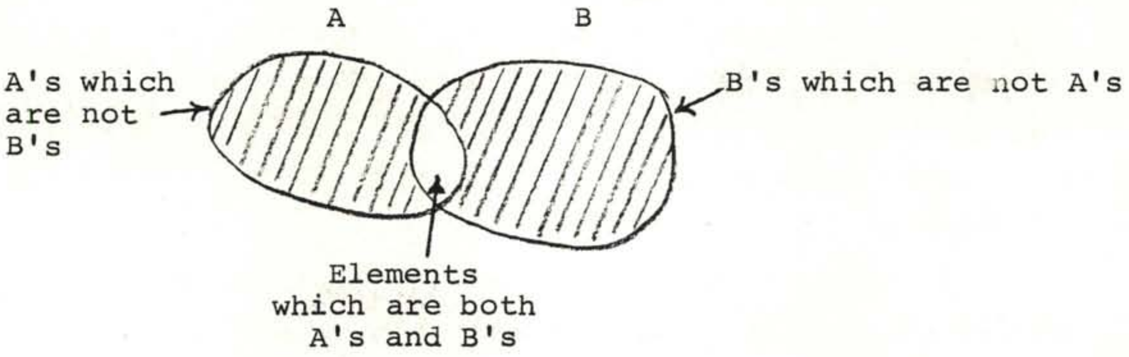
All A's are B's

nor the statement

All B's are A's

need be true.

In terms of our circle diagrams, this situation would be depicted by:



(Here we see a good reason, pictorially, for referring to "closed curves" rather than "circles." Namely suppose we draw A and B as circles and insist that all sets be circles. Then those objects which belong to both A and B form a set; but as the above diagram indicates, that region would not be a genuine circle, even though it would be a closed curve.)

In line with the representation of sets by circles, it is quite natural that the idea that A and B are equal would translate into the picture that the curves A and B enclose precisely the same region. This is exactly what is implied when we say that the A-circle is contained within the B-circle and that at the same time the B-circle is contained within the A-circle.

Notice how we have used the word "within" in the above paragraph. It is certainly natural from a geometric point of view to consider any region as being contained within itself. In terms of sets, this translates into the fact that it is perfectly proper to say that all A's are A's. It may be a truism - but it's still a fact. The important point here is that this gives us another important difference between ϵ and $\subset$. Namely, while a well-defined set is not a member of itself there is no such restriction placed on a well-defined set being a subset of itself. In other words, if A is a well-defined set, we have that $A \notin A$ but $A \subset A$*. Another way of visualizing this important result is to think of a set as being any collection and a subset as anything we can form by choosing members of the collection. In this context, if someone says to us "Take whatever you want", one choice we can make is to take everything.

Of course, there is another extreme, and that is that we may elect to take none of the collection. This has no bearing on what we have just discussed, but it plays a large role in motivating our next section.

C. Two Special Sets.

We have just intimated that it was conceivable that a set might have no members. Why would we allow a set to have no members? For one thing, if there are no members, why

*For example, the United States is not a state; nor is the National Football League a team in the National Football League.

bother naming it; for a second and more important thing, if a set is a collection, doesn't the term "collection" imply at least one member?

The answer to these questions centers around the idea of THE TEST FOR MEMBERSHIP that is implied for any set. That is, in terms of the test for membership, it is possible that the test for membership is so stringent that nothing can survive it. For instance let us consider the set of all numbers which are greater than 5 but less than 3. Certainly, we have an objective - indeed, meaningful - test for membership. Namely, given any number we see if it exceeds 5 and then we see if it is less than 3. If both of these things happen then our number belongs to the above set; otherwise, it doesn't. Of course, it happens that no number can survive the test for membership in this case; nevertheless, the test is objective and well-defined. Moreover, it is a significant piece of knowledge to discover that a test is so severe that no element can survive it.

To strengthen our case even more, notice how often (especially in mathematics) we are more interested in the test for membership than we are in knowing the names of all the members. For instance, we might be given a particular number and wish to know whether this number is a prime. At this point, we are more interested in knowing the recipe by which we objectively test the given number for being a prime than we are in wanting to see a complete listing of all the primes, for, among other things, the number of primes is infinite!

Later we shall supply even more reasons for allowing a set to contain no members. For now we wish only to establish the fact that such a set is meaningful.

In any event, it should be clear that such a set is the smallest possible collection. In other words, we do not say that a collection is so tiny that even if it had three more members there still wouldn't be any! Less colloquially, we do not have a notion of "negative" sets.

All summed up, then, we define THE EMPTY SET to be a set which has no members; and we denote the empty set (also called the NULL SET) by Ø.

By the way, do not confuse Ø with 0. For example, consider the set of all numbers which are neither positive nor negative. This particular set happens to have one member - namely 0. In other words, the set whose only member is 0 is not the empty set because the empty set has no members - not even 0. What is true, however, is that 0 denotes the NUMBER of elements in Ø. On the other hand the set of all numbers which are <u>both</u> even <u>and</u> odd is the empty set. (Notice that 0 is an even number, but 0 is <u>not</u> odd.)

To see what the empty set means in terms of grammatical structures, consider a sentence like:

> No dog has two heads.

In the language of sets, this says:

> The set of dogs which have two heads is the empty set.

Well, enough said about Ø for now. Instead, let us turn our attention to the other extreme mentioned in the last section on subsets. That is, you can't take more than you have. In other words, given a set A, the smallest subset of A is Ø and the largest subset of A is A itself.

This can be generalized, among other ways, by thinking of the following stupid question: "Does the color blue belong to the set of all lawyers?". At first glance the answer is no. At second glance the answer is even more emphatically no! In fact, we get the feeling that when we talk about the set of all lawyers, it is implicitly understood that only people are eligible for even the test for membership.

In other words, at certain times not only do we wish to limit membership in a set but also we even wish to limit those things that are eligible for the test for membership. To formalize this idea, we introduce the following definition:

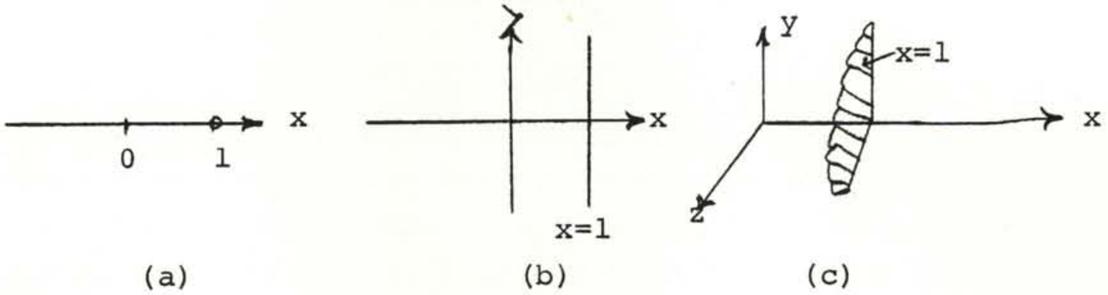
> By THE UNIVERSE OF DISCOURSE or THE UNIVERSAL SET, usually denoted by I, we mean the set such that for any element b and every set A which is being considered, it is true that $b \in I$ and $A \subset I$.

Thus, $\emptyset$ and I serve as "upper" and "lower" bounds in our discussion in the sense that no set being studied can have fewer than no elements nor can it contain any element not already contained in I. That is, for each set A, it is true that

$$\emptyset \subset A \subset I.$$

If we accept the fact that we can use a concept even though we don't know its name, it turns out that we have made use of the universal set many times in our treatment of elementary mathematics. Consider, for example, the fact that in ninth grade algebra we are told that $x^2 + 1$ cannot be factored; yet in the eleventh grade, we are taught that $x^2 + 1 = (x + i)(x - i)$. Certainly the body of knowledge in mathematics did not change that drastically during the two years! What happened was that in the ninth grade we were dealing with real numbers, while in the eleventh grade we were using the complex numbers. In terms of our above discussion, what we were really taught was that if the universe of discourse were the real numbers, then $x^2 + 1$ could not be factored, but if the universe of discourse were the complex numbers, then it could be factored.

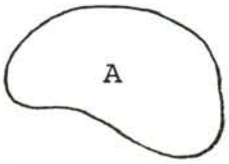
As a second example, let us refer to a problem in analytic geometry. Suppose we wish to know the graph of the equation $x = 1$. If our universe of discourse is the x-axis then the graph of our equation is precisely a single point. On the other hand, if our universe of discourse is the xy-plane, then our graph is a straight line. Finally, if our universe of discourse is 3-dimensional space, our graph is a plane. Pictorially:



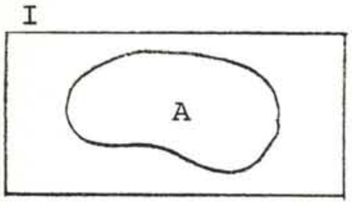
(a) (b) (c)

There are many such examples that can be supplied, but for our purposes the present two examples should suffice to show the meaning of the universal set.

In terms of Venn diagrams, it is conventional to represent I by a rectangle and to make sure that all sets under consideration are drawn within this rectangle. For example, instead of denoting A by



we would write



D. Set-builder vs. Roster method.

Now that we have established a few basic terms and the concept of a well-defined set, it might be in order to describe some general methods for describing sets. The two major methods are known as

(1) The Roster Method
and
(2) The Set-Builder Method

Each method has properties which, depending on the particular circumstances, make it more desirable than the other. Let us begin our discussion with the roster method. As the name may imply, the roster method is nothing more than an explicit listing of the members of a set. For example, if A were to denote the set of natural numbers which are less than ten, then use of the roster method would yield

$$A = \{1,2,3,4,5,6,7,8,9\}$$

It should be noted that the use of the braces to "enclose" the set is a universally-accepted convention. Logical or not, we agree never to use such notation as A = 1,2,3,4,5,6,7,8,9.

It should also be noted that in our definition of the equality of two sets, we merely specified that the two sets contain exactly the same members. We never required that the elements be listed in any particular order. This certainly agrees with our intuition and past experience. For example, if we were talking about the set of all men who were senators of the United States during 1964, we would not quibble about whether they should be listed by age, or by years of service, or in alphabetical order. Indeed, while one listing may be more convenient than another for a particular purpose, the fact remains that each is a listing of the required set. In terms of our example, we would write, for instance, $\{1,2,3,4,5,6,7,8,9\} = \{9,1,3,2,7,6,4,8,5\} = A$. In this same context, we agree that a set doesn't change merely by counting the same element more than once. For example, no matter how many times we wished to count "Monday" there are still only seven days in a week. Thus, for simplicity, we agree never to list the same element more than once. For example, we would write $\{1,2,3\}$ rather than $\{1,1,2,2,3,3,3\}$. (There are some exceptions; for example, suppose we wish to list the set of all letters which occur in the word, "Mississippi" and investigate the various rearrangements. In this case, we would most likely <u>not</u> write $\{M,i,s,p\}$, but rather $\{m,i,s,s,i,s,s,i,p,p,i\}$ since we are distinguishing between the first i and the second, etc. If this seems awkward, notice we may, for purposes of identification, imagine the i's, p's, and s's to be colored differently so that we can tell them apart.)

In contrast to the roster method, the set-builder notation describes the members of the set <u>implicitly</u> rather than <u>explicitly</u>. In other words, while the roster method actually lists the members of the set, the set-builder method describes, or emphasizes, the test for membership.

Specifically, in set-builder notation, we write $\{x:x.....\}$, which is read as "The set of all elements, x, such that......" For example, we might write $\{x:x \text{ is a real number}\}$ to indicate that we are referring to the set of real numbers. Or if we were going to make reference to the real numbers very often, we might say let R denote the real numbers, and we could then write, very compactly, that $\{x:x \,\epsilon\, R\}$ denotes the set of real numbers. As a non-mathematical example, let I denote the set of all Americans who were alive on January 1, 1960; and let B denote the set of all Americans who were born on April 2, 1929. Then by

$$\{x:x \,\epsilon\, I \text{ and } x \,\epsilon\, B\}^*$$

we would mean the set of all Americans who were alive as of January 1, 1960 and who were born on April 2, 1929. Notice the compactness of the set-builder method as, it clearly emphasizes the test for membership.

Now that the two methods have been described, let's discuss their relative strengths and weaknesses. We begin with the roster method. The obvious strength of this method is that it tells us outright what the members of the set are. That is, we have the easiest possible test for membership. Namely, if an element appears on the list, it belongs to the set; otherwise, it doesn't.

There are basic weaknesses, however, that can cause us great difficulty. For one thing, we cannot list an infinite set if only because life is finite. That is, suppose we wish to list the set of natural numbers. How shall we go about it? We could begin by writing: 1,2,3,4,5,6,7,8,9, 10, 11, but we would never come to an end. Some people write: 1,2,3,4....., where "..." may be viewed as standing for "etc."

*We usually abbreviate $\{x:x \,\epsilon\, I \text{ and } x \,\epsilon\, B\}$ by $\{x:x \,\epsilon\, B\}$. That is, when I is clear from context we assume $x \,\epsilon\, I$. In still other words: $\{x:x \notin I\} = \emptyset$.

However, "etc." can be vague and subjective, to say the least. As an example, consider the numbers generated by the following recipe:

$$a_n = n + (n - 1)(n - 2)(n - 3)(n - 4)$$

where a_n is merely an abbreviation for indicating the nth number. For example, if we wished the first number in our progression, we would choose n = 1, and write

$$a_1 = 1 + (1 - 1)(1 - 2)(1 - 3)(1 - 4)$$ (where we have merely replaced every n by 1)

Thus:

$$\begin{aligned} a_1 &= 1 + (0)\ (-1)\ (-2)\ (-3) \\ &= 1 + 0 \\ &= 1, \end{aligned}$$

and then we see that 1 is the first member of our sequence.

To obtain the second member, we replace every n by 2.

$$\begin{aligned} a_2 &= 2 + (2 - 1)\ (2 - 2)\ (2 - 3)\ (2 - 4) \\ &= 2 + (1)\ (0)\ (-1)\ (-2) \\ &= 2 + 0 \\ &= 2, \end{aligned}$$

and we see that our second term is 2.

Replacing n by 3, we next see that

$$\begin{aligned} a_3 &= 3 + (3-1)\ (3-2)\ (3-3)\ (3-4) \\ &= 3 + (2)\ (1)\ (0)\ (-1) \\ &= 3 + 0 \\ &= 3 \end{aligned}$$

Similarly

$$\begin{aligned} a_4 &= 4 + (4-1)\ (4-2)\ (4-3)\ (4-4) \\ &= 4 + (3)\ (2)\ (1)\ (0) \\ &= 4 + 0 \\ &= 4 \end{aligned}$$

In this way, we see that we have a well-defined test which tells us that the first four members in our sequence, in order, are 1,2,3,4. Suppose we told this to a person, but we did not tell him the "recipe" we were using. What do you guess he'd choose for the fifth member of the sequence? Well, it seems reasonable that when someone says 1,2,3,4...., we expect 5 to occur next. Yet

$$\begin{aligned} a_5 &= 5 + (5-1)\ (5-2)\ (5-3)\ (5-4) \\ &= 5 + (4)\ (3)\ (2)\ (1) \\ &= 5 + 24 \\ &= 29! \end{aligned}$$

a far cry from being "self-evident," if the recipe is withheld.

As other examples of this type of confusion, consider the sequence: o,t,t,f,f,s,s,..... What letter comes next? We claim that this sequence consists of the first letter in the name of each natural number starting with one, that is, one, two, three, four, five, six, seven, eight,...

Thus, in this example we were looking for "e" as our next entry, which is again, not so self-evident when the rule is witheld!

As a final example along these lines, consider the following sequence and try to decide what comes next:

31, 30, 31, 30, 31, 31, 30, 31, 30, 31, 31,...

Wrong if you said either 30 or 31! We weren't looking for either 30 or 31 but rather 28. For the above list is the number of days in a month, starting with March in a non-leap year.

These examples should serve as excellent reasons for why we have the right to shun the subjective "etc." Moreover, these examples should also serve to illustrate, at least indirectly, one advantage of the set-builder notation. Namely, notice how the "riddle" effect of these examples vanishes as soon as we let the other fellow in on the explicit rule we were following.

Aside from the fact that the roster method cannot be used without some ambiguity for trying to list infinite sets, there is even trouble in listing non-infinite sets. For example, the people who are named in the Boston telephone directory this year constitutes a finite set, yet a set large enough to be undesirable for the average man to list! Not only that, but even if a set has few elements, it might still be very difficult to list, because we may have trouble "locating" its members. To this end, consider the set of all living people who were born in New England on May 14, 1865. It is reasonable to assume that this set has relatively few numbers. In fact, it might well be empty. Yet, look at the difficulty that could arise if we were to try to determine the list of the actual members of this collection!

The very weakness of the roster method is the strength of the set-builder method, for, quite often, we are more interested in the test for membership than in the members themselves. For example, we might be more interested in knowing whether a particular number is divisible by 7, 11, and 13 than in knowing the entire set of numbers which are divisible by 7, 11, and 13. By the same token, the strength of the roster method is the weakness of the set-builder method. That is, there are times when we require the members of a set explicitly. For example, the lawyer might be more interested in the names of the people mentioned in a will than in the set-builder idea of "all my living relatives".

ONE OF THE UNIFYING THREADS IN THE PRESENTATION OF THIS CALCULUS COURSE WILL BE THAT OF JUXTAPOSITIONING THE ROSTER METHOD AND THE SET-BUILDER NOTATION SO THAT WE MAY BETTER UNDERSTAND THE OVERALL PROBLEM. THAT IS, WE SHALL USE THE SET-BUILDER METHOD TO EMPHASIZE THE PARTICULAR TEST FOR MEMBERSHIP; THEN WE SHALL APPLY VARIOUS COMPUTATIONAL TECHNIQUES WHICH WILL ALLOW US TO TRANSLATE FROM THE IMPLICIT FORM TO THE MORE EXPLICIT FORM OF THE ROSTER METHOD.

In order to illustrate this idea more concretely, we have chosen an example from elementary algebra so that we may highlight the technique without having it obscured by a maze of cumbersome computational devices.

In this section, it shall be our aim to show how the roster method and the set-builder method can be combined to give a "modern" meaning to "traditional" topics. In particular, let us focus our attention on the problem of finding roots of algebraic equations.

Consider the following "traditional" problem:

Find the roots of $x^2 - 4x + 3 = 0$.

In this problem one was expected to show that 1 and 3 were the required roots, whether the technique employed was that of factoring, or using the quadratic formula, or for that matter even trial and error. In terms of factoring, one used the fact that $x^2 - 4x + 3$ and $(x - 1)\ (x - 3)$ were "synonyms" and then replaced the original equation by the "equivalent" one,

$$(x - 1)\ (x - 3) = 0$$

This latter equation lent itself more readily to solution , by virtue of the theorem that if the product of two numbers is zero then at least one of the factors is zero, whence the result that either $x = 1$ or $x = 3$ followed immediately.

So much for that! In modern language, the same problem might read:

Find the solution set, S, for the equation $x^2 - 4x + 3 = 0$.

In this event, we would be expected to write that $S = \{1,3\}$.

Now, as we mentioned earlier, if we agree to be sensible about the whole thing, we will soon admit that it is wasteful to introduce new expressions just to express the same things. That is, what difference does it make whether we refer to the roots of an equation or whether we refer to its solution set, if we must still understand such techniques as the quadratic formula, factoring, etc., regardless of the language employed?

Our claim, of course, is that there is much more to this than meets the eye and involves our two major ways for describing sets. For example, we used the roster method when we said that the solution set, S, of the equation $x^2 - 4x + 3 = 0$ was given by $S = \{1,3\}$. On the other hand, had we wished to use the set-builder notation for expressing the solution set, S, of the equation $x^2 - 4x + 3 = 0$, we would have written

$$S = \{x: x^2 - 4x + 3 = 0\}$$

As we have mentioned, these two methods for describing sets are basically different. For example, $S = \{1,3\}$ tells us explicitly that our set consists of the numbers 1 and 3, but it does not tell us what property they share in common to make themselves members of the same set. On the other hand, $S = \{x: x^2 - 4x + 3 = 0\}$ tells us implicitly the property which all members of S must have in common, but it does not explicitly name for us the members of S.

Thus, we have two methods for describing the same set, much as in ordinary arithmetic we often have two (or more) ways for describing the same number. At the same time, again as in ordinary arithmetic, which of the methods is "best" depends on the specific problem being solved.

We shall now show how "traditional" algebra is enhanced, both for the student and the teacher, by an effective use of these two major methods for describing sets. Still working within the framework of the equation $x^2 - 4x + 3 = 0$, observe that the set-builder notation gives us an ultra-simple way for describing the solution set, S, of this equation. Namely, we need only write $S = \{x: x^2 - 4x + 3 = 0\}$.

It is now urgent to point out that the above solution is not a gimmick! Granted that the form of our answer may not be as reassuring to our intuition as the answer afforded by the roster method, and granted that it looks as if we really haven't done anything by the set-builder method, the important fact is that it focuses our attention on the meaning of the answer and an understanding of the question. That is, even if we have never heard of factoring or the quadratic equation we can use the test for membership to decide whether a given number belongs to S. For example, if we are given the number four to test, we need only check to see whether it is true that $(4)^2 - 4(4) + 3 = 0$. Since this is a false statement, the test for membership allows us to conclude that $4 \notin S$. On the other hand, since it is true that $(1)^2 - 4(1) + 3 = 0$, we can conclude that $1 \in S$, EVEN THOUGH WE MAY HAVE LACKED THE COMPUTATIONAL TOOLS TO DISCOVER WITHOUT TRIAL-AND-ERROR THAT $x = 1$ HAD TO BE A SOLUTION!

Indeed, one could now motivate algebra, at least in part, by observing that algebra is a system of techniques whereby one learns logically how to convert a solution set from the implicit set-builder notation into the more explicit roster form.

In other words, with respect to our illustrative problem, whether the technique be factoring or the quadratic formula or anything, it is a computational device which allows us to convert the set-builder form $S = \{x: \ x^2 -4x + 3 = 0\}$ into the roster form $S = \{1,3\}$. From a pedagogical point of view, this process allows us to focus our attention on two separate, but equally important, parts of the problem; namely, knowing what it means to solve the problem (that is, being able to recognize a solution when we see it), and finding techniques for helping us discover solutions more "conveniently".

Finally, if we recall that two sets are "equal" if they are "aliases" for one another, we can then get a rather meaningful interpretation concerning the equivalence of two or more equations. For instance, we are not born with the knowledge that $x^2 - 4x + 3$ and $(x - 1)\ (x - 3)$ "look alike"; thus, we should beware of such statements as "The equation $x^2 - 4x + 3 = 0$ is the same as the equation $(x - 1)\ (x - 3) = 0$". What we really mean (and notice how much less ambiguous things become this way) is that the two different equations have "the same" solution set. That is, if we let $S = \{x: \ x^2 - 4x + 3 = 0\}$ and $T = \{x: \ (x - 1)\ (x - 3) = 0\}$, then $S = T$. Even more explicitly, we have $S = T = \{1,3\}$. Notice also here that when we say $S = T$ we are not just being frivolous and giving the same set two different names. Rather, we have showed that the two different descriptions are actually equivalent. Moreover, when such is the case, it will frequently happen that one of the equivalent descriptions will be more advantageous to us in a certain situation than the other.

Chapter II
THE ARITHMETIC OF SETS

A. Unions, Intersections, and Complements

In Chapter I we introduced the concept of set along with a few basic properties. We pointed out that the study of sets served a multipurpose function in mathematics. More specifically, we used the concept of sets to show the inner mechanism whereby one proceeded from an implicit to an explicit form of an answer, and how the "recipes" usually associated with algebra fit into this overall pattern.

In this section we intend to introduce a few additional concepts in the study of sets, and we shall then show how sets may be used to unify apparently unrelated topics in the study of mathematics.

We begin by pointing out that, while we have applied the concept of sets to arithmetic, we have not applied the concept of arithmetic to sets. That is, as of now we have in no way explained how we may combine sets to form new sets. Thus, before proceeding further, let us first introduce the arithmetic of sets. For the present, we shall study this idea for its own sake and then later we shall see how this applies to other topics in mathematics.

From a very informal point of view, suppose that two organizations, which we shall call A and B, wish to form a merger. This means that a new organization incorporating A and B is formed. Calling this newly-formed organization C, we see that C consists *precisely of those elements which belonged to at least one of the original groups A or B*. In a sense, C may be viewed as the *union* of A and B. Notice that if there are well-defined tests for membership in A and B, then there is also a well-defined test for membership in C. Namely, given any element in our universe of discourse, we test to see whether it belongs to either A or B. If it belongs to neither then it will not belong to C; otherwise, it will.

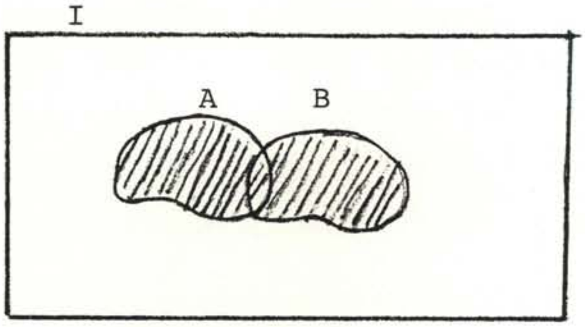
Using this as an intuitive motivation, we generalize this idea to cover all sets. First of all, to minimize any chance for paradoxes, we assume that we have a specific universe of discourse which we shall denote by I.

Definition 1: Let A and B be subsets of I. Then by the *union of A and B*, written $A \cup B$, we mean the *set* of all elements which belong to *either* A *or* B (unless otherwise specified, by "either .... or ...." we mean *at least one*[*]). In the language of sets:

$$A \cup B = \{x\colon x \in A \ \underline{\text{or}}\ x \in B\}$$

In terms of our circle diagrams:

[*] While this may seem strange to some, the fact remains that "either... or ..." is often used in this non-mutually exclusive sense. For example, when we say that either Tom or Jerry will go to the store, we do not preclude the possibility that both boys might go. When we reach into a deck of cards and say that we shall draw either a spade or a face card, we do not feel that we have lied should we draw the king of spades. On the other hand, there are times when "either ... or ..." means "one or the other, but not both." In this event, there is still no contradiction, for "one or the other, but not both" is covered by the expression "at least one." Thus, our only precaution is that we will say "either ... or ... but not both" when we mean "either ... or ..." in the exclusive sense.



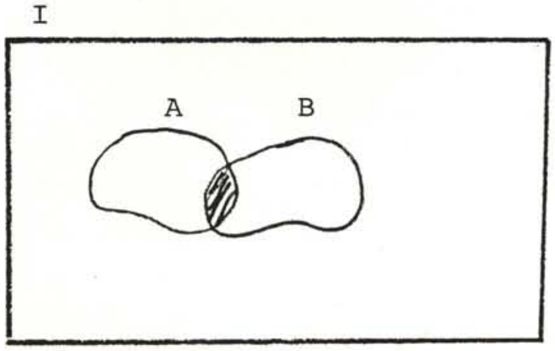
/// denotes $A \cup B$

Now we could have combined the sets A and B in a way almost completely "opposite" to that of union. For example, rather than a merger, we could have formed a "*shrinker*". That is, we could have consolidated our two organizations A and B by forming a new organization D, characterized by the fact that its members were those which belonged simultaneously to both A and B. That is:

Definition 2: Again let A and B be subsets of I. Then by the *intersection* of A and B, written $A \cap B$, we mean the *set* of all elements which belong to *both* A *and* B. More symbolically:

$$A \cap B = \{x: \ x \in A \ \underline{\text{and}} \ x \in B\}$$

The choice of the word "intersection" can be motivated in one way from the circle-diagrams by observing that $A \cap B$ actually is the intersection of the two circles. That is:

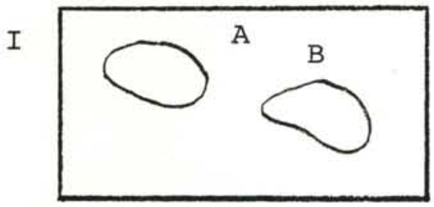


/// denotes $A \cap B$

Notice that Definitions 1 and 2 introduce operations on sets which are closed. That is, the union of two sets is again a set; and similarly, the intersection of two sets is a set.

This leads to still another reason for introducing the empty set, $\emptyset$. For, first observe that since $A \cup B = \emptyset$ if and only if $A = \emptyset$ and $B = \emptyset$ (why?), it would not be necessary to "invent" the empty set for unions, since the only way for a union to be empty is for the sets forming the union to be empty. However, it is possible for the intersection of two non-empty sets to be empty. For example, if I denotes the integers and if A denotes the even integers while B denotes the odd integers, we see that both A and B contain infinitely many elements; yet, their intersection is empty, since there are no integers which are simultaneously even and odd. That is, each integer is either even or odd, but not both. At any rate, this shows that if we did not consider the empty set, it might well happen that the intersection of two sets might not be a set.

$A \cap B = \emptyset$ is equivalent to the more familiar "No A's are B's" and translates into the following circle diagram:



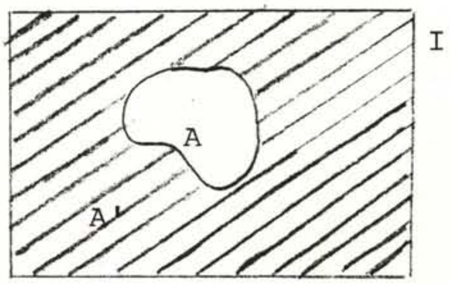
The final operation that we wish to introduce in this section is the concept of the complement of a set. Observe that whenever we choose a subset of the universe of discourse, we actually choose two subsets. For example, if I denotes the set of natural numbers and A denotes the subset consisting of the perfect squares, then we have induced the set B whose members are the natural numbers which aren't perfect squares. If I denotes all American citizens and A denotes the set of American citizens who reside in New York, then we can induce a set B, by allowing it to denote all American citizens who do not reside in New York. More generally:

Definition 3: By the complement of A, written A', we mean the set of all elements which belong to the universe of discourse but not to A. That is

$$A' = \{x: \ x \in I \text{ but } x \notin A\}$$

In terms of our circle-diagrams:

/// denotes A'



Remarks: (1) "But" logically has the same meaning as "and" in many contexts. In this sense, $A' = \{x: \ x \in I$ and $x \notin A\}$. (Thus, we may say that

$$A' = I \cap A'$$

This is not surprising since all elements belong to I, that is, if B is any subset of I then $B = I \cap B$.)

(2) The concept of complement does not depend on the concept of universal set. However, the complement of a given set does depend on the universal set. In other words, we do not say that A' means all non A's but rather, all I's which are non-A's.

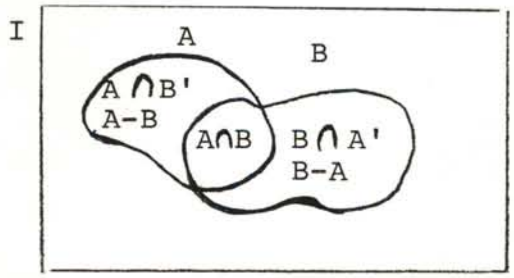
(3) One often extends the concept of complement to that of relative complement. Given any two sets A and B, we define the relative complement of A in B, written B - A, to be all B's which are non-A's. That is

$$B - A = \{x: \ x \varepsilon B \text{ but } x \not\varepsilon A\}$$

Combining this with our observation in Remark (1) we see that

$$B - A = B \cap A'$$

In terms of circle-diagrams:

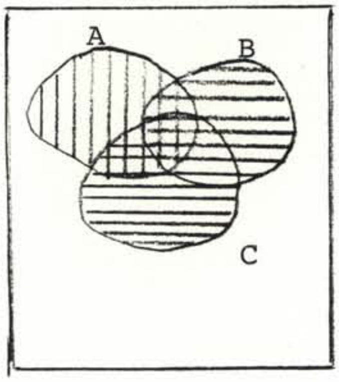


$A \cup B =$
$(A \cap B')\cup(A\cap B)\cup(B\cap A')$
(why?)

In terms of a basic structure, observe that arithmetic seems to hinge on rules of combination that are closed. In this sense, we see that we are in a good position to introduce the arithmetic of sets, for we have our three basic operations of union, intersection and complement whereby we can form new sets from old ones. For example, if A, B, and C are sets, we can form the new set $D = B \cup C$ and then form

$A \cap D$. That is, we can talk about such "combinations" as $A \cap (B \cup C)$

In terms of circle-diagrams:



A is denoted by |||

$B \cup C$ is denoted by ≡

$\therefore A \cap (B \cup C)$ is denoted by ▦

Example:

Let $I = \{1,2,3,4,5,6,7,8,9,10\}$

$A = \{1,2,3,5,7,9\}$

$B = \{1,3,5,6,8,9\}$

$C = \{1,2,3,4,5,7\}$

Then:

$A \cup B = \{1,2,3,5,6,7,8,9\}$

$A \cap B = \{1,3,5,9\}$

$(A \cup B) \cap C = \{1,2,3,5,7\}$

$(A \cup B)' = \{4,10\}$

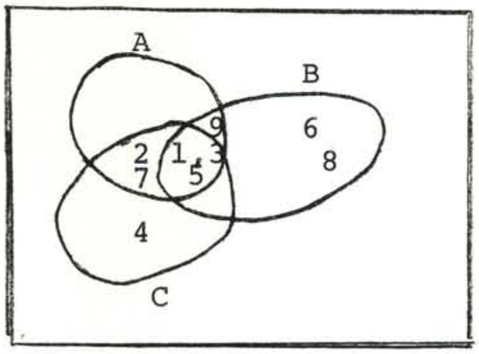
$A' = \{4,6,8,10\}$

$B' = \{2,4,7,10\}$

$A' \cup B' = \{2,4,6,7,8,10\}$

$A' \cap B' = \{4,10\}$

In terms of circle diagrams, we could have represented the above problem as follows:



Before concluding this example, it is worth observing that certain "natural" things do not seem to be true. For example, while one might "suspect" that $(A \cup B)' = A' \cup B'$, the above example shows this to be false. On the other hand, in terms of the above example, $(A \cup B)' = A' \cap B'$ seems to be true. We shall be interested in those things that are true for all sets, not just for special cases. Thus, we shall be interested in such statements as: For all sets A and B, $(A \cup B)' = A' \cap B'$. That is, we are interested in finding universally true recipes about sets. (For example, in the case of ordinary arithmetic, it turns out that 1 x (2 + 3) = (1 x 2) + 3; yet this is not true in general. That is, we can find numbers a, b, and c for which a x (b + c) is not a synonym for (a x b) + c.)

B. Contrast Between Addition and Union

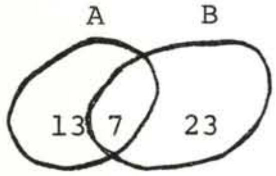
In many ways, especially since union seems to incorporate the idea of COMBINING sets, there is a tendency to associate union with addition. That is, suppose we have two sets A and B and we know the number of members in each set. Say A has m members and B has n members. Then it is not necessarily true that $A \cup B$ has $m + n$ members.

As a particular example, let us consider the case wherein A has 20 members and B has 30 members. We then wish to ascertain the number of members in $A \cup B$. Let us first introduce the following notation:

> If X denotes any set, let us use the notation N(X) to denote the number of elements that belong to X.

In other words, for the problem we are now describing, we wish to compute $N(A \cup B)$, knowing that $N(A) = 20$ and $N(B) = 30$.

Often, since the first impulse is to think of union in terms of addition, we decide that $N(A \cup B) = 50$, since $20 + 30 = 50$. We do not deny that $20 + 30 = 50$! The point is that it is not clear that we really want to add 20 and 30 in this problem. Why? Well, for example, suppose the college registrar finds that 20 students have enrolled in Math 209 and that 30 have enrolled in Bio 232. The registrar records the names of all these students. It should not be difficult to see that the resulting list will contain 50 <u>different</u> names if and only if <u>no</u> student takes both Math 209 and Bio 232. However, for the sake of demonstration, let us assume that 7 students take both of the courses. Then the list would contain only 43 different names. To see this, let A denote the set of students taking Math 209 and let B denote the set of students taking Bio 232. In terms of a circle-diagram, we see that:



The diagram shows us that there are 13 elements that belong to A but not B, 7 that belong to both, and 23 which belong to B but not A. In all, there are 43 elements, despite the fact that, separately, A has 20 members and B has 30. The controlling factor is $A \cap B$, for if an element belongs to both A and B, it contributes 1 to N(A) by virtue of belonging to A; and 1 to N(B) by virtue of belonging to B, even though it is just one element. In summary, each element in $A \cap B$ is counted twice when we form N(A) + N(B); but it is only one element of $A \cap B$. For example, in the problem we just considered, notice that the difference between the correct answer (43) and what may have been the impulsive answer (50), is exactly the number of elements in the intersection of the two sets. That is, we have 50 - 43 = 7.

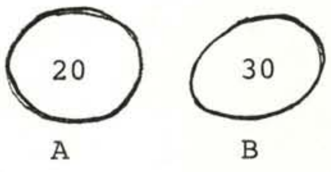
It thus appears that our recipe should have been

$$N(A \cup B) = N(A) + N(B) - N(A \cap B).$$

Applying this result to our problem, we have

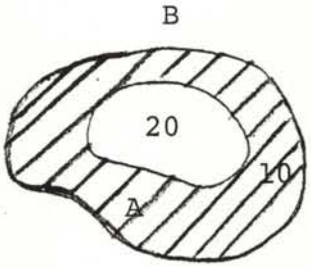
$$N(A \cup B) = 20 + 30 - 7 = 43,$$

which agrees with the proper result. To carry our example one step further, the given information that $N(A) = 20$ and $N(B) = 30$ does very little to determine $N(A \cup B)$. We do know that $N(A \cap B)$ is between 0 and 20, but little else. These two extreme cases correspond to the events that (1) A and B share no elements in common, and (2) A is a subset of B. Pictorially:



$N(A \cup B) = 50$

(1)



$N(A \cup B) = 30$

(2)

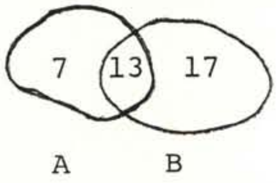
In terms of our recipe, these cases lead to

(1) $N(A \cup B) = 20 + 30 - 0 = 50$

and

(2) $N(A \cup B) = 20 + 30 - 20 = 30$

In other words, with regard to the present problem, unless $N(A \cap B)$ is given, we can only conclude that $N(A \cup B)$ is at least as great as 30, but, in no event, in excess of 50. Moreover, we can obtain a correct answer between 30 and 50 merely by appropriately choosing the value for $N(A \cap B)$. In general, for this problem, we need only let $N(A \cup B) = 50 - N(A \cap B)$. For example, if we wish that $N(A \cup B) = 37$, we let $N(A \cap B) = 13$. Thus:



In summary, the trouble with writing $N(A \cup B) = N(A) + N(B)$ is that A and B need not be mutually exclusive. That is, it may happen that $A \cap B \neq \emptyset$. In the above diagram, we ran into no trouble when we added 7, 13, and 17, but this was because the regions in question were mutually exclusive. In other words, as we shall speak more about later, if X, Y, and Z are mutually exclusive, in pairs (that is, $X \cap Y = Y \cap Z = X \cap Z = \emptyset$) then $N(X \cup Y \cup Z) = N(X) + N(Y) + N(Z)$. With regard to the above diagram, think of $X = A \cap B'$, $Y = A \cap B$, and $Z = A' \cap B$. Then, $N(A \cup B) = N(X \cup Y \cup Z) = N(X) + N(Y) + N(Z)$, since in this case X, Y, and Z are mutually exclusive in pairs.

So far, we have made no restriction as to the finiteness of the sets under consideration. To avoid ambiguity and/or misinterpretation, we shall now impose the restriction for the remainder of this section that all sets under consideration be finite. To see why, let us consider the following situation. Look at the expression 5 - 3. We viewed 5 - 3 as the number which must be added to 3 to yield 5. From a physical interpretation point of view, we might have viewed 5 - 3 as the process of deleting three tally marks from a collection of five. More generally, in terms of sets, suppose that $B \subset A$ and that $N(B) = 3$ while $N(A) = 5$. Then 5 - 3 could be viewed as being the number of members in the set that results when B is deleted from A. (Recall that this set is called A - B, where A - B is merely another name for $A \cap B'$). In general, then, if $B \subset A$ then we can view $N(A) - N(B)$ as being $N(A - B)$. The problem occurs if B is a subset of A where A is an infinite set; for in this case, it is not so easy to describe the number of members in A - B.

For example, let A denote the set of whole numbers. Then A is certainly an infinite set. Now let B denote the set of even whole numbers. Then it is clear that B is an infinite set which is also a subset of A. In this case, A - B would be the set of odd whole numbers, which is an infinite set; hence N(A - B) would be infinite. On the other hand, suppose B were the set of all whole numbers greater than 10. That is: $B = \{11, 12, 13, 14, 15, 16, 17, 18, \ldots\}$ In this case, B would also be an infinite subset of A. If we now delete B from A to form A - B, we find that:

$$A - B = \{1,2,3,4,5,6,7,8,9,10\} = J_{10};$$

or

$$N(A - B) = 10.$$

In summary, if A and B are infinite sets, and no further specifications are made, then we cannot, without the risk of misinterpretation, give a well-defined definition of N(A) - N(B).

In other words, the formula:

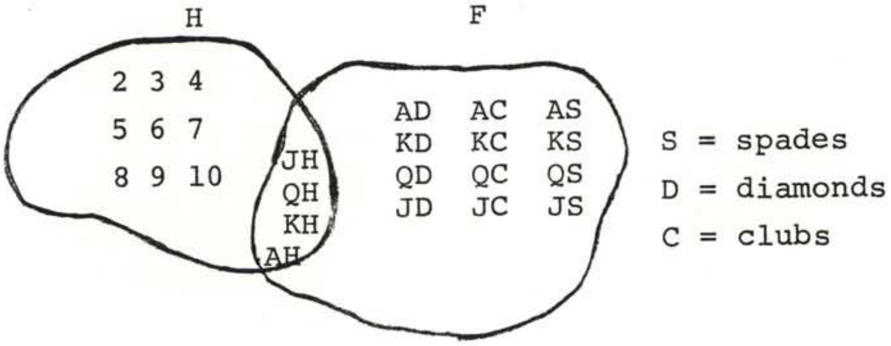
$$N(A \cup B) = N(A) + N(B) - N(A \cap B)$$

becomes "troublesome" if A and B are infinite sets since (1) we can't be sure whether $A \cap B$ is finite or infinite, and (2) if $A \cap B$ is infinite, we must in effect subtract "infinity" from "infinity," and, as mentioned above, this is not well-defined. At any rate, for these reasons, we shall only consider finite sets in this section.

Let us return to

$$N(A \cup B) = N(A) + N(B) - N(A \cap B)$$

By way of additional drill, consider the following situation. We have a standard deck of playing cards (52 cards). We define the face cards to be: Ace, King, Queen and Jack. We are to reach into the deck and randomly extract a card. What is the likelihood (probability) that we chose either a face card or a heart? Well, there are 13 hearts in the deck and 16 face cards. 13 + 16 = 29. Hence, since there are 52 cards in the entire deck, and since 52 - 29 = 23, it appears that the likelihood of drawing either a heart or a face card is 29 chances in favor, to 23 chances against. However, we have made an error if this is our chain of reasoning. Certainly, we do not deny that 13 + 16 = 29, or that 52 - 29 = 23. We merely point out that we should not perform these operations in arriving at the correct answer. Why? To begin with, there are four cards that are counted both as hearts and as face cards, namely, the Ace, King, Queen and Jack of Hearts. In terms of a circle-diagram, letting H denote the set of Hearts and F the set of face cards:



More abstractly, let H and F be as in our diagram; then, the answer to our problem is represented by $N(H \cup F)$. We see that $N(H) = 13$, $N(F) = 16$, and $N(H \cap F) = 4$. The formula now reads: $N(H \cup F) = 13 + 16 - 4 = 25$.

In other words, an "intuitive" count might make one think that he has 29 chances out of 52 of accomplishing his objective, while a "proper" count shows that the chances are only 25 out of 52. As we said, we shall discuss this idea more later; for now we wish only to point out one more use of the knowledge of the theory of sets in solving problems in other branches of mathematics.

It is next our endeavor to extend these results beyond the intersection and union of two sets. For example, suppose we have three sets A, B, and C, and we wish to compute

$$N(A \cup B \cup C).$$

We shall show that

$$N(A \cup B \cup C) = N(A) + N(B) + N(C) - N(A \cap B) - N(A \cap C) - N(B \cap C) + N(A \cap B \cap C)$$

In line with our objection to the subjectivity of intuition, we shall not call the above result self-evident, but rather we shall show a few ways of visualizing this result:

(1) Suppose that A, B, and C were three lists containing names of people, and we wished to amalgamate the three lists into one, but never count the same name twice. Then if we just "stapled" the lists together we would see that:

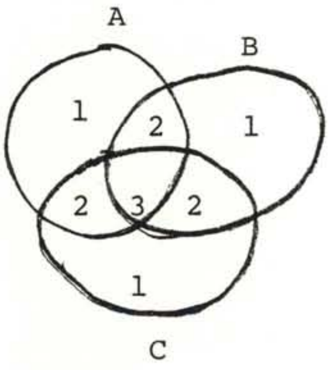
(i) If a name appeared on exactly one of the lists, it was counted the correct number of times.

(ii) If the name appeared on exactly two of the lists, it was counted once too many, and hence should be subtracted once.

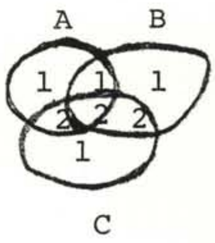
(iii) If the name appeared on all three lists, then it should be subtracted twice.

Thus, for example, if an element belongs only to A, it is counted only once in the sum $N(A) + N(B) + N(C)$. If it belongs to just B and A it is counted twice in the sum $N(A) + N(B) + N(C)$, but it is subtracted once in $N(A \cap B)$. Finally, if the element belonged to A and B and C, it is counted three times in $N(A) + N(B) + N(C)$, then subtracted out three times in-$N(A \cap B) - N(A \cap C) - N(B \cap C)$. Now, however, it isn't counted at all, hence, we add, $N(A \cap B \cap C)$ which counts it once.

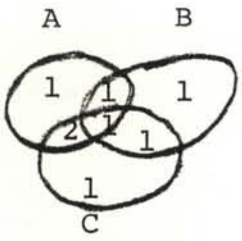
(2) In terms of circle-diagrams, let us indicate by 1, 2, or 3 the number of times an element is counted in arriving at the sum $N(A) + N(B) + N(C)$. Thus:



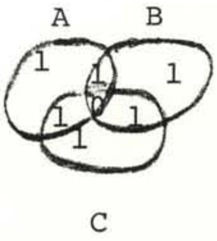
If now we subtract out those that appear in the pairs of intersections, we have



(A $\cap$ B)

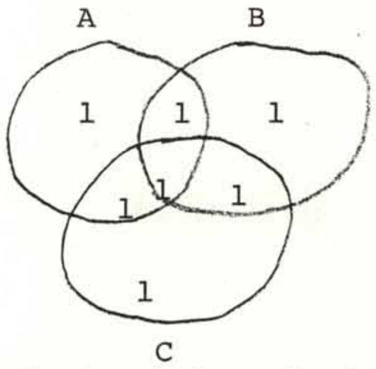


(B $\cap$ C)



(A $\cap$ C)

If we then add in those that belong to all three sets, we have



and this is precisely what we desire, namely, each element is counted once, no matter where it appears.

For more than three sets, there is a rather interesting pattern that prevails, which we present without proof. For example:

$$\begin{aligned} N(A \cup B \cup C \cup D) = {} & N(A) + N(B) + N(C) + N(D) \\ & - [N(A \cap B) + N(A \cap C) + N(A \cap D) + N(B \cap C) + \\ & \qquad N(B \cap D) + N(C \cap D)] \\ & + [N(A \cap B \cap C) + N(A \cap B \cap D) + N(A \cap C \cap D) + \\ & \qquad N(B \cap C \cap D)] \\ & - N(A \cap B \cap C \cap D) \end{aligned}$$

The pattern is that we alternately add and subtract all possible combinations of intersections ranging from taking the sets one at a time to all at one time.

We conclude this section with an example:

In a certain school it is required to take at least one of the three languages, French, German, or Spanish, in order to graduate. In a certain graduating class, we find that 20 students took all three languages, 35 took French and German, 40 took both French and Spanish, 50 took both German and Spanish, 90 took French, 80 took German, and 110 took Spanish. How many were in the graduating class?

Solution:

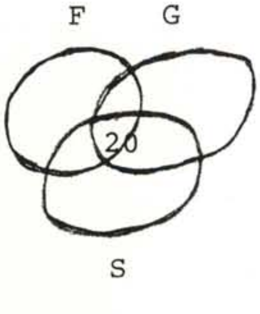
We must be careful to curb our enthusiasm and not add the given numbers. For if we do, we count certain students more than once. One solution, letting F, S, and G denote the set of students taking French, German, and Spanish respectively, is to use our formula with

$N(F) = 90$, $N(G) = 80$, $N(S) = 110$, $N(F \cap G) = 35$. $N(F \cap S) = 40$, $N(G \cap S) = 50$, and $N(F \cap G \cap S) = 20$. Then, the number in the graduating class is $N(F \cup G \cup S)$ (why?), and we have

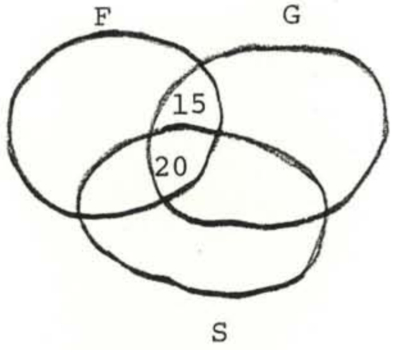
$N(F \cup G \cup S) = 90 + 80 + 110 - 35 - 40 - 50 + 20 = 175$

Thus, there were 175 members in the graduating class. It might have been more intuitive had we used circle-diagrams. In this event:

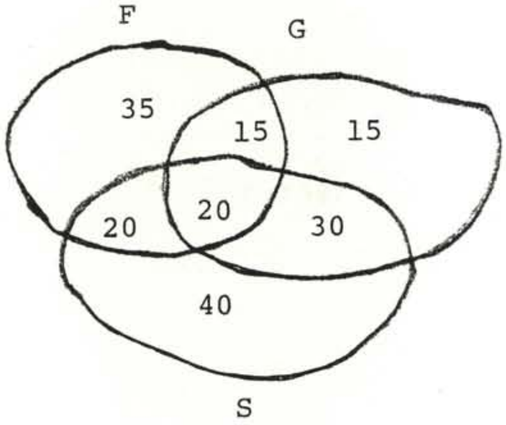
(i) since 20 take all three



(ii) then since 35 take both French and German (be careful; this number includes the 20 who take all three)



(iii) continuing in this way (the details are left to the reader)



Not only does this give us the same answer, but since the regions in the diagram are mutually exclusive, we can pick off such other results as: There were 15 members of the graduating class who took German but neither French nor Spanish.

Chapter III
AN INTRODUCTION TO FUNCTIONS AND GRAPHS

A. Introduction

Mathematics has been described as: the study of relationships, the language of science, the basic tool of technology, the logical quest for truth, the study of exact measurement. More subjectively, it has been called a strict discipline, a way of life, and a philosophy. In a sense, the attempt to answer the question, "What is Mathematics?" reminds one of the fable concerning the blind men and the elephant. Each man touched a different part of the animal and, depending on which part he touched, each gave a different description of the elephant.

From an engineering point of view, however, one definition seems particularly appropriate. Namely:

Mathematics is the study of relationships.

Certainly this is a true statement. In geometry, for example, one studies the relationship between the area of a region and its various demensions. In the traditional "John and Bill"-type algebra problem, one is usually investigating the relationships of greater than and less than. In fact, it is probably true that wherever we turn in mathematics, somehow or other we are concerned with the study of relationships.

However, it should not be difficult to see that to define mathematics as the study of relationships would make virtually every academic endeavor of man a branch of mathematics. For, in physics, Galileo studied the relationship between the distance that an object fell and the time during which it was falling; Newton studied the force of attraction between two objects in relation to their sizes and the distance between them. Such quantitative studies of relationships are well known in all of the physical sciences. Indeed they are the backbone of many investigations. However, an equally important point is that the study of relationships is by no means restricted to mathematics and the physical sciences. For example, the philosopher studies the relationship between a concept and the word used to denote that concept, the economist studies the relationship between various forms of supply and demand, the student of literature studies writing in relationship to the society of the times, the historian judges the success of a particular society in relationship to the aims upon which the society was formed, the psychologist studies the scores in certain tests in relationship to the environmental background of those who took the test.

Such examples are numerous, and the point we are trying to make is that the phrase, "the study of relationships," permeates every, or nearly every, field of human endeavor. Thus, it is in this sense it would seem that such a definition of mathematics would be too comprehensive.

Notice, however, that such a definition of mathematics--even though it does not separate mathematics from other subjects--is rather worthwhile; at least in the sense that it reflects a general trend in one prevalent form of mathematical usage of the day. Namely, more and more subjects, at one time only thought to have a minimal need for mathematics, are beginning to require the studying of relationships in more precise quantitative ways. Such a study brings mathematics into play as a very strong computational tool.

It shall be our aim in this chapter to exploit the idea that mathematics is the study of relationships. While such a definition might not apply to every aspect of mathematics, and while such a definition may be much too comprehensive, the fact is that this aspect of mathematics, perhaps more than any other aspect, has been the unifying thread by which man has tried to explore and to understand the "real" world around him.

In this context, it is particularly easy to introduce the meaning of a FUNCTION. Stripped of all embellishment, a function is a RULE. In the classical sense, it was a rule which assigned to one number, another number. In the modern sense, it is a rule which assigns to an element of one set an element of another set.

We shall study functions from both points of view. In particular, we shall choose to introduce the topic from the modern point of view. While this is not chronologically correct, we prefer the generality of the modern approach. Afterwards, we shall look into the classical viewpoint.

B. Functions and Sets

Since it is usually difficult to think in abstract terms, let us pave the way for our present discussion by introducing a trivial but concrete environment.

Consider a collection of salesmen (which we shall call set A) who are having a convention at a particular hotel (the rooms of which we shall call set B). The entire hotel has been reserved for the salesmen and each salesman will reside at the hotel for the duration of the convention. As each man enters the hotel, he is assigned to a room by the room-clerk.

In the discussion that follows, we shall illustrate all of our definitions in terms of the above "situation".

DEFINITION 1:

Let A and B denote sets. Then, by a function, f, from A to B, written $f:A \to B$, we mean that f is a rule which assigns to each element $a \in A$ an element $b \in B$. The fact that f assigns $a \in A$ to $b \in B$ is denoted by $f(a) = b$ (read as: "f of a equals b")

In terms of our illustration, the room-clerk plays the role of the function from A to B; that is, he is a "rule" which assigns to each element of A (each salesman) an element of B (a hotel room).

DEFINITION 2:

> If $f:A \rightarrow B$ then we call A the domain of f (abbreviated as dom f or D_f) while B is called the range of f (also written R_f).

Again, in terms of our illustration, our domain is the salesmen, and our range is the hotel rooms.

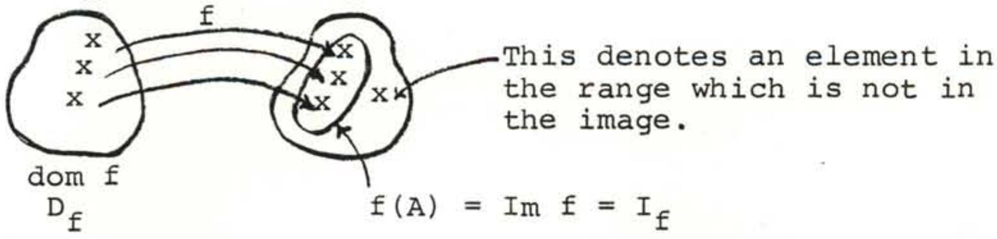
To motivate our next definition, let us observe that it is possible that the salesmen do not completely fill the hotel rooms. In this sense, our notion of range is a bit deceptive since it gives us no hint as to how much of the range is "used up" by the function. In still other words, we might be more interested in knowing what rooms are being used by the salesmen than to know that all the rooms being used are in the hotel. At any rate, with this in mind:

DEFINITION 3:

> Given $f:A \rightarrow B$, we define the image of f (usually abbreviated by Im f or I_f) to be the set: $\{f(a):a \in A\}$. This set is also denoted by f(A).

In other words, f(A) is precisely that subset of B that is "used up" by f. More precisely, if $f(a) = b$, we often call b the image of a (with respect to f). In the context, f(A) denotes that subset of B which consists of those elements of B which are images of elements in A (with respect to f). In terms of our illustration f(A) is the set of rooms to which salesmen are actually assigned.

In terms of a diagram:



This denotes an element in the range which is not in the image.

Of course, it's possible that by the time all the salesmen check in they have managed to use up all the hotel rooms. That is, nothing excludes the possibility that if $f:A \rightarrow B$, then $f(A) = B$. This leads to:

DEFINITION 4:

> Given $f:A \rightarrow B$, we say that f is *onto* B if and only if $f(A) = B$ (otherwise, the usual vocabulary says that f is a function from A *into* B).

In many situations, it is clear that one is more interested in the image of f than in its range. That is, given $f:A \rightarrow B$, we usually wind up concentrating on $f:A \rightarrow f(A)$.

While the distinction between onto and into (or equivalently, between range and image) is important, it turns out that in many "real-life" situations the problem takes care of itself. What happens in these cases is that we usually have some set A and we define some rule which we apply to members of A *WITHOUT EXPLICITLY MENTIONING ANY OTHER SET* B. For example, suppose we let $A = \{x:2 \leqslant x \leqslant 3\}$. Suppose next we decide to square each member of A. In this respect, we are actually defining a function, say f, on A in which $f(x) = x^2$ for each $x \in A$.

Notice that while we haven't said it in so many words, we have induced a set B which is the image of f. That is, $f(A) = \{f(x):x \in A\}$ or $f(A) = \{x^2:2 \leqslant x \leqslant 3\}$. Without verifying the details it is not hard to conclude that f(A) in this case is precisely the set $\{y:4 \leqslant y \leqslant 9\}$ since as x takes on all values from 2 to 3, x^2 takes on all values from 4 to 9. In any event it would now seem natural to let $B = \{x:4 \leqslant x \leqslant 9\}$* and in this case, we have $f:A \rightarrow B$.

*Notice that $\{x:4 \leqslant x \leqslant 9\}$ and $\{y:4 \leqslant y \leqslant 9\}$ denote the same set. The name of the generic element (in this case, $x \cap y$) is irrelevant.

By our very construction, f is onto and thus it really makes no difference whether we refer to B as the range or the image of f. For this reason it seems to be an accepted practice (although we don't condone it) to use the words range and image interchangeably. The important point is that they are different concepts, but in the case that B is to be "induced" from A and f, it is most natural to choose B to equal f(A).

Returning once again to our salesmen, it is reasonable to assume that we don't assign a salesman to more than one room. However, it is equally reasonable that we might assign more than one salesman to the same room.

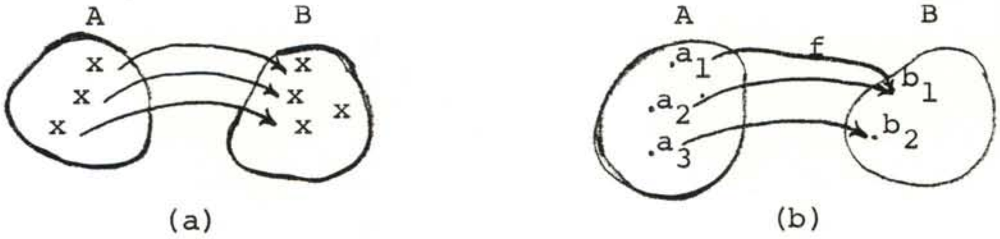
This leads us to still another basic definition:

DEFINITION 5:

The function $f: A \to B$ is called one-to-one (often written as 1-1) if no element of B is the image of more than one element of A. More precisely, f is 1-1 means that if a_1 and a_2 are elements of A then $a_1 \neq a_2$ implies that $f(a_1) \neq f(a_2)$; or from a different emphasis:

$$f(a_1) = f(a_2) \text{ implies that } a_1 = a_2$$

Pictorially, Definition 5 has the following translation:



In (a) f is depicted as 1-1. The point is that no member of B is the image of more than one member of A. In terms of a numerical example let R denote the real numbers and consider $f: R \to R$ given by $f(x) = 2x$ for each $x \in R$.

Recall that $f(x_1)$ denotes the image of x_1 while $f(x_2)$ denotes the image of x_2. In this example, $f(x_1) = 2x_1$ while $f(x_2) = 2x_2$. Then $f(x_1) = f(x_2)$ means that $2x_1 = 2x_2$, and this can only happen if $x_1 = x_2$. Thus the function f defined by $f(x) = 2x$ is a 1-1 function. (In "plain" English, unequal numbers have unequal doubles.)

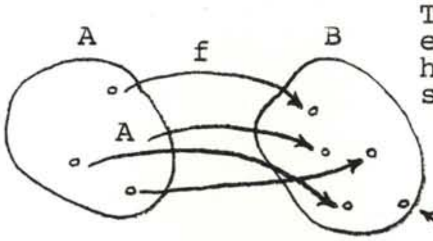
As for (b) we see here an example of a function which is not 1-1. In particular, b_1 is the image of both a_1 and a_2 in our diagram. In terms of an example, if we define f by $f(x) = x^2$ we see that f is not 1-1 since x^2 denotes the square of both x and -x. In other words, for $x \neq 0$, $f(x) = f(-x) = x^2$ but $x \neq -x$. It is worth noting here that the idea of 1-1 depends on the <u>domain</u> of the function. For example, in terms of an earlier illustration, let $A = \{x:2 \leqslant x \leqslant 3\}$ and let $B = \{x:4 \leqslant x \leqslant 9\}$. If we define f by $f(x) = x^2$, we see that $f:A \rightarrow B$ is 1-1. In this case while $x^2 = 4$ for both $x = 2$ and $x = -2$, the fact remains that $-2 \notin$ dom f (that is, $-2 \notin A$). We will have more to say about this in a later section

While we shall also say more about the following topic later, a few words might still be in order now. Notice that in our definition of $f:A \rightarrow B$ we <u>insisted</u> that f assign to each member of A <u>a</u> (meaning <u>one</u> <u>and</u> <u>only</u> <u>one</u>) member of B. In the classical treatment of functions no such restriction was ever made. By way of an example which we shall explore in more detail later, consider the idea of taking the square root of a number. In the old days, one said that the square root of 4 was 2 or -2 since either 2 or -2 had its square equal to 4. However to say that $\sqrt{4} = \pm 2$ would not allow the square root to be considered a function by our "modern" definition since the square root assigns to 4 <u>two</u> real numbers. To avoid this problem, we could have let, say, R^+ denote the non-negative real numbers and then defined $f:R^+ \rightarrow R^+$ by $f(x) = \sqrt{x}$ for all $x \varepsilon R^+$. In this event $f(4) = \sqrt{4}$ would be only 2 since -2 doesn't belong to the image of f.

At any rate the point we are trying to make is that what we call a function by modern terminology would have been called a SINGLE-VALUED function in the classical terminology. In the classical terminology if a function were not single-valued, it was called MULTI-VALUED. In the modern language there is no analog of a multi-valued function. As we have said, we shall have more to say about this later, but for now let us accept the fact that we shall, unless otherwise specified, use function to mean single-valued function.

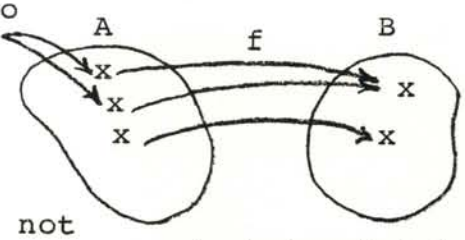
The important thing to keep in mind is that single-valued and one-to-one are entirely different concepts. All functions are, by our definition, single-valued, but not all are 1-1.

The following diagrams illustrate pictorially the meaning of 1-1 and onto. Observe that a particular function can be both, neither, or one but not the other. Thus:



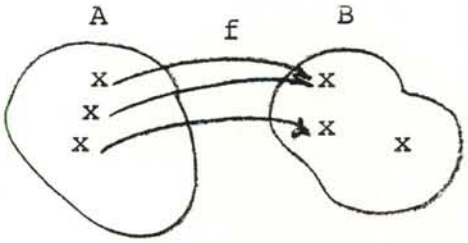
1-1 but not onto

This is not the image of anything in A.

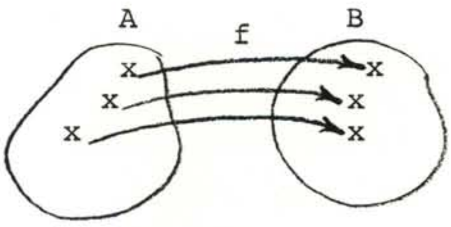


These two elements have the same image

onto but not 1-1



neither 1-1 nor onto



Both 1-1 and onto

We would like to conclude this section by introducing just enough more definitions so that we can begin to view the study of functions as a mathematical structure.

In any mathematical system we must have a criterion for equality* so that we may distinguish between the elements of our system. To this end:

DEFINITION 6:

Given two functions f and g, we say that $f = g$ if and only if:

(1) dom f = dom g

and

(2) $f(a) = g(a)$, for each $a \varepsilon A$ (where A = dom f [or g])

That is:

(1) We insist that we don't even begin to compare functions unless they "operate on" the same set of elements.

(2) Once the domains are the same, we insist that the images be the same, element for element.

Referring to our salesmen-hotel model, observe that different groups of salesmen can come to the hotel at different times. Moreover, for a given group of salesmen there is more than one way for the room clerk to assign rooms. If we now talk about equal functions (that is, in terms of our model - equal room assignments) we insist (1) that each assignment involves the same set of people and (2) each person who receives a room by one assignment receives the same room by the other assignment.

Notice especially well that we require more than just that the same rooms be used in each assignment. In still other words to say that for each $a \varepsilon A$, $f(a) = g(a)$ says much

*See Appendix I if more details about equality are desired.

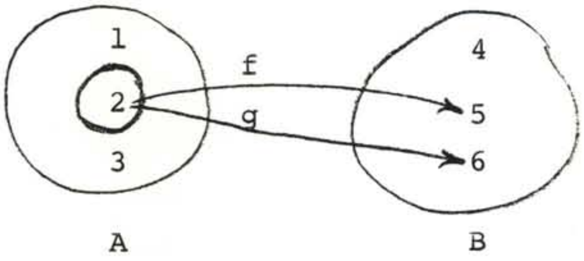
more than just saying that f(A) = g(A). In terms of a more specific illustration let A = {1,2,3} and let B = {4,5,6}. Define f:A → B by:

$$f(1) = 4$$
$$f(2) = 5$$
$$f(3) = 6$$

and define g:A → B by

$$g(1) = 4$$
$$g(2) = 6$$
$$g(3) = 5$$

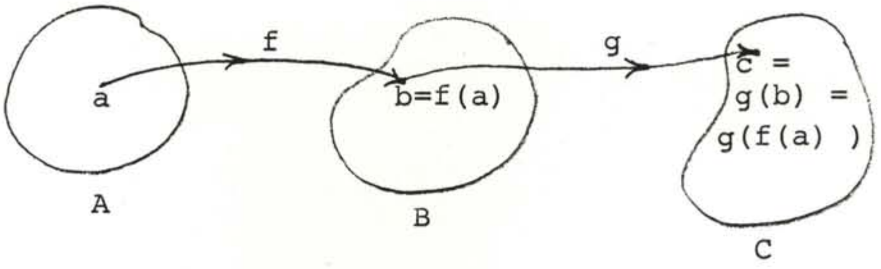
In particular, since both f and g are onto, we have that f(A) = g(A) = B. However, f ≠ g; for while they have the same domain, f(2) ≠ g(2). Pictorially:



Our next quest is to define how we may combine functions to form other functions. To this end, suppose f:A → B and g:B → C. f and g can be composed so as to "induce" a function from A to C. For example, we can start with a ε A and then look at f(a). This is an element of B. Let's call it b (that is, b = f(a)). Now by definition of g, g maps b into some element of C, say c. That is, g(b) = c. Putting these two separate operations into one symbol (specifically, by replacing b by f(a)), we obtain:

$$g(f(a)\ ) = c$$

In terms of a picture:

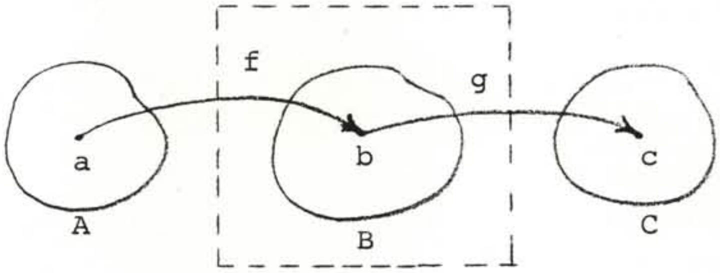


This leads to

<u>DEFINITION 7</u>:

Let f:A → B and g:B → C be given. Then by the composition of f and g, written g∘f, we mean the function h:A → C such that for each a ε A, h(a) = g(f(a)).

In terms of the previous diagram:



h = g∘f

Let us make a few comments about this last definition.

(1) Observe that by our definition g∘f has been defined so that its domain is the domain of f and its image is the image of g. In this sense, then, we must be very careful not to confuse g(f(x)) with f(g(x)). For example, by definition of g, g(x) belongs to C, but C is not necessarily the domain of f. In other words, f(g(x)) might not even make sense, if f:A → B and g:B → C since we can't be sure that g(x) ε A.

(2) While our definition refers to the sets A, B, and C, nothing excludes the possibility that A = B = C. In this event, the problem described in (1) cannot occur since the domain and range of each function is then the same set, say, A. Even then, however, we must not confuse f∘g with g∘f. By way of illustration, let A = {1,2,3} and define f and g as follows:

$$\begin{array}{ll} f(1) = 1 & g(1) = 2 \\ f(2) = 3 & g(2) = 1 \\ f(3) = 2 & g(3) = 3 \end{array}$$

Then;

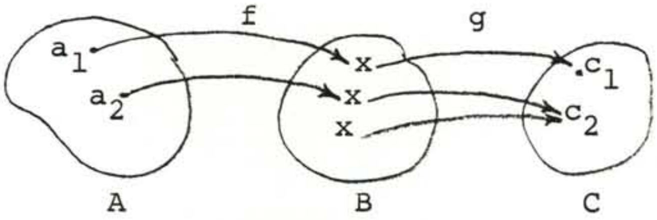
$$\begin{aligned} g(f(1)) &= g(1) = 2 \\ g(f(2)) &= g(3) = 3 \\ g(f(3)) &= g(2) = 1 \end{aligned}$$

while:

$$\begin{aligned} f(g(1)) &= f(2) = 3 \\ f(g(2)) &= f(1) = 1 \\ f(g(3)) &= f(3) = 2 \end{aligned}$$

Thus f∘g and g∘f are both 1-1 and onto functions from A to itself, but they are not equal. (Why?)

(3) If f:A → B and g:B → C are both 1-1 and onto then g∘f is also 1-1 and onto from A to C. This is an important result, but it is not as self-evident as it might seem - at least, in the sense that its converse is not true. That is, g∘f can be 1-1 and onto even though not both f and g are. An example is shown in the diagram below.



Letting h = g∘f, we see that $h(a_1)=c_1$ and $h(a_2)=c_2$ so that h= g∘f is indeed both 1-1 and onto

Here g∘f is both 1-1 and onto
$(g(f(a_1))=c_1$ and $g((fa_2))=c_2)$
But f is not onto and g is not 1-1

In the next section, we shall explore a few properties of functions which are both 1-1 and onto.

While our comments about composition of functions are self-contained within the present context, we should mention that such compositions play a very important role in the study of calculus. Without delving into the calculus, the idea is that in many physical situations we have a variable y which is expressed in terms of (as a function of) a second variable u, and u, in turn, is expressed in terms of a third variable x. We then wish to study y in terms of x. More symbolically, we might have that $y = f(u)$ and $u = g(x)$; then $y = f(g(x))$. Another version of this idea is in the guise of what is known as PARAMETRIC EQUATIONS. Sometimes we wish to study the relationship between y and x when each is expressed in terms of another variable (parameter) u. This is precisely equivalent to our earlier example in the sense that if u is expressed in terms of x, the implication is that we may view x as being expressed in terms of u (although there are a few sticky problems that we will discuss shortly). A common illustration might be that we want to study the relationship between the velocity and the acceleration of a particle when all we have explicitly is how each behaves as a function of time.

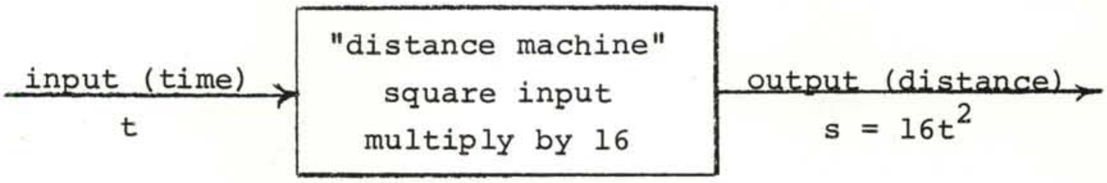
C. Application of Functions to Real Numbers

The study of calculus or, more generally, mathematical analysis begins with the special case in which both the domain and the range of our functions are subsets of the real numbers. Before pursuing this idea further it might be best to illustrate our ideas in terms of a concrete, practical situation - especially in the light of how abstract our earlier remarks about functions may have seemed.

Let us consider the classic problem of Galileo, concerning a freely falling body near the surface of the earth in the absence of air resistance. Recall that, in this case, the body falls a distance of s feet at the end of t seconds according to the rule

$$s = 16t^2 \tag{1}$$

Notice that we can interpret (1) as defining a function whose domain is a set of numbers (representing time) and whose image is a set of numbers (representing distance). More pictorially, we may think of a "distance machine" where the inputs are time and the outputs are distances, and the machine generates an output for a given input by squaring the input and multiplying the result by sixteen. Thus:



In referring to (1), the traditional language is to say that s is a function of t, and to write:

$$s = f(t) = 16t^{2*} \tag{2}$$

*Since the letter f in no way seems to suggest either time or distance, the convention used in many books is to write $s(t) = 16t^2$, where s(t) is an abbreviation for saying that s is a function of t; and for a given t, s is determined by $16t^2$.

With respect to (2), one often refers to t as the independent variable and to s as the dependent variable. To correlate this language with our function-machine, observe that the independent variable plays the role of the input while the dependent variable plays the role of the output. Perhaps an easy way to remember the difference is that the output *depends* on the input.

Of course the entire notion of independent variable versus dependent variable hinges on the fact that one of the variables is *explicitly* expressed in terms of the other. For example, if we wrote

$$x + y = 1 \tag{3}$$

neither x nor y is explicitly stated in terms of the other, even though there is clearly an implicit relationship between the two variables. That is, once either x or y is given, equation (3) determines the value of the other.

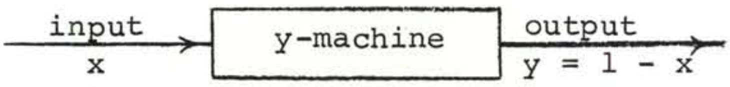
Notice that we can rewrite (3) in either of the following equivalent forms:

$$y = 1 - x \tag{4}$$

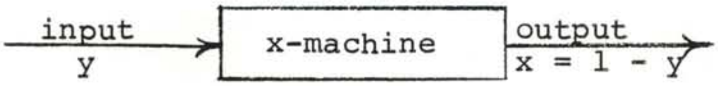
or

$$x = 1 - y \tag{5}$$

If we use (4), we would refer to x as the independent variable and to y as the dependent variable, and we might write $y(x) = 1 - x$. The associated function machine would be:



Similarly, if we use (5), we would refer to y as the independent variable and to x as the dependent variable. We would also write $x(y) = 1 - y$, and the function machine would be given by:



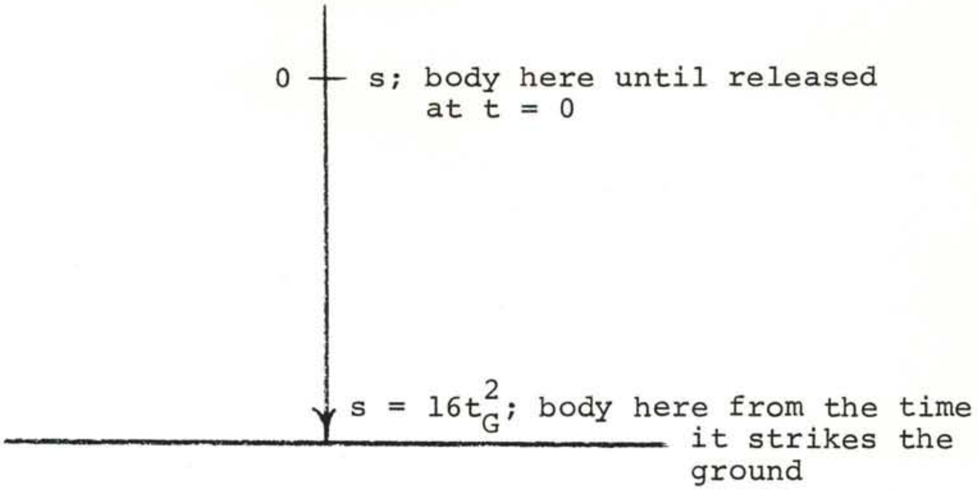
It should be noted, however, that not only does the role of independent and dependent variable depend on how we elect to resolve the implicit relationship between the two given variables, but also that there are many times when it is eather extremely difficult or even impossible to find one variable explicitly in terms of the other from a given implicit relation. For example, consider

$$x^8 + x^6y^4 + y^6 = 3$$

To find y as an explicit function of x requires that we be able to solve a sixth degree polynomial equation in y - a feat which, <u>if</u> possible, would be extremely tedious. Similarly, to find x in terms of y would require our solving an eighth degree polynomial equation in x.

More important at this stage than which is the dependent and which is the independent variable is the question concerning the domain and image of our function. For example, if we return to equation (1) it is quite clear that any real value of t will yield a non-negative real value for s. "Inversely," we can show that for each non-negative value of s there is at least one real number t such that $s = 16t^2$. In fact, by elementary algebra we see that $s = 16t^2$ implies that $t = \pm\sqrt{\frac{s}{16}}$. In still other words, if $s = 0$, $t = 0$, and if $s > 0$ then $t = \pm\sqrt{\frac{s}{16}}$ (and if $s < 0$, $\sqrt{\frac{s}{16}}$ is imaginary). Notice that $t = \pm\sqrt{\frac{s}{16}}$ violates the modern definition of a function. In the "old days", we would have shrugged things off with the observation that we had a multi-valued function. The implications of not allowing multi-valued functions in modern mathematics are extremely profound but peripheral to our present discussion. Consequently, further discussion of this point is left to the next two sections of this chapter.

In any event, then, from (1) we could justifiably deduce that if $f(t) = 16t^2$ then the domain of f is the set of all real numbers, while the image of f is the set of all non-negative real numbers. On the other hand, we know that (1) doesn't tell the whole story. For example, it is physically clear that $s = 0$ until the body begins to fall, and that $s = 16t^2$ applies only for the time that the body is in free fall. More mathematically, if we let t_G denote the time at which the body strikes the ground, then $s = 16t^2$ is true only for the time interval $0 \leqslant t \leqslant t_G$, and once t exceeds t_G then s remains constant. In particular, for $t > t_G$, $s = 16t_G^2$. Pictorially,


0
s; body here until released at t = 0
s = 16t_G^2; body here from the time it strikes the ground

The point we wish to make is that equation (1) most likely was meant to be an abbreviation for

$$s = 16t^2,\ 0 \leqslant t \leqslant t_G \qquad (6)$$

If we rewrite (6) as $f(t) = 16t^2$, $0 \leqslant t \leqslant t_G$, we seem to be saying that the domain of f is the time interval from 0 to t_G, which we write more mathematically as the set $\{t:0 \leqslant t \leqslant t_G\}$. Such a set is also referred to as the *closed* interval from 0 to t_G and is usually abbreviated by $[0,\ t_G]$.

More generally, if a and b are any real numbers such that $a < b$, we define *the closed interval from a to b* to be the set $\{x:a \leqslant x \leqslant b\}$ and we abbreviate this set by writing $[a,b]$.

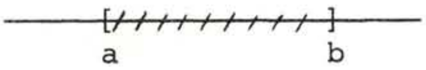
In a similar way, *the open interval from a to b* is defined to be the set $\{x:a < x < b\}$. The open interval from a to b is usually abbreviated by the notation (a,b).

In summary,

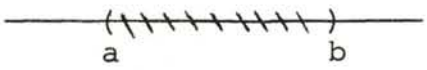
$$[a,b] = \{x:a \leqslant x \leqslant b\}$$

$$(a,b) = \{x:a < x < b\}$$

The only difference between the open and the closed interval is the exclusion (inclusion) of the "end points". Pictorially,



closed interval from a to b } points a and b are included



open interval from a to b } points a and b are excluded

Notice here the difference between a "point" and a "dot". Clearly, there is a difference between [a,b] and (a,b). Yet, if we rely on a picture we cannot see the difference since a point has no thickness. That is, merely from a picture we cannot tell whether the points a and/or b have been deleted. This is why we write $-\underset{a}{[}\!-\!-\!-\!-\underset{b}{]}-$ if we mean the closed interval and $-\underset{a}{(}\!-\!-\!-\!-\underset{b}{)}-$ if we mean the open interval. Just how important the difference is between open and closed intervals hopefully will be made clear as our course progresses.

Notice, too, that one end point could be excluded and the other included. That is, we can talk about half-open or half-closed intervals. For example,

$$[a,b) = \{x : a \leqslant x < b\}$$

$$(a,b] = \{x : a < x \leqslant b\}$$

Our main point is that in virtually every physical situation we encounter, it turns out that a particular relation holds only for some specific (time) interval. In other words, while there is no requirement for the domain of a function to be an interval, it turns out that in most important situations, the domain will be an interval [as in the present illustration exhibited by equation (6)] or else it will be a union of intervals.

In fact, our $s = 16t^2$ example may be viewed as a function that is defined for all real numbers but which behaves as if it were defined on a union of intervals. More specifically, we can summarize some of our previous remarks about the relation $s = 16t^2$ by writing

$$s = \begin{cases} 0 \text{ if } t < 0 \\ 16t^2 \text{ if } 0 \leqslant t \leqslant t_G \\ 16t_G^2 \text{ if } t > t_G{}^* \end{cases} \qquad (7)$$

While (7) might look as if it were defining three different functions, it is in reality defining only one. Namely, for each input, the function machine must decide whether $t < 0$ or $0 \leqslant t \leqslant t_G$ or $t > t_G$. Notice that exactly one of these three conditions must prevail for a given t, and then (7) tells what output the machine will yield for the given input once the input is located in the proper category. In terms of the more pictorial number line we are saying that our domain will usually be a "connected" portion (a line segment) of the number line or perhaps a union of such segments.

With respect to the above remark, observe that we may view both the domain and the image as subsets of real numbers (pictorially as points on the number line). This, in turn, might suggest the use of Cartesian coordinates whereby we could use the x-axis to denote the domain and the y-axis to denote the range.

This results in the concept of the graph of a function.

*Technically speaking, this is not an interval in the sense of our definition of an interval since a and b were always finite numbers. In spirit, however, since an expression such as $x > a$ defines a "continuous" set of x's, it is customary to call such expressions intervals also. To make things look a bit more uniform the mathematician uses the symbol ∞ (infinity) to his advantage by writing $x > a$ as $a < x < \infty$ (and similarly, $x < a$ is written as $\infty < x < a$).

D. Graphs

Even on a subconscious level, we frequently think of non-geometric ideas in terms of a geometric picture. For example, consider the expression "profits rose". The only way profits can "rise" is, for example, if the office safe blows up! Obviously what we mean when we say that profits rose was that profits INCREASED. Why then did we interchange rise (geometric) with increase (arithmetic)? The answer centers around the idea of a GRAPH.

How is the concept of a graph related to the concept of a function? The answer lies, at least for a first approximation, in the concept of ordered pairs (which in turn suggests the [Cartesian] Plane). Namely, our so-called function machine is determined once we know the output for each given input. In short, we "can abbreviate" our function by using ordered pairs, where, for example, the first member of the pair could name the input while the second member of the pair could name the output. In terms of the Cartesian Plane, this says we may use the x-axis to indicate the domain of the function while we may use the y-axis to indicate the range of the functions. In still other words, we are saying that we may use the point $(x,f(x))$ to represent the fact that for an input of x the output is $f(x)$. Conceptually the graph and the function are two entirely different things. One is an analytic relation and the other is a picture of it. What does the picture do for us? Well, for example, suppose we know that for a certain real number x, $f(x)$ is positive (this is an arithmetic statement). But if $f(x)$ is positive, we know that the POINT $(x,f(x))$ lies above the x-axis. In a similar way, if $f(x)$ is negative, the point $(x,f(x))$ is below the x-axis; and if $f(x) = 0$, the point $(x,f(x))$ is on the x-axis. Thus, the idea of a graph replaces the analytic terms "greater than 0", "equal to 0", and "less than 0" by "above the x-axis", "below the x-axis", and "on the x-axis", respectively.

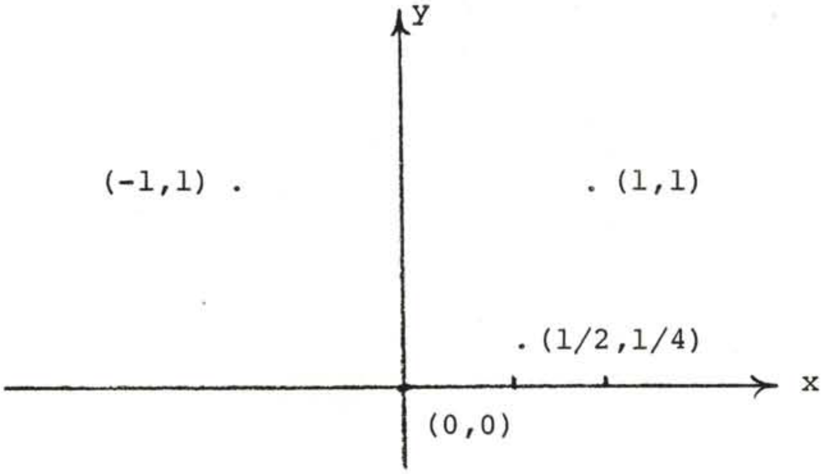
In a similar way, it replaces "increasing" by "rising", "decreasing" by "falling" and "constant" by "horizontal" (that is, if f(x) doesn't vary, the point (x,f(x)) always has the same height above the x-axis. Hence, all such points are parallel to the x-axis or horizontal if we define the direction of the x-axis as being horizontal).

Rather than continue to ramble on in abstractions, let us now turn our attention to a specific situation. For example, let us discuss the function f given by:

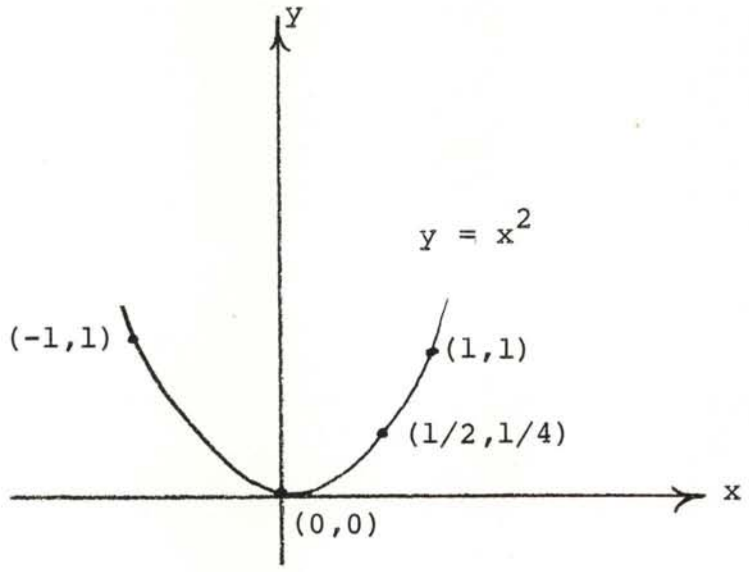
$$f(x) = x^2$$

We know at once that the input x yields the output x^2. This means that in terms of our graph, the input x will give rise to the point in the plane (x,x^2).

Among others then, some of the points on the graph would be (0,0), (1,1), (-1,1), (1/2,1/4), etc. In terms of a picture, we would have:

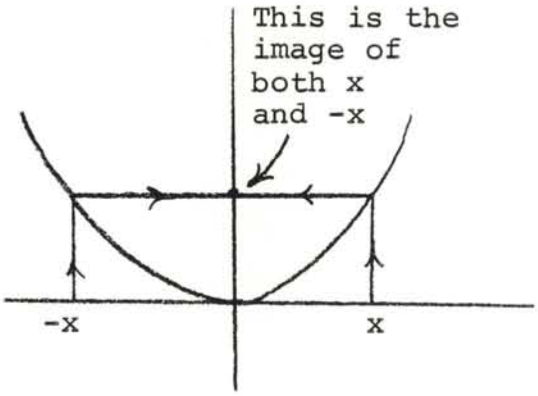


If we now use our "intuition" we might conjecture* that the graph of $f(x) = x^2$ is given by



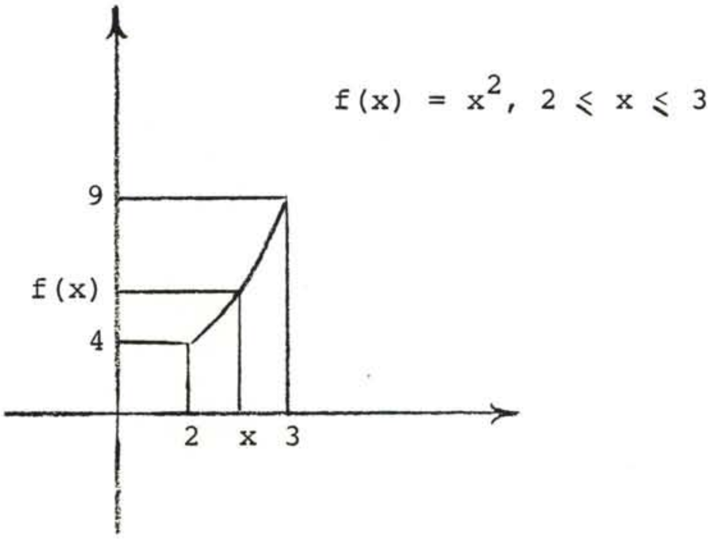
We may study one-to-oneness and single valuedness very nicely in terms of graphs. Namely, if each line parallel to the y-axis and which passes through a point in the domain intersects the graph at only one point then our function is single-valued; and if each line parallel to the x-axis and which passes through a point in the image intersects the graph in only one point then our function is 1-1. By way of illustration, the following diagram indicates that if $f(x) = x^2$ then f is single valued but not 1-1. In fact every positive number in the image is yielded by two members of the domain.

*Actually, we only have a conjecture no matter how intuitive things may seem. That is, as long as we locate points which have spaces between them, we are only conjecturing as to what goes on "in between". We shall have much more to say about this later in the course.

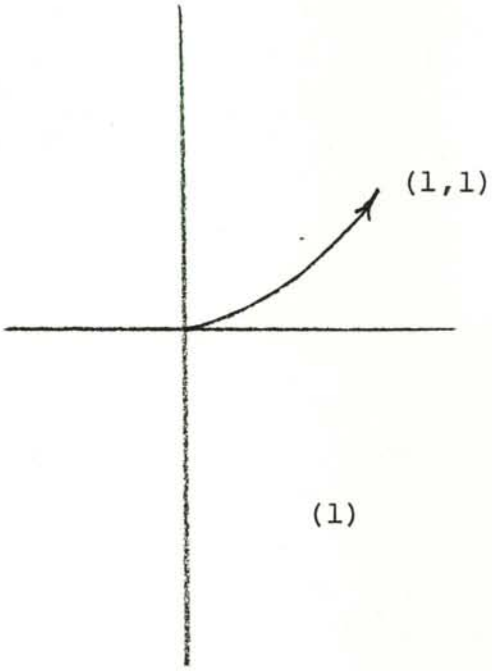


Each value of x yields one, and only one, point on the graph; but each non-zero value in the image comes from two numbers in the domain.

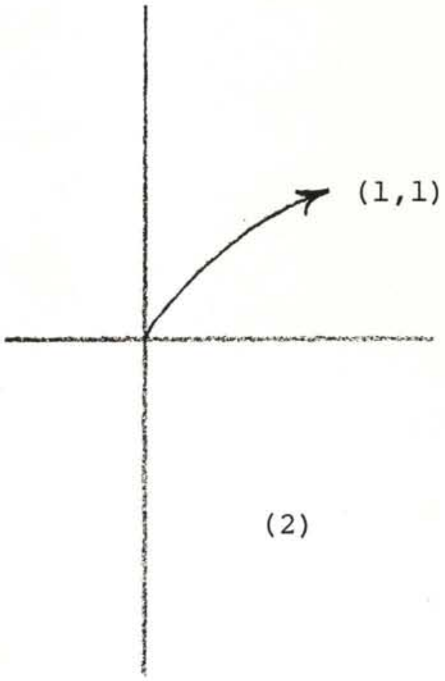
Do not confuse the graph with the domain and the image of the function. Recall that the domain is located as a subset of the x-axis while the image is a subset of the y-axis. For example, referring again to our function f where $f(x) = x^2$, and the domain of f $\{x:2 \leqslant x \leqslant 3\}$. Then the image of f in this case would be the closed interval $\{x:4 \leqslant x \leqslant 9\}$. In fact, our picture also shows us that in this example f is both single valued and 1-1.



While this result in no way depends on the graph, the graph does, however, give us much important information at a glance. For example, the fact that the graph is always rising tells us that the "output" increases as the "input" increases. That is, f(x) is increasing as x increases. Notice also that the graph not only rises but it seems to be "accelerating" - that is, it appears to be rising at a faster and faster rate. To illustrate this more pictorially, observe that in each of the curves depicted below we have that the curve is rising. However, in the first case the curve seems to be rising faster and faster, while in the second case it seems to rise more and more slowly. We shall discuss this in more detail later, but for now let us merely observe that the first case corresponds to acceleration while the second case corresponds to deceleration.



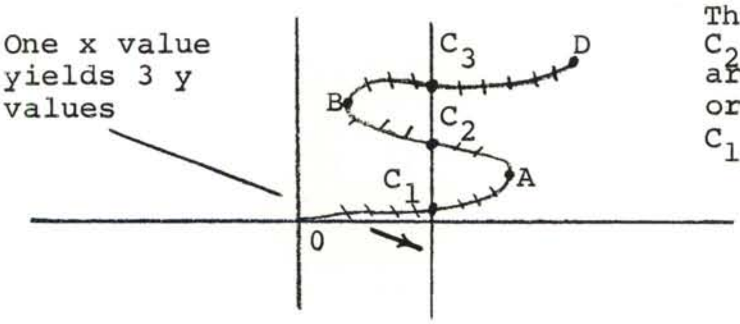
(1)



(2)

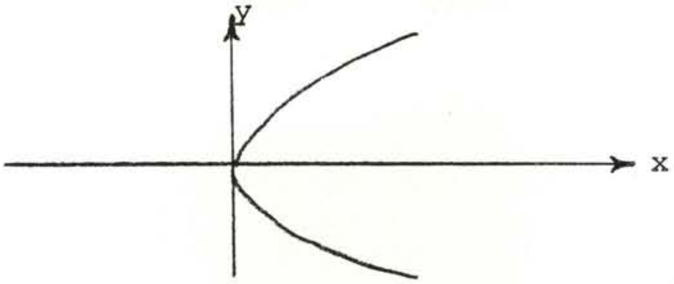
Returning to $f(x) = x^2$ observe that we do have a genuine acceleration. For example when x changes from 1 to 2 f(x) changes from 1 to 4; thus, in this case an increase in x by 1 unit produces an increase in y to 3 units. Yet when x changes from 2 to 3 (which is still only a 1 unit change in x), f(x) now changes from 4 to 9, or a change in 5 units.

Graphs also supply us with a nice answer to a deep question; namely, as to why we may always deal with single-valued functions without loss of generality. The point is that the graph of multi-valued function has the property of "doubling back". In other words, this is what is implied when we say that a line parallel to the y-axis intersects the graph in more than one point. At any rate, for the purpose of an illustration let us suppose that our graph is a "smooth" curve. Intuitively, it should be clear that the places at which the graph doubles back are those at which the curve possesses a vertical tangent line. In terms of a picture:

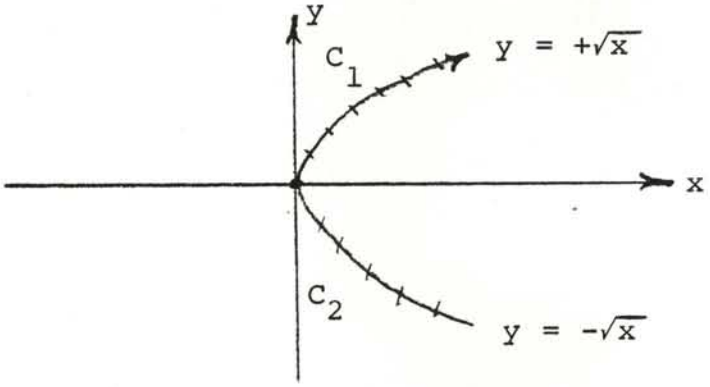


The curves C_1 (from 0 to A), C_2 (A to B) and C_3 (B to D) are single valued and our original curve is $C_1 \cup C_2 \cup C_3$

Notice that these points of tangency partition the curve into a UNION of single-valued curves. As a specific example, we can now discuss the convention that says $\sqrt{x}$ means $+\sqrt{x}$ unless otherwise specified. Let us start with the equation $y^2 = x$. The graph of this equation is:



For positive values of x, we see that the curve is double valued; and the curve possesses a vertical tangent at the point (0,0). We can thus break the curve into two mutually exclusive subsets. These two pieces are called BRANCHES. One branch lies above the x-axis and the other below. We shall denote the upper branch by $y = +\sqrt{x}$ and the lower one by $y = -\sqrt{x}$. This is in accord with the usual idea that if $y^2 = x$ then $y = \pm\sqrt{x}$. At any rate once we know one branch in detail, the other is just the mirror image with respect to the x-axis. Thus:



$y = x^2$ is then given by $C_1 \cup C_2$ where C_1 and C_2 are both single-valued.

To summarize what we have said about one-to-oneness and single-valuedness:

(1) If the graph "doubles back" the function is not single valued. In this case we can partition our function into a union of single-valued functions by noting the points at which the graph has vertical tangents.

(2) If the curve passes from rising to falling (and for a smooth curve this is characterized by those points at which we have a horizontal tangent) then the function is not 1-1.

(3) The upshot of these last two points is that if the graph is either always rising or else always falling, the function is both 1-1 and onto.

There is much more that we would like to say about functions. But for our immediate purposes we have completed our list of preliminary basic ideas, and we are now ready to reinforce and expand these ideas in terms of the more concrete treatment given in the text.

E. Inverse Functions

The concept of inverse functions is usually associated with functions that are both one-to-one and onto, and very shortly we shall discuss this idea. First, however, we want to emphasize that from an intuitive point of view the concept of inverse functions has been with us for quite some time, in various disguises, in the traditional curriculum. Roughly speaking, we may associate inverse functions with a "switch in emphasis". For example, when one refers to subtraction as being the inverse of addition, one is, in effect, saying that, for example, 3 + 2 = 5 and 5 - 3 = 2 say the same thing but from a different point of view. In this context, every subtraction problem can be paraphrased as an equivalent addition problem. That is, we may think of 5 - 3 as naming that number which must be added to 3 to give 5. More generally, we may define a - b by

$$b + (a - b) = a$$

In a similar way, to form the inverse of, say

$$c = \log_b a \qquad (1)$$

we would write

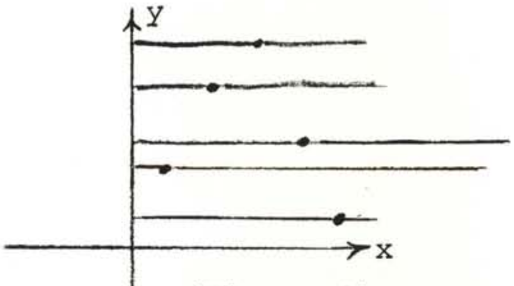
$$a = b^c \qquad (2)$$

The point is that (1) and (2) are different ways for saying the same thing; all that differs is that in (1) c seems to be emphasized while in (2) a seems to be emphasized. (In the more familiar functional notation $y = \log_b x$ and $x = b^y$ have the same meaning but the roles of the dependent and independent variables are reversed.)

The emphasis on one-to-oneness lies in the fact that without it we can get into a bit of trouble when we try to "invert." For example, when we try to switch the emphasis in $s = 16t^2$ to form $t = \pm\sqrt{\frac{s}{16}}$ we wind up with a multi-valued function, even though the original function was single-valued.

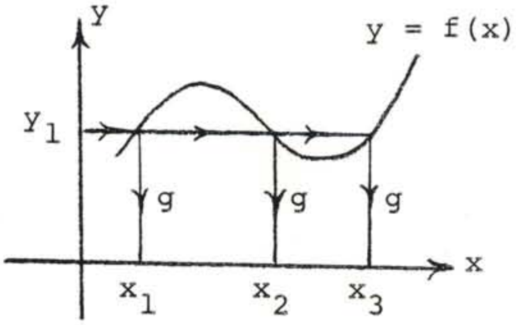
In terms of a graph, as we have described in the previous section, if the curve which represents the equation $y = f(x)$ is "unbroken" then f is one-to-one if and only if the curve never "doubles back"*. That is,

*If the graph consists of data which are made up of separate points (more formally, this is called discrete data), the function may be one-to-one even though the graph "doubles back". For example:



(Figure 2)

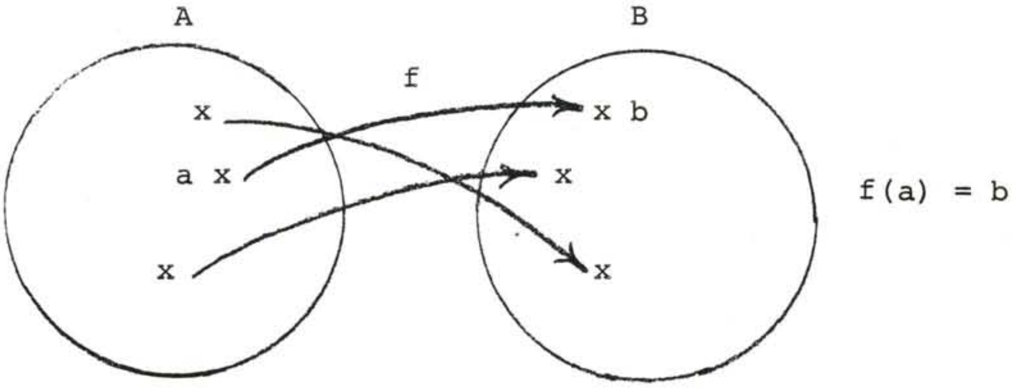
No two points lie on the same horizontal line (though in other cases they could have).



(Figure 1)

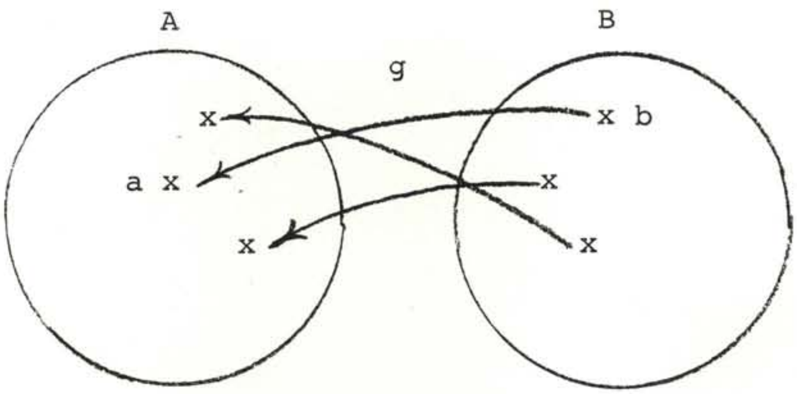
This curve is single-valued but not 1-1. If we try to "invert" the roles of x and y and write the same curve in the form x = g(y) we see that g is not single-valued since one value of y determines several values of x.

In terms of circle diagrams it is rather easy to describe inverse functions. Suppose f:A → B is both one-to-one and onto. Then,



(Figure 3)

We can induce a function g:B → A merely by reversing the sense of each of the arrows in Figure 3 (that is, by interchanging the head and the tail of each arrow). Thus,

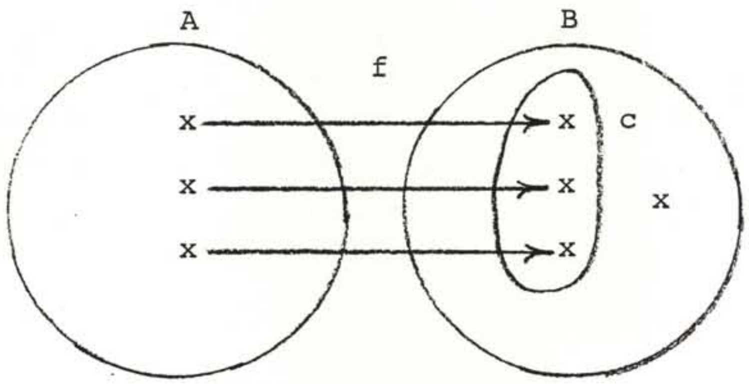


$g: B \to A$

$g(b) = a$

(Figure 4)

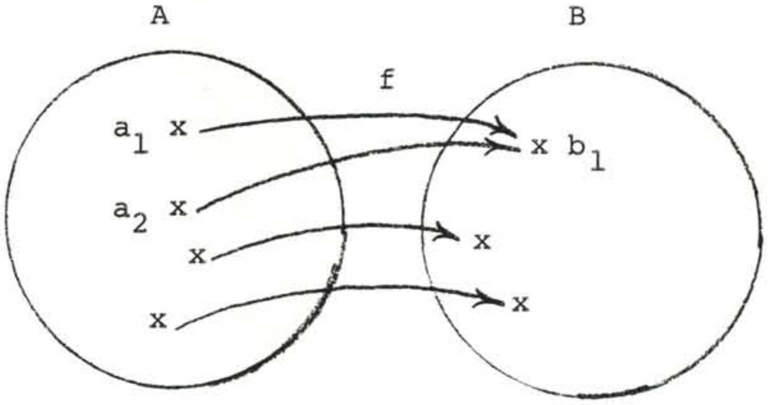
Before proceeding further, let us observe that the existence of g depends on f being both one-to-one and onto. For example, if f is not onto, the induced function g does not have B as its domain. Pictorially:



If we form g by reversing the sense of our arrows, $B \neq \text{dom } f$. Rather $c = \text{dom } f$. This prevents g from being a function from B to A.

(Figure 5)

On the other hand, if $f: A \to B$ is not one-to-one, even if it is onto, the induced function g does not exist since at least one element in B has more than one image in A. Again, pictorially,



$f(a_1) = b_1$

$f(a_2) = b_1$

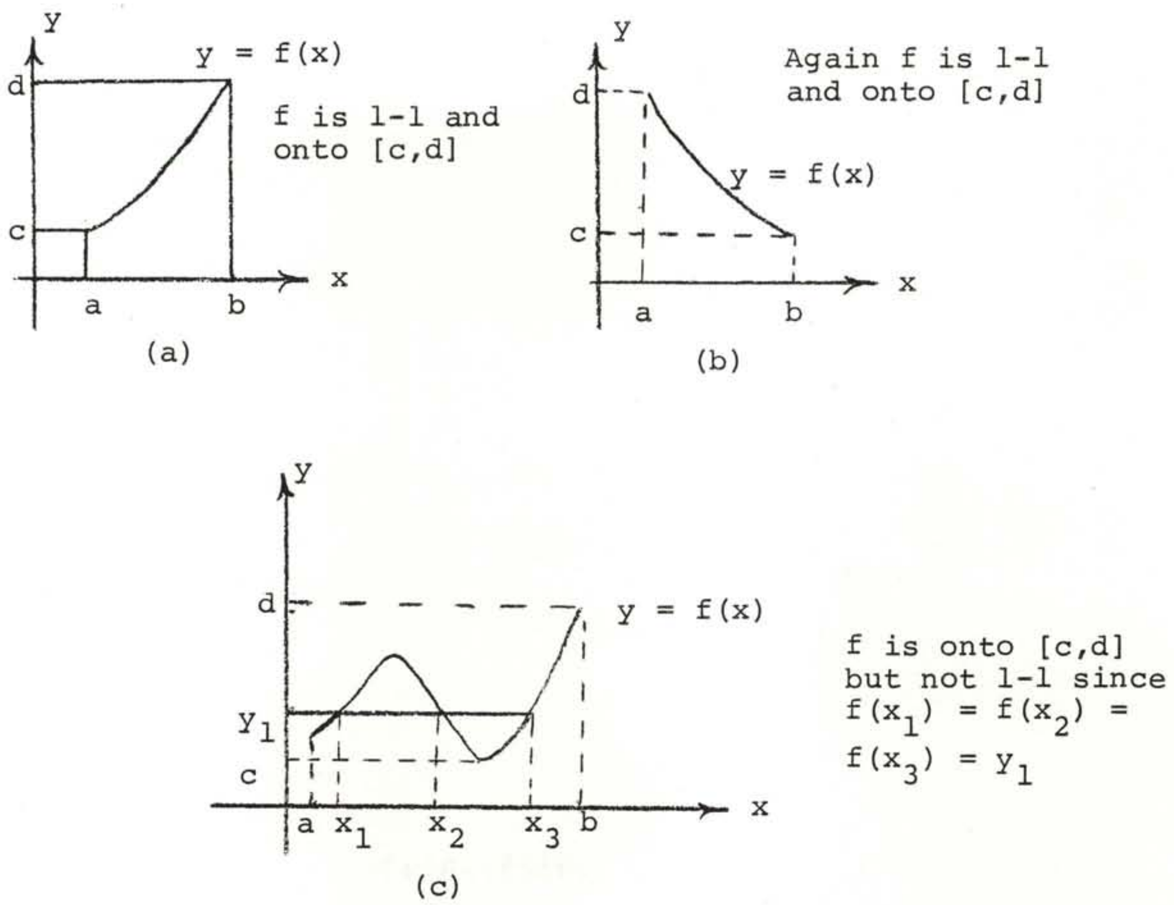
Hence

$g(b_1) = a_1$ *and* a_2

(Figure 6)

In any event, however, if f:A → B is both one-to-one and onto, then we can form the inverse function g. Notice that there is, of course, a rather strong relationship between f and g and for this reason g is usually denoted by a very special symbol, namely, f^{-1} (read as "f inverse"). The reason for this notation is to emphasize the role of "inverting". As seen from Figures 3 and 4, if we let $g = f^{-1}$ then $f(a) = b$ and $a = f^{-1}(b)$ are merely two different ways of saying the same thing, but with a switch in emphasis (see note at the end of this section for further clarification of the notation f^{-1}).

Lest the preceeding discussion seem a bit too abstract, let us once again return to the special case of functions of real variables. Let us assume that the domain of f is the closed interval [a,b] and that its range is the closed interval [c,d]. We have already seen that if f is to be one-to-one the graph can never double back. In terms of graphs,



f is 1-1 and onto [c,d]

(a)

Again f is 1-1 and onto [c,d]

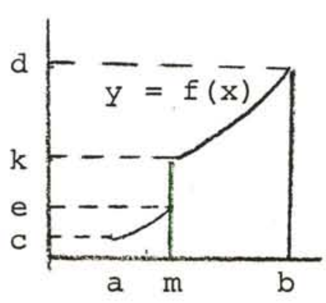
(b)

f is onto [c,d] but not 1-1 since $f(x_1) = f(x_2) = f(x_3) = y_1$

(c)

(Figure 7)

To see what it means for f not to be onto [c,d], we need only observe that this can happen whenever our curve has a "break" or a "jump" in it. That is,



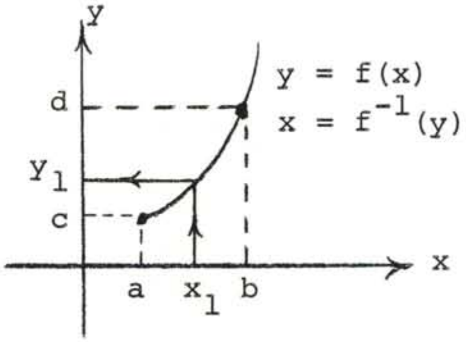
For any $y_1 \varepsilon (e,k)$ there exists <u>no</u> $x_1 \varepsilon [a,b]$ such that $f(x_1)=y_1$
Hence $y_1 \notin \mathrm{dom}\, f^{-1}$

The break in the curve at x=m results in f not being onto [c,d]

(Figure 8)

Thus, as we have mentioned before, if we wish to think of an invertible function, its graph must be similar to either Figure 7(a) or 7(b).

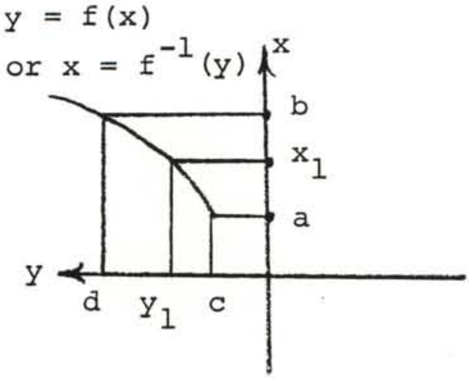
Let us now consider the case of a real function f for which f inverse exists. If we denote its graph by $y = f(x)$, let us observe that the graph is also denoted by $x = f^{-1}(y)$. That is,



We may view f as the function which maps x_1 into y_1 or we may view f^{-1} as the function which maps y_1 into x_1.

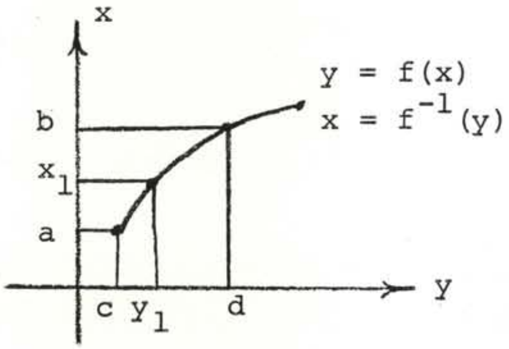
(Figure 9)

The main obstacle in Figure 9 is that in trying to "read" f^{-1} we are not used to having the y-axis (vertical axis) denote the independent variable [which it does when we write $x = f^{-1}(y)$]. We could rotate the system in Figure 9 through a positive 90° and this would give us:



(Figure 10)

Again, the problem with Figure 10 is that the sense of the horizontal axis is the opposite of what we are used to. This is easily rectified by reflecting the graph in Figure 10 about the vertical (in this case, x) axis. Thus:

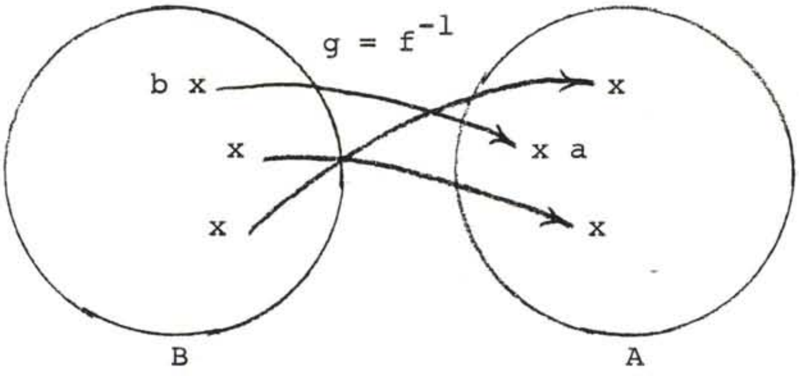


(Figure 11)

It is important to observe and to understand that in all three diagrams (Figures 9, 10, and 11) the functional relation denoted by $y = f(x)$ or $x = f^{-1}(y)$ are the same. The only reason that the curves look different is that our axes are labelled differently. In other words, we hope that it is clear that while the relation $y = f(x)$ is independent of any coordinate system, how the graph looks will certainly depend on how the axes are labelled. In terms of our earlier remarks, Figure 10 is only an intermediary step between Figure 9 and 11, and Figures 9 and 11 are two different ways of saying the same thing with a different emphasis. That is, if we accept the convention that the independent variable must always be plotted along the horizontal axis and in the left-to-right sense, then Figure 9 is the appropriate graph for $y = f(x)$, while Figure 11 is the appropriate graph for $x = f^{-1}(y)$. Notice that nothing forces us to accept this convention, and had we wished, we could have deduced any properties of $x = f^{-1}(y)$ just by looking at Figure 9.

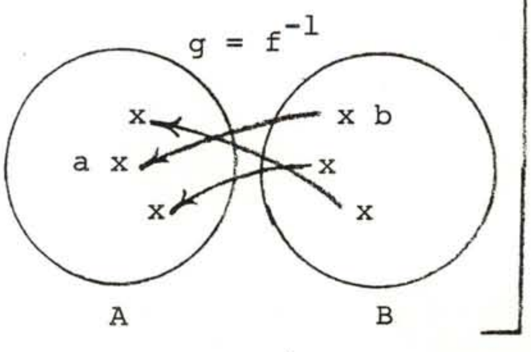
The analog of these last few remarks in terms of circle diagrams may be explained in terms of Figures 3 and 4. Once we knew f we automatically could find g (as we did in Figure 4) just by reversing the arrowheads. On the other hand, we might feel strange using a diagram that seemed to suggest the notation $A \leftarrow B:g$, suggested in Figure 4 since the domain is listed to the right of the range.

This causes no trouble in the sense that if we insist that the domain of the function appears to the left of the range, we could recopy Figure 4 interchanging the positions of A and B and obtain:



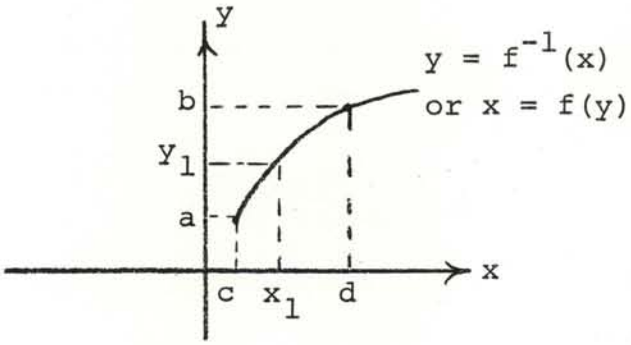
(Figure 12)

For your convenience, Figure 4 was:



All we are saying is that the function is not changed by changing from Figure 4 to Figure 12; what is changed is that if we disregard how the domain and range are labelled in Figures 4 and 12, the pictures look quite different, and this difference is due to the fact that we have changed the position of the domain and the range.

In any event, returning to our mainstream of discussion, there may still seem to be one psychological drawback to the diagram in Figure 11 and that is that we are not used to calling the horizontal axis the y-axis. If this is the case, and we are bothered by this, then we can overcome this obstacle by merely interchanging the roles of the x- and y-axes. In other words, we may "relabel" Figure 11 to form:



(Figure 13)

That is, the names we give to the x and y axes are really immaterial, except that we must remain consistent in a given application. To help clarify this idea, let us consider a notation such as $f(x) = x^2$. f is defined as the rule which assigns to any number (x), its square (x^2). Clearly, this could just as well have been written as $f(y) = y^2$. More generally, we could have written $f([\ ]) = [\ ]^2$, where [ ] denotes the "input" to the "f-machine".

In a similar way, we may view an expression like $y = f(x)$ to mean that x is the generic name of the input to the f-machine, while y is the name for the resulting output. We could have denoted this by writing:

$$\{\ \} = f([\ ])$$

Then, if we replace { } by y and [ ] by x we obtain

$$y = f(x)$$

while if we replace { } by x and [ ] by y we obtain

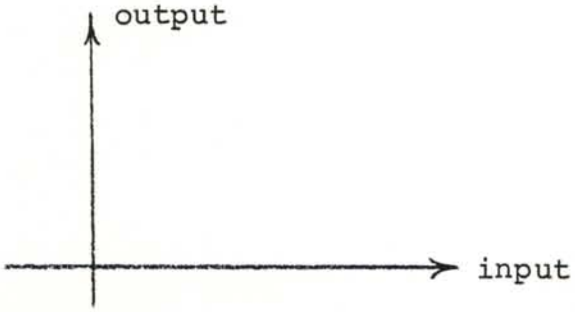
$$x = f(y)$$

As far as defining f is concerned, it makes no difference whether we write $y = f(x)$ or $x = f(y)$ since each says

$$\text{output} = f(\text{input})$$

What does matter, of course, is that the graph of f will depend on whether the independent variable (input) is represented by the horizontal axis or the vertical axis.

This is precisely what happened in our transition from Figure 9 to Figure 13 where we "transformed" the graph of $y = f(x)$ into the graph of $y = f^{-1}(x)$ [or more suggestively, perhaps, $x = f(y)$]. The point is that had we labelled our coordinate system by:



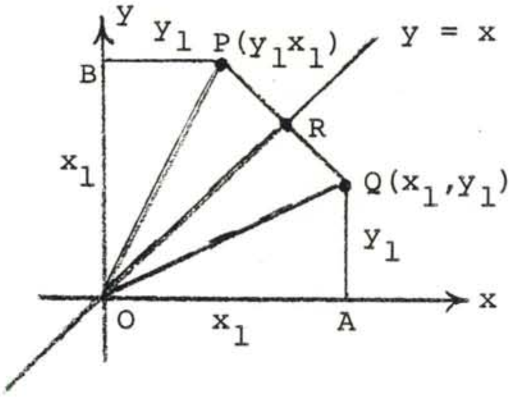
then Figure 9 could have been used to denote either the graph of $y = f(x)$ or the graph of $x = f(y)$, since in the first case x is the input, while in the second case y is the input. The problem occurs when we agree that the axes must be "frozen" as they appear in Figure 9 but that we still want the graph of the relation $x = f(y)$.

With the convention that the horizontal axis is the x-axis and the vertical axis the y-axis, there is an interesting relation between the graphs of $y = f(x)$ and $y = f^{-1}(x)$.

To see this, we may proceed as follows. Suppose (x_1,y_1) satisfies $y = f(x)$, that is (x_1,y_1) is a point on the curve $y = f(x)$. Since $y = f(x)$ is another name for $x = f^{-1}(y)$, it follows that (x_1,y_1) also satisfies $x = f^{-1}(y)$. If we now agree to interchange the names of our variables so that we keep the desired orientation and names of the axes, then we observe that (x_1,y_1) satisfies $x = f^{-1}(y)$ says the same thing as (y_1,x_1) satisfies $y = f^{-1}(x)$.* In this way, we see that (x_1,y_1) belongs to the curve $y = f(x)$ if and only if with the same orientation of axes (y_1,x_1) belongs to the curve $y = f^{-1}(x)$. [By the way, this type of observation makes sense even if f^{-1} doesn't exist; that is, as we have tried to emphasize before, $y = f(x)$ and $x = f(y)$ are obtained from one another merely by reversing the roles of input and output. The point is that as long as f is single-valued but not necessarily one-to-one, we may meaningfully write $x = f(y)$ even though we cannot rigorously write $y = f^{-1}(x)$].

*It may not seem clear that y_1 is the input of the f^{-1} machine. That is, it may seem unnatural to say things like let $x = y_1$. If this is the case, think, instead, of the pair of points, say, (a,b) and (b,a). All we are saying is that the input of the f-machine (a) is the output of the f^{-1}-machine while the output of the f-machine (b) is the input of the f^{-1}-machine.

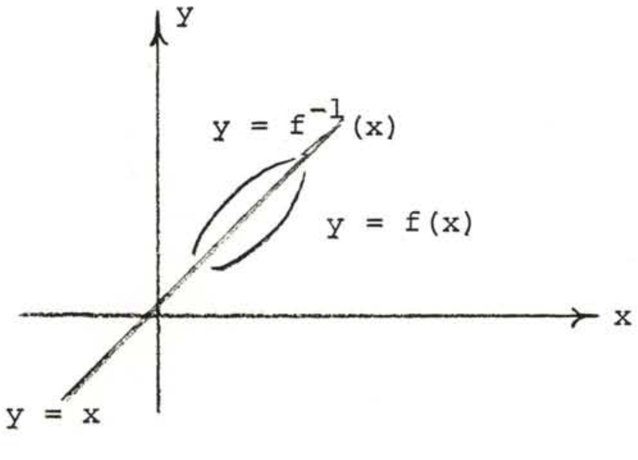
We then observe that in a given Cartesian coordinate plane the points (x_1, y_1) and (y_1, x_1) are symmetrically located with respect to the line $y = x$ (see Figure 14) and hence that the curves $y = f(x)$ and $x = f(y)$ (or equivalently $y = f^{-1}(x)$ if f^{-1} exists) are mirror images of one another with respect to the line $y = x$.



$\Delta OBP = \Delta OAQ \therefore \overline{OP} = \overline{OQ}$
$y = x$ bisects $\sphericalangle POQ$ and ΔPOQ
is isosceles $\therefore \overline{PR} = \overline{RQ}$
and $PR \perp OR$, $\overline{RQ} \perp \overline{OR}$

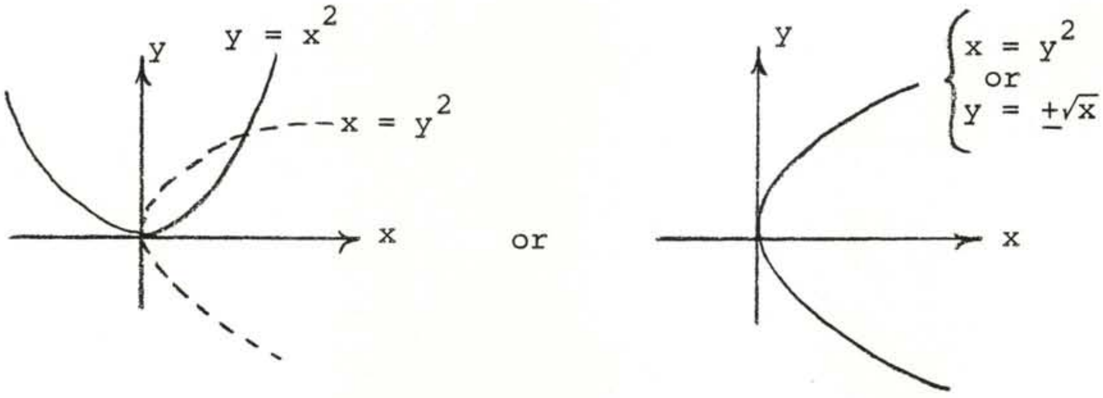
(Figure 14)

These remarks can be concretely illustrated by superimposing Figures 9 and 13, whereupon it is easily seen that



(Figure 15)

To see what happens if f is not one-to-one, we may consider the explicit case where $f(x) = x^2$. If we plot the graph and then reflect it with respect to the line $y = x$, we obtain

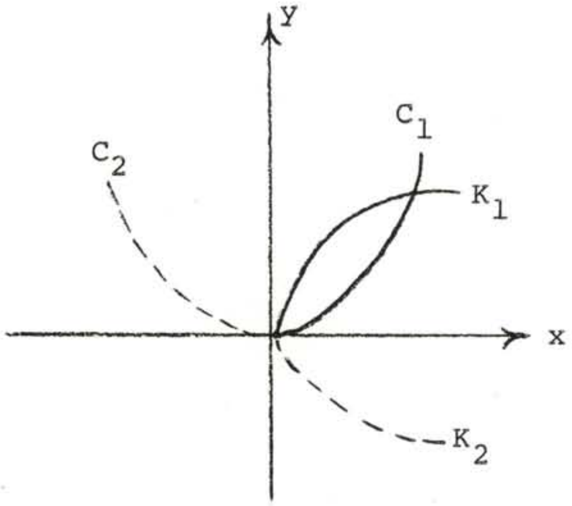


(Figure 16)

This is the curve $x = y^2$ or, more familiarly, the double-valued function $y = \pm\sqrt{x}$.

Now, in the same way that we can decompose a multi-valued function into a union of single valued functions, we can decompose a non one-to-one function into the union of one-to-one functions. (Notice that in terms of reversing the roles of the variables, we can now relate single-valued andone-to-one. Namely, if $y = f(x)$ is not one-to-one, then $y = f^{-1}(x)$, or $x = f(y)$ is not single-valued.)

Again, with reference to Figure 16 we can see what must be done if f is not one-to-one. Namely, we decompose the curve $y = x^2$ into pieces say C_1 and C_2, each of which is one-to-one. Then, each of the curves C_1 and C_2 is invertible, yielding K_1 and K_2 respectively. That is,



If, $y = x^2$, $x \geqslant 0$ then $x = +\sqrt{y}$ is the inverse.

If $y = x^2$, $x \leqslant 0$ then $x = -\sqrt{y}$ is the inverse.

(Notice that $x = -\sqrt{y}$ cannot be the inverse of $y = x^2$, $x \geqslant 0$; among other things, notice that C_1 and K_2 are not symmetric with respect to $y = x$.)

(Figure 17)

In many physical situations we can distinguish between C_1 and C_2 in the sense that we might be given the information that x must be positive, but if we are not given such information, we have no way of knowing, without the risk of ambiguity, whether it was 2 or -2 that yielded 4 as the output.

We shall say more about inverse functions later in the course in a more direct application of calculus. For now, we only want you to feel at home with: (1) what is meant by an inverse function, (2) what this means pictorially, and (3) what goes wrong if our function is not both one-to-one and onto.

A NOTE ABOUT THE f^{-1} NOTATION

Many of us are used to the notation that $a^{-1} = \frac{1}{a}$. In this sense, there might be a tendency to confuse $y = f^{-1}(x)$ with $y = \frac{1}{f(x)}$. To be sure, there are times when one might write f^{-1} to mean just this, but all other considerations aside, it should be clear from context when such a thing happens.

It may be interesting to note that f^{-1} is used to denote the inverse of f for the same reason that a^{-1} is used to denote the inverse (reciprocal) of the number, a. Observe that $a(a^{-1}) = 1$, and 1 may be viewed as the identity element for multiplication, meaning that with respect to multiplication we do not change the "identity" of a number when we multiply it by 1. In this sense, let us define the identity function I to be the function which does not change the identity of any input. That is, define I by $I(x) = x$ for all x. If we then agree to combine functions by the usual rule of composition; that is, $f \circ g$ means $f(g(x))$, we see that $f \circ f^{-1} = I$.

Let us illustrate our remarks with a specific example. Suppose $f(x) = 2x - 7$. In the language of graphs, we have

$$y = 2x - 7 \qquad (1)$$

If we invert (1) (and in the precalculus curriculum, this was known as solving for x in terms of y) we obtain

$$x = \frac{y + 7}{2} \qquad (2)$$

If we then employ our remarks about wanting x always to denote the independent variable, we may rewrite (2) as:

$$y = \frac{x + 7}{2} \qquad (3)$$

If we let $g(x) = \frac{x + 7}{2}$, then it follows that $g = f^{-1}$. As a check, we then note:

$$f\left(f^{-1}(\underline{x})\right) = f\left(\frac{x + 7}{2}\right)$$

$$= 2\left(\frac{x + 7}{2}\right) - 7$$

$$= \underline{x}$$

$$\therefore f \circ f^{-1} = I$$

$$f^{-1}\left(f(\underline{x})\right) = f^{-1}(2x - 7)$$

$$= \frac{(2x - 7) + 7}{2}$$

$$= \underline{x}$$

$$\therefore f^{-1} \circ f = I$$

In summary then, if $f(x) = 2x - 7$, then $f^{-1}(x) = \frac{x + 7}{2}$ while $\frac{1}{f(x)} = \frac{1}{2x - 7}$.

The relationships are:

$$[f(x)] \left[\frac{1}{f(x)}\right] = 1 \qquad (f(x) \neq 0)$$

$$f^{-1} \circ f \; (= f \circ f^{-1}) = I$$

Chapter IV
LIMITS

A. Introduction

Let us begin our discussion with the following contrived experiment. We have three beakers filled with water. One is filled with ordinary tap water, a second is filled with hot water, and the third is filled with very cold water. We place one hand in the hot water and the other in the cold water, and after a while, we plunge both hands into the tap water. It is clear that the hands transmit different sensations in that the hand that was in the hot water will find the tap water quite cold, while the hand that was in the cold water will find the tap water quite warm. From one point of view both hands depict an inaccurate picture, yet each hand "is telling the truth" based on its experience.

Now, in contrast to this approach of putting the different hands into the tap water, suppose, instead, that we place a thermometer in the tap water and the thermometer yields a reading of 68°F. To be sure, 68°F may seem quite cold to one observer and quite warm to another, but the fact is that 68°F is an objective measurement that transmits the same knowledge to either observer.

With this example in mind, perhaps it will be easy for us to understand what is meant when we say that a major problem in scientific investigations is to translate a well-known qualitative concept into a more objective, well-defined quantitative form. In any event this problem is at the very foundations of calculus.

For example, differential calculus is primarily concerned with the concept of instantaneous rate of change. That is, it concerns itself, for example, with how fast a particle is moving at a given instant (as opposed to pre-calculus mathematics wherein we talk about average rate of change during a certain time interval). Certainly, we all have some notion as what an instant is. While our manner of wording our idea might differ from person to person, by and large we would agree that an instant was a time interval of "extremely short" duration - but "extremely short" to whom? Clearly what might be "extremely short" to one observer might seem only "moderately short" to another. Thus, just as in our previous discussion when we saw the need of a thermometer, we are now in need of some way of describing an "instant" so that it will have the same quantitative meaning to all observers - and at the same time, we must make sure that in our quest for objectivity we have not destroyed our intuitive feeling about what we "know" an instant is supposed to be. In summary, then, the fundamental problem of differential calculus is to find an objective definition of an instant that agrees with what our intuition tells us an instant is.

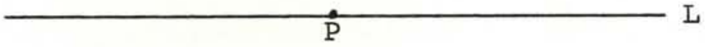
It is in this environment that we must introduce the notion of limits. Limits form the building blocks of calculus. From one point of view calculus is nothing more than pre-calculus mathematics - AUGMENTED BY THE CONCEPT OF A LIMIT.

In this course our approach toward introducing new concepts shall be the following. Since most people can think better in terms of concrete situations rather than in abstractions, we shall always try to introduce a new concept "intuitively". That is, we shall try to show in terms of our experience why it was crucial for the concept to have been invented. Such an attempt relies heavily on subjective interpretations. The problem is that it is difficult to learn or to teach subjective ideas. So, in order that we can better handle these new concepts we will try to "pin them down" objectively once they have been properly introduced. In other words, the second phase will be to treat the new concept from a more objective point of view - a point of view that is more independent of the subjective differences we encounter in going from one student to another.

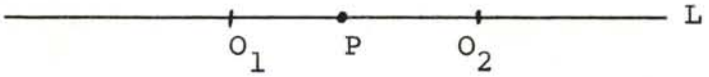
In particular, as far as the topic of limits is concerned, we shall try to discuss the idea informally first and then gradually develop a more and more refined objective treatment which will seal in the flavor of the more intuitive approach but at the same time eliminate those places which tend to be ambiguous or otherwise highly subjective.

B. The Problem of Zero Divided by Zero (0/0)

Suppose a particle is moving along the line L and we wish to compute its speed at the instant it is at the point P on L. Thus:



Now in logical investigations our first approximation is usually to try to reduce the new situation to a collection of more familiar situations. In this case, recall that it is assumed that we already know how to handle the problem of average speed. For example, suppose instead of the given problem we were told that there were two observers, O_1 and O_2, on either side of P and we wished to find the average speed of the particle as it moved from O_1 to O_2. Experimentally, we need only measure the distance between O_1 and O_2 and then measure the time that it took for the particle to travel from O_1 to O_2. By definition, then, the average speed of the particle in going from O_1 to O_2 is just the distance it travels divided by the time it took to travel this distance.



The main problem is that we have found the RIGHT ANSWER to the WRONG PROBLEM. We were not asked to find an average speed, but rather an instantaneous speed. We could then argue that as the distance between O_1 and O_2 diminished the average speed of the particle would seem to be a better estimate of the instantaneous speed - at least in so far as our intuitive notion of instantaneous speed is concerned. In fact, we might even begin to believe, since the average speed becomes a better and better approximation to the instantaneous speed as the distance between O_1 and O_2 is made less and less, that the instantaneous speed would be determined exactly when there was no distance between the two observers. Alas, the rub is that if there is <u>no</u> distance between the observers then the particle travels <u>zero</u> distance in going from O_1 to O_2 and this trip, clearly, takes <u>no</u> time. Thus, in this case, if we were to keep our definition of average speed, we would find that the average speed of the particle in going from O_1 to O_2 is given by 0/0.

Let us now digress and discuss the meaning of 0/0. Our claim is that, as it stands, 0/0 is <u>indeterminate</u>, and we shall try to illustrate what we mean from two entirely different points of view. From a more "applied" point of view, let us view 0/0 as the quotient of two "very small" numbers (which captures the flavor of the two observers being moved closer and closer together). Now it is not difficult to imagine that when we add "very small" numbers the result is still a "very small" number. In a similar way the difference and product of "very small numbers" are still "very small". (Notice that we are not going to get too concerned over the subjective interpretation of "very small" here. All we are saying is that if we add, subtract, or multiply tiny amounts the result is a tiny amount and leave it to the individual to interpret "tiny".) The point we do wish to make, however, is that when we divide very small amounts by one another, it is no longer clear what size the quotient will be. It is in terms of division that our intuitive idea of "small" causes trouble. By way of illustration suppose our numerator is 10^{-12} and that our denominator is also 10^{-12}. In this case it is clear that the quotient is 1. In this case the quotient of two "tiny" amounts is still small but hardly as "tiny" as either of the two numbers. On the other hand if our denominator is still 10^{-12} but our numerator is now 10^{-6} our quotient is $10^{-6}/10^{-12} = 10^{6} = 1{,}000{,}000$, and, in this case, by ordinary standards we would agree that the quotient of two "tiny" numbers was "very large". More specifically in this example, we are noticing that while 10^{-6} may be considered small, it is mammoth (one million times larger) when compared with 10^{-12}. In any event this should indicate why from an objective point of view we must steer clear of an expression such as 0/0.

To see this from a more abstract, mathematical point of view, let us recall the basic definition that a/b denotes that number which when multiplied by b yields a. That is, a/b is defined mathematically by:

$$b \times (a/b) = a$$

(To see that this definition agrees with what we already know, observe that by our definition 6/2 is defined by $2 \times (6/2) = 6$ whence it follows that $6/2 = 3$.)

Now if we apply this definition to 0/0, we have that 0/0 is defined by:

$$0 \times 0/0 = 0$$

In words, 0/0 is that number which when multiplied by 0 yields 0. The point is that <u>any</u> number has this property, and in this sense 0/0 does nothing to determine a specific number. That is why 0/0 is called indeterminate. (In still other words, if we were trying to guess a number and we were given as a clue that when multiplied by 2 the product was 6, we should have no trouble in deciding that the number was 3. If, however, the clue is that this number when multiplied by 0 yields 0, we know nothing that we didn't already know.)

In any event, it should now be clear as to why in mathematics we specifically exclude the expression 0/0. With these remarks in mind we can safely return to the main stream of our discussion.

It appears that our approximation to the instantaneous speed gets better and better as we allow the observers to get closer and closer together - but that we lose all our information if we allow the two observers to coincide. Thus, somehow or other, we are going to have to come to grips with the problem of allowing the two observers to get arbitrarily close - <u>BUT NEVER TOUCH</u>. Herein lies the whole kernel of the problem, for if the two observers can be allowed to get arbitrarily near to each other without touching there are infinitely many positions they can be in. (Here we neglect the "thickness" of the observer; geometrically, this corresponds to the difference between a point and a dot.) That is, as long as there is a space between two observers there is always room to fit in another pair of observers.

Thus the central problem of differential calculus turns out to be that of translating the notion of "getting arbitrarily close to but never equal" into a precise, objective, computational form. The resulting concept is known as the limiting process and will be discussed more fully in the next section.

What is most important to note is that we are not saying that we can't define instantaneous speed because of the 0/0 form. Rather we must come up with a different definition of "instantaneous" that agrees with what we believe intuitively to be the correct definition, but which at the same time eliminates the 0/0 form.

C. A Semi-Rigorous Mathematical Approach to Limits

Let us revisit some of the ideas of the previous section from a more quantitative point of view. Suppose we have an object falling toward the earth, and we know that the distance s (in feet) it has fallen at the end of t seconds is given by:

$$s = 16t^2 \qquad (1)$$

Suppose we would like to find the speed of the object at the instant that $t = 1$. (Notice that we still make no attempt to define an instant rigorously, but rather rely on our intuition that we know what is meant by an instant.) From a pre-calculus point of view we could have solved a problem such as that of finding the average speed of the object between the time $t = 1$ and $t = 2$; or for that matter, more generally we could have tried to find the average speed of the particle between time $t = 1$ and $t = 1 + h$,* where h is any real number OTHER THAN 0 (since as we mentioned in the previous section, we are in trouble if no time transpires).

Clearly

$$s(1 + h) = 16(1 + h)^2 = 16 + 32h + 16h^2 \qquad (2)$$

and

$$s(1) = 16(1)^2 = 16 \qquad (3)$$

Subtracting (3) from (2) shows us that $\Delta s = 32h + 16h^2$. At the same time $\Delta t = (1 + h) - 1 = h$, and, accordingly, the average speed of the object between $t = 1$ and $t = 1 + h$ is given by

$$\Delta s/\Delta t, \text{ or } (32h + 16h^2)/h \qquad (4)$$

Since $h \neq 0$ we can divide through by it in (4) and obtain the result that:

$$\frac{\Delta s}{\Delta t} = 32 + 16h \qquad (5)$$

From (5) we can compute the average speed of the object on any time interval between $t = 1$ and $t = 1 + h$. For example, when $h = 0.5$ (which means that our time interval is from $t = 1$ to $t = 1.5$) we see that the average speed of the object during this time interval is $32 + (16)(0.5) = 40$ ft./sec.

*Notice that nowhere do we assume that h must be positive except in how we worded the problem. All that would happen if h were negative is that we would talk about the time interval from $t = 1 + h$ to $t = 1$. The important thing to note is that none of the analytic statements that we make depend on whether h is positive. What is important is that $h \neq 0$.

We next notice that the qualitative statement that we want the distance between the observers to become smaller and smaller but never equal to 0 translates, in this problem, into the fact that in (5) we let the value of h get as close to 0 as we wish but that $h \neq 0$. Making a chart of sorts, we obtain from (5):

Time Interval:	Average Speed
$t = 1$ to $t = 2$	48 feet per second
$t = 1$ to $t = 1.5$	40 feet per second
$t = 1$ to $t = 1.1$	33.6 feet per second
$t = 1$ to $t = 1.01$	32.16 feet per second

At the same time to emphasize that h need not be positive, notice that the choices $h = -1, -0.5, -0.1$ and -0.01 make (5) become:

Average Speed from $t = 0$	to $t = 1$	is 16	ft. per second
$t = 0.5$	to $t = 1$	is 24	ft. per second
$t = 0.9$	to $t = 1$	is 30.4	ft. per second
$t = 0.99$	to $t = 1$	is 31.84	ft. per second

Thus we see that the average speed seems to "tend" to 32 feet per second as h "tends" to 0. Indeed looking at (5) directly and letting h approach the value of 0 we sense that 16h also tends to 0 and hence that 32 + 16h tends to 32. In such a case, the "official" jargon is to say that "32 + 16h approaches 32 as h approaches 0". We also say that "the limit of 32 + 16h <u>is</u> 32 as h approaches 0". We write this as:

$$\lim_{h\to 0} (32 + 16h) = 32 \qquad (6)$$

(Notice especially the use of the equal sign. As h gets closer and closer to 0, 32 + 16h gets closer and closer to 32, but the limiting value is EXACTLY 32.)

Finally, since we sense that our average speed serves as a better and better approximation for the instantaneous speed as Δt approaches 0, *we define the instantaneous speed to be this LIMIT of an average rate of speed as the size of the time interval gets arbitrarily nearly equal to 0 but is never allowed to equal 0*. In terms of our illustration, the instantaneous speed of the object when t = 1 is exactly (not approximately) 32 ft./sec. Notice that had we wished to steer clear of calculus the fact that we don't anticipate anything drastic happening between t = 1 and t = 1.01 would make it rather easy for us to accept the fact that at the instant t = 1, the speed of the particle is very close to 32.16 (the average speed in this time interval) - even though from this information alone we would not be able to claim that the instantaneous speed is exactly 32 feet per second.

If we now return to (6), our intuition would tend to tell us that all we did to form the limit was to replace h by 0. By way of further examples, if we follow the type of reasoning that led to (6), we would be tempted to conclude:

$$\left.\begin{aligned} &\lim_{x\to 3} x^2 = 9 \\[2ex] &\lim_{x\to 3} (x^2 + 2x + 1) = 16 \end{aligned}\right\}$$

Notice here that 0 is not crucial. The expression $\lim_{x\to a}$ has meaning for any real number, *a*.

As far as we've gone, we would get what seems to be the right answer merely by replacing x by 3. To generalize still further, we might be tempted to define $\lim_{x\to a} f(x)$ to be simply $f(a)$.

The really serious problem exists when our "friend" 0/0 enters the picture. For example, let us define f by:

$$f(x) = \frac{x^2 - 9}{x - 3}$$

and let us try to investigate f(3). Clearly we arrive at 0/0. Should we, therefore say that $\lim\limits_{x\to 3} f(x) = 0/0$? The answer is no. In particular, let us recall that in our feeling that the time interval could not be zero, the notion of $h \to 0$ meant that h got as close to 0 as we wished but that h could NEVER equal 0. In this spirit then our definition of $x \to a$ should mean that x gets as close to a as we wish but it can never equal a.

In other words, it is true that if we replace x by 3 then $\frac{x^2 - 9}{x - 3}$ becomes 0/0. On the other hand, $\lim\limits_{x\to 3} \frac{x^2 - 9}{x - 3}$ does not mean that we replace x by 3. Rather we must see what happens as x is allowed to get "as close to" 3 as we wish without equaling 3.

From another point of view, then we may think of $\frac{x^2 - 9}{x - 3}$ as being $\frac{(x+3\ (x-3)}{(x-3)}$.

The key point now is that contrary to what you may have believed:

$$\frac{(x+3)\ (x-3)}{(x-3)} \text{ is NOT the same as } x + 3,$$

for as we have discussed earlier we are not allowed to divide by zero. That is, we may cancel (x-3) from both numerator and denominator if and only if x - 3 is NOT EQUAL TO ZERO. The point now is that the only time x - 3 is equal to zero is when x = 3 AND THIS IS PRECISELY THE VALUE THAT x IS NOT ALLOWED TO EQUAL WHEN WE WRITE $x \to 3$. In other words, what is true is that for $x \neq 3$

$$\frac{(x+3)\ (x-3)}{x-3} = x+3$$

Therefore since $\lim\limits_{x\to 3}$ implies that $x \neq 3$, we can say that

$$\lim_{x\to 3} \frac{x^2 - 9}{x - 3} = \lim_{x\to 3} (x+3)$$

Thus, $\lim\limits_{x\to 3} \frac{x^2 - 9}{x - 3} = 6.$

More formally, if we define f and g by:

$$f(x) = \frac{x^2 - 9}{x - 3} \text{ and } g(x) = x + 3$$

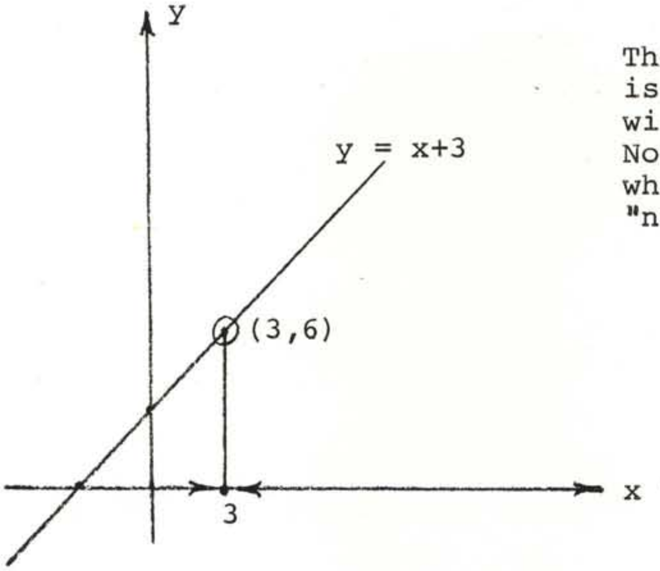
then f and g are not equal. In particular 3 belongs to the domain of g but not to the domain of f. Still another way of saying this is to observe that:

$$f(x) = \begin{cases} g(x) \text{ if } x \neq 3 \\ \text{undefined when } x = 3 \end{cases}$$

Now since $x \to 3$ excludes the possibility that $x = 3$, it follows that even though f and g are not equal we can say:

$$\lim_{x\to 3} f(x) = \lim_{x\to 3} g(x)$$

In terms of a picture:



The curve $y = f(x) = \dfrac{x^2 - 9}{x - 3}$ is the straight line $y=x+3$ with the point (3,6) deleted. Notice from the picture that when x is "near" 3, f(x) is "near" 6.

Of course, it does turn out that if we are asked to compute $\lim\limits_{x\to a} f(x)$ it will frequently happen that $\lim\limits_{x\to a} f(x) = f(a)$. In those cases, however, where we find that $f(a) = 0/0$ we are in trouble. One might feel that given f at random it is very unlikely that we will wind up with a 0/0 form. Yet the point we have tried to make clear is that we will ALWAYS wind up with 0/0 when we use limits to compute instantaneous speed, since we will always be dividing a "tiny" distance by a "tiny" time!

To write our results as generally as possible, let us get away from the specific example of $s = 16t^2$ and turn to the more general form $s = f(t)$, where all we mean here is that there is some "recipe" which allows us to express s for each value of t.

Then if we want the instantaneous speed at time $t = t_1$, we compute the average speed over the interval from $t = t_1$ to $t = t_1 + h$. This average speed is given by:

$$\frac{\Delta s}{\Delta t} = \left[\frac{f(t_1+h) - f(t_1)}{h}\right] \qquad (7)$$

Then by <u>definition</u>, the instantaneous speed at $t = t_1$ is given by the expression:

$$\lim_{h\to 0}\left[\frac{f(t_1+h) - f(t_1)}{h}\right] \tag{8}$$

and it is crucial to see that in (8) we do not replace h by 0. For if we did, the bracketed expression clearly becomes 0/0.

The "standard" technique which helps us prevent (8) from taking on the form 0/0 is to observe that as long as h is unequal to 0 we can divide through by it, say, in (7). To help make this point a bit clearer, let us return to (4) and (5) of our previous example. While we made no great issue about it, we wrote that $(32h + 16h^2)/h$ was equal to $32 + 16h$ <u>since $h \neq 0$</u>. Notice that we carefully included in our definition of $\lim\limits_{h\to 0}$ the specific restriction that $h \neq 0$.

Again returning to the more general expression (7) we <u>never</u> say that

$$\lim_{h\to 0}\left[\frac{f(t_1+h)-f(t_1)}{h}\right] = \left.\frac{f(t_1+h)-f(t_1)}{h}\right|_{h=0} = \frac{f(t_1)-f(t_1)}{0} = \frac{0}{0}$$

The great temptation to replace h by 0 has caused more than one author to define differential calculus to be the study of 0/0. More precisely when we wind up with the form 0/0 what we have done is the intuitive equivalent of having no distance between our two observers. From a more computational point of view, what we do is form the quotient

$$\frac{f(t_1+h) - f(t_1)}{h}$$

and divide through by h BEFORE we take the limit.

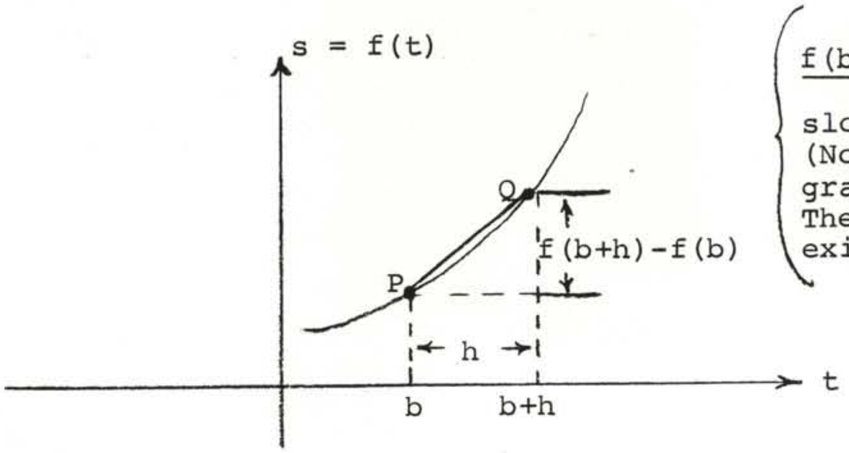
As our course progresses it will be important to note that this basic concept never changes. Essentially what happens is that we develop the computational "know-how" to handle the expression

$$\lim_{h\to 0}\left[\frac{f(t_1+h) - f(t_1)}{h}\right]$$

for a wide range of possibilities for f(t). That is, the basic concept remains the same but certain functions f(t) are more difficult to handle in the above expression than are others.

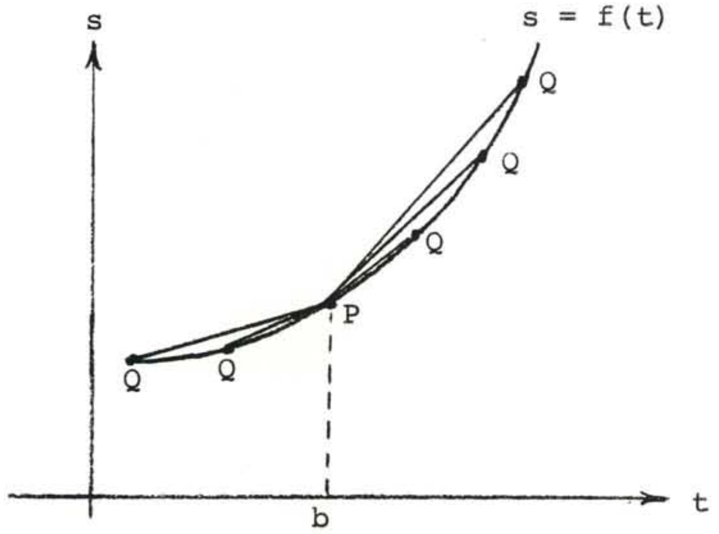
D. The Limit Concept in Terms of Graphs

The concept of instantaneous speed has a nice interpretation in terms of graphs. For example, if we take the graph of s = f(t) we have that



$\frac{f(b+h)-f(b)}{h}$ is the slope of the chord PQ (Notice that our diagram depicts h > 0. The same interpretation exists if h < 0.)

Thus, average rate of change may be identified pictorially with the slope of the chord that joins two points on the curve; while instantaneous rate of change may be identified with the slope of a tangent line to the curve. That is:



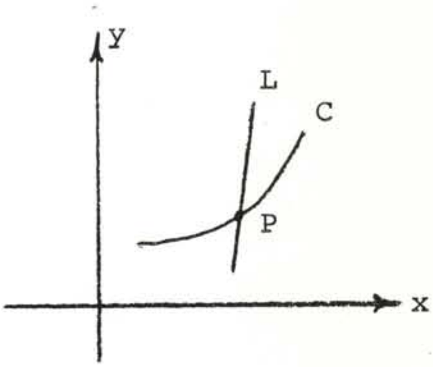
The closer Q gets to P, the more PQ looks like the tangent line to the curve at P.

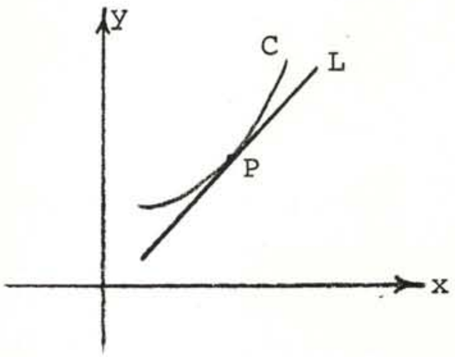
As Q gets close to P, h gets close to 0 since h is the horizontal distance between P and Q.

This shows us still another advantage of analytic geometry. Namely, if we so desire we may identify the analytic concept of instantaneous rate of change with the geometric notion of the slope of a tangent line to a curve; and, conversely, we may identify the geometric notion of slope with the analytic notion of instantaneous rate of change. The point is that there will be times when one interpretation will be preferred to the other.

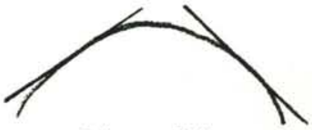
As an application of this idea let us discuss the notion of tangent lines, for it might well be that this geometric approach is easier to visualize than the more analytic notion of instantaneous rate of change.

To parallel our earlier discussion, notice that we all have an intuitive notion as to what is meant by a tangent line to a curve. Qualitatively we know that it is a line which "touches" the curve at the point of contact. Yet how can we objectively distinguish between a line which "touches" a curve and a line which "crosses" a curve. For example, in either of the two cases depicted below, the line L "meets" the curve C at one and only one point. How do we distinguish between these two cases objectively?

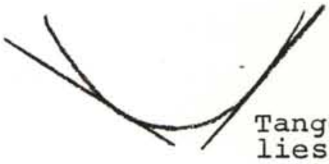




To complicate matters even more, consider the case of the line tangent to a smooth curve at a point where the curve changes its concavity. For example, the tangent line lies above the curve if the curve "spills water" at the point of contact, while it lies below the curve if the curve "holds water". Pictorially:

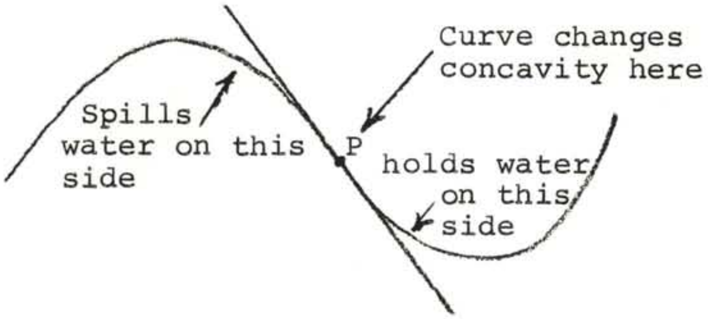


Tangent line lies above the curve



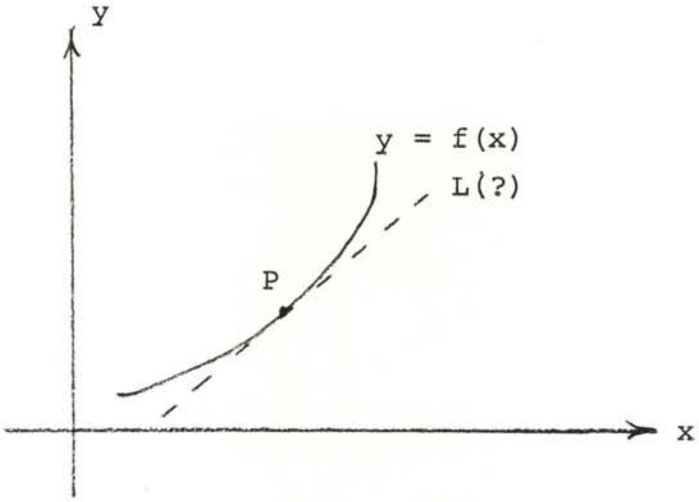
Tangent Line lies below the curve

Thus at a point at which the curve changes concavity the tangent line lies above the curve on one side and below the curve on the other side (a point at which the concavity changes is called a point of inflection). At a point of inflection the tangent line CROSSES the curve. That is:



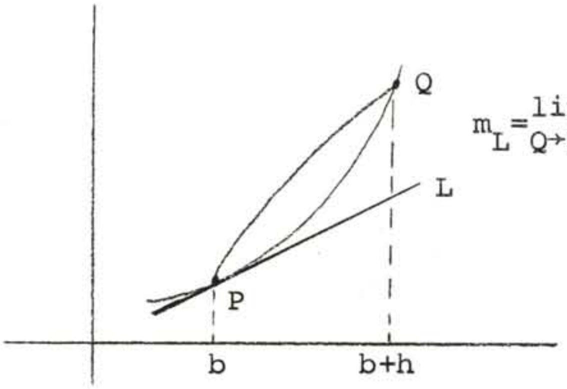
From a geometric construction point of view, notice that in the usual plane geometry course the only "curves" that are studied are circles and straight lines, and in these cases the idea of a tangent is unusually simple. Namely, a straight line is its own tangent and for a circle we can construct a tangent line at a point on the circle simply by drawing a line perpendicular to the radius at that point. How, then, would we construct a tangent line at a point of an arbitrarily given curve? Certainly, one way is the subjective technique of placing a ruler on the curve at the given point and sliding it around until the ruler seems to "touch" the curve at the given point. Obviously such an approach is subjective. It not only depends on the person who is drawing the line, but even the same person may see things from a different perspective at different times.

The point is that we can apply the idea of limits to this geometric situation in a way that is completely analogous to what we did analytically earlier. That is, suppose we wish to locate the line tangent to curve C at point P (this is tantamount to finding the slope of the line since the point P is a point on the line). That is:



We could pick a "near-by" point Q on the curve and find the slope of the line PQ. This would again give us a correct answer to a wrong problem since we are not interested in the line PQ. As we allow the point Q to be chosen closer and closer to P, however, the line PQ seems to become a better and better approximation to the tangent line (if indeed there is a tangent line). Thus we might define the slope of the tangent line to be the limit of the slope of PQ as Q is allowed to get as close to P as we wish PROVIDED THAT P AND Q ARE NOT ALLOWED TO COINCIDE. For if P and Q coincide, we have only one point; and it takes two points to determine a line. (The notion that the points P and Q were to be arbitrarily close but never equal reflects in an early [seventeenth century] definition that a tangent line to a curve is a line which "passes through 'two consecutive points' on the curve".)

Again in terms of a picture:



$$m_L = \lim_{Q\to P}(m_{PQ}) = \lim_{h\to 0}\frac{f(b+h)-f(b)}{h}$$

In summary we conclude this section with the result that $\lim_{x\to a} f(x) = L$ means that we can make the difference between f(x) and L as small as we wish (but not necessarily zero) by choosing x "sufficiently close to" but not equal to a, and that we may visualize this either analytically or geometrically.

In the next section we shall try to formulate this same result in purely analytic language. That is, we shall try to give a precise, rigorous, mathematical definition that captures the meaning of our previous discussion but which affords us an objective way of computing limits.

E. Limits - A Rigorous Approach

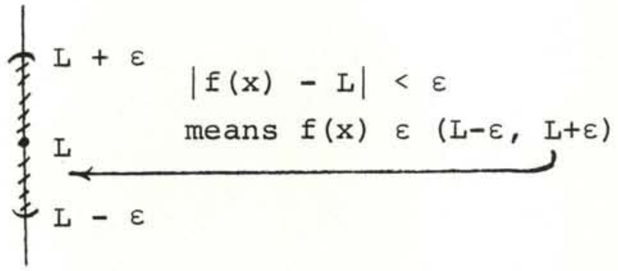
Somehow or other we would now like a way of saying $\lim_{x\to a} f(x) = L$ in a well-defined mathematical way which preserves the mood that already prevails about limits.

Thus we must find some way of translating "the difference between f(x) and L can be made as small as we wish" and "when x is sufficiently close to a" in precise mathematical language.

We shall find that the use of absolute values helps us immensely in this quest. To make f(x) as close to L as we wish means that we should be able to make the DIFFERENCE between f(x) and L smaller than any arbitrarily prescribed value. Let us introduce the symbol ε (epsilon) to generically name an arbitrary positive number. Then the mathematical statement for "f(x) is within ε of L" simple becomes

$$|f(x) - L| < \varepsilon$$

Pictorially:


L + ε
$|f(x) - L| < \varepsilon$
means $f(x)\ \varepsilon\ (L-\varepsilon, L+\varepsilon)$
L
L - ε

To indicate that x must be sufficiently close to a, we could say that we can find another number, generically named by δ (delta), such that $|x - a| < \delta$. Now to capture the flavor that $x \neq a$, we observe that $x = a$ if and only if $|x - a| = 0$. Thus we impose the condition that $|x - a| \neq 0$. Since absolute value cannot be negative this is the same as saying that $|x - a| > 0$.

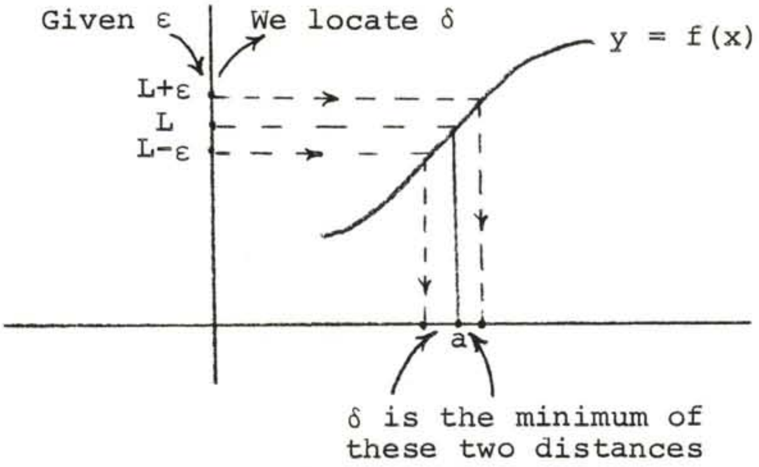
Putting this all together we now state the following formal definition:

$$\lim_{x \to a} f(x) = L$$

means that for each $\varepsilon > 0$ we can find $\delta > 0$ (where the choice of δ may depend on the choice of ε) such that $|f(x) - L| < \varepsilon$ whenever $0 < |x - a| < \delta$. (2)

Again, from an informal point of view, this simply says that for any given $\varepsilon > 0$ we can find a number δ such that whenever x is within δ of a (but not equal to a), f(x) will be within ε of L.

In terms of graphs, we are saying that:



Rather than proceed abstractly let us consider a particular example. Let f(x) = (2 + x) (3 - x), and let us compute $\lim_{x\to 0} f(x)$.

Certainly, it seems fairly obvious that as $x \to 0$, f(x) approaches 6. Thus we might conjecture that $\lim_{x\to 0} f(x) = 6$.

What we would like to do is gain some experience with the so-called "epsilon-delta" method for verifying this result.

By way of review, recall that the limit of (2 + x)(3 - x) as x approaches 0 is EXACTLY 6, not approximately 6. That is, f(x) is approximately 6 for values of x which are "near" 0, but not equal to 0; but the limiting value of f(x) as $x \rightarrow 0$ is 6.

This suggests a rather interesting difference between the "pure" and the "applied" worlds. For example, in the "pure" world we can talk about a piece of string being EXACTLY six inches long; but in the "real" or "applied" world we can never measure accurately enough to ascertain whether something is exactly six inches long. Among other things, the "lines" on our measuring device themselves have thickness and if the string terminated on the marking called 6 on our ruler, we would not be sure whether it was the beginning of the 6 mark, the end of the mark, etc. This is one reason that we talk about SIGNIFICANT FIGURES. That is, to the "pure" mathematician 6, 6.0, 6.00, 6.000, etc. are all synonyms (i.e., different numerals), all of which name the number 6; but to the engineer these numerals are different ways of saying "CLOSE ENOUGH!". In other words, when he writes that the length of the string is 6.0 he is saying that he does not know what the EXACT length is, but any error in his measurement that would indicate that the length was not exactly 6 will not occur until at least the second decimal place. Similarly, when he writes 6.000 he is asserting confidence that any error must wait until at least the fourth decimal place before it can occur.

This same idea is extended when one specifies TOLERANCE LIMITS in indicating the dimensions of certain objects. Thus when we write that the length of a certain part is to be 6 ± 0.001, we are really saying that anything that is made to this specification is "close enough" for our purposes.

Getting back to our immediate problem, in the real world we might be content if we could only be sure that the difference between f(x) and 6 was no greater than 0.0001. In this case we would in turn be interested in knowing how much x could deviate from 0 and still have f(x) in the required "range". In other words, we might be interested from a practical point of view in knowing how small x had to be in order to guarantee that f(x) was "close enough" to 6. Of course, there is nothing sacred about the choice of 0.0001. We could have chosen any amount of "tolerance"; and we will denote this amount by ε. [While we observe that ε could be any size, from a practical point of view we usually think of ε as being very small. This is because in most problems where we would be interested in finding a tolerance limit, it is clear that there is not too much tolerance given. That is, to think of ε as being large would correspond in practice to saying that we would like a length of 6 inches give or take 10 yards!]

Correspondingly, we denote by δ how close to zero x must be chosen to guarantee that we meet the tolerance limits.

Thus from a practical point of view, when we are given an expression such as (2), what we are really interested in is for a given ε to be able to determine δ such that f(x) will be within ε of 6 provided that x is within δ of 0.

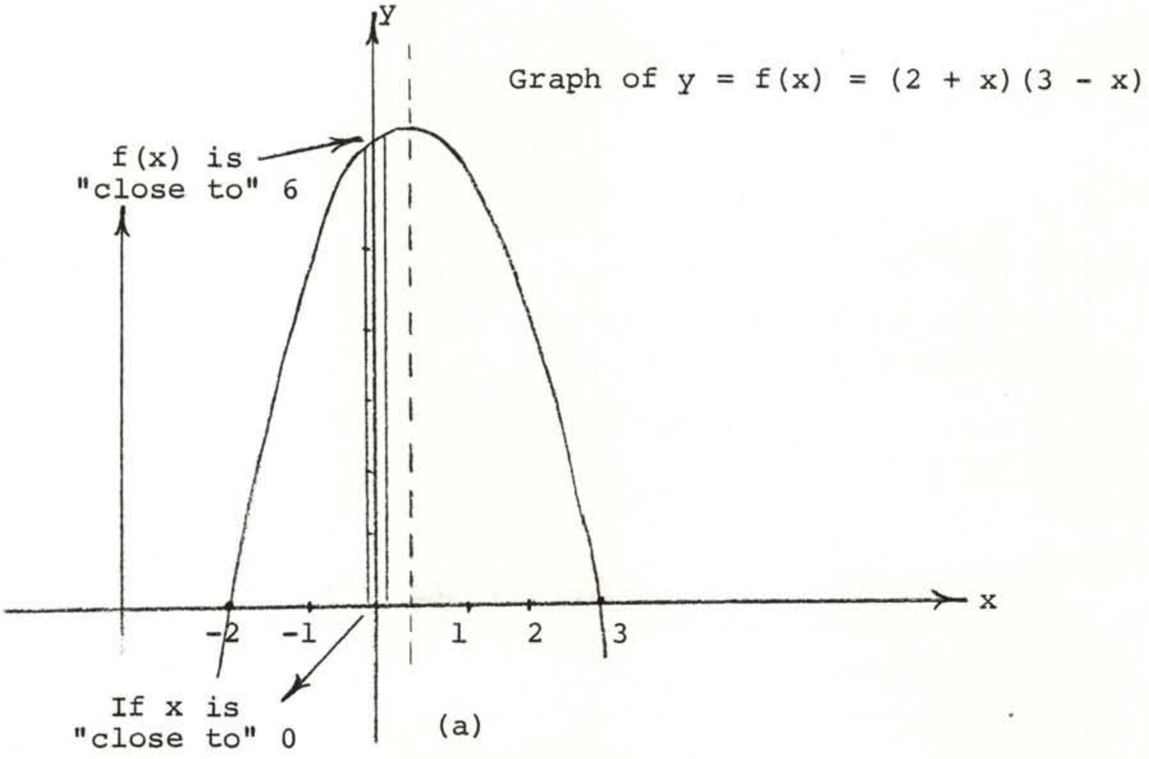
While this is not too difficult a point to grasp from a qualitative point of view, it does often turn out that it is an exceedingly difficult problem from a quantitative point of view. We feel that it is worth the experience of our "plowing through" one such problem in the hope that you will get a better feeling for what ε and δ mean from a practical point of view.

In what follows we shall be working specifically with

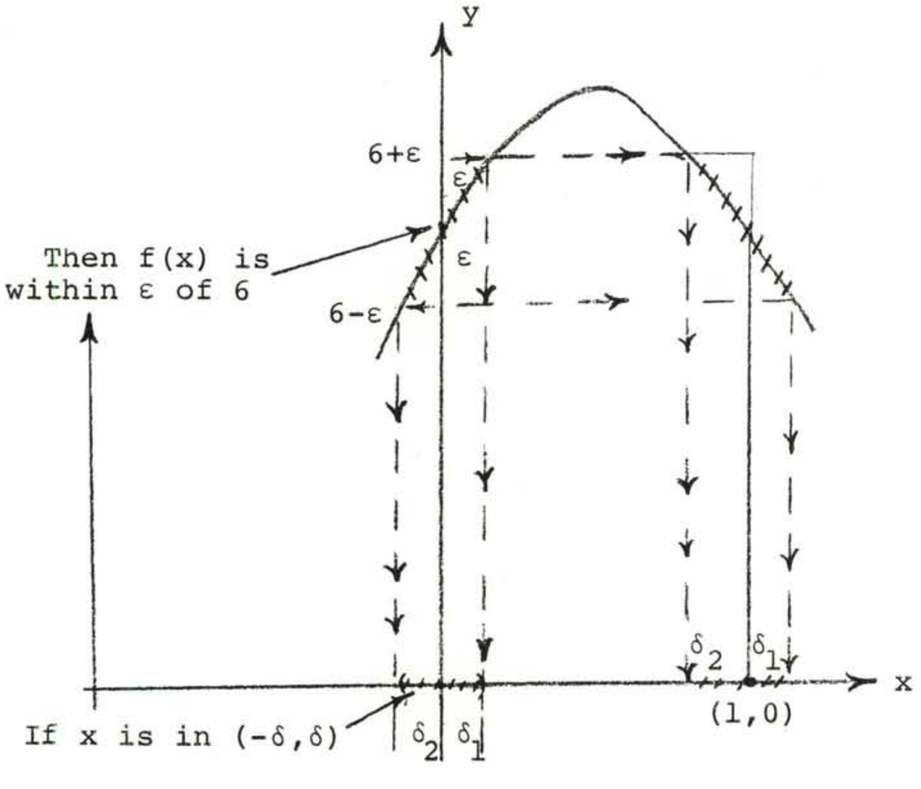
$$f(x) = (2 + x)(3 - x)$$

and we want to determine δ so that for a given ε, $f(x)$ will differ from 6 by no more than ε provided that x differs from 0 by no more than δ. In more formal language, given ε, we want to find δ so that $|x| < \delta$ implies that $|f(x) - 6| < \varepsilon$.

Graphical situation:



"Enlarged Scale"



(b)

Some Notes:

(1) Notice a difference between "Local" and "Global". There are two neighborhoods (one near 0 and the other near 1) where f(x) is within ε of 6.

We are only interested in what happens near 0. That's what $x \to 0$ means.

(2) Note that δ_1 need not equal δ_2. All we have to do to "correct" this is to let

$$\delta = \min(\delta_1, \delta_2)$$

Notice once again how helpful graphs are in helping us "size up" what is happening. In the present example, diagram (b) shows how to construct δ given ε. We simply locate 6 + ε and 6 − ε on the y-axis and project these horizontally until we intersect the curve. At the points of intersection we drop perpendiculars to the x-axis and we thus determine the intervals in which f(x) is within ε of 6. In the present example we find two such intervals one of which includes 0 and the other of which includes 1 (this is because f(x) = 6 for both x = 0 and x = 1); and there are no other values of x for which f(x) differs from 6 by less than ε. In terms of our specific problem we are interested only in the neighborhood which includes 0.

We are interested, however, in the analytical approach and this is much harder, for now we want to determine δ as a FUNCTION of ε so that we can actually compute δ for the given ε. (In many real-life examples, the expression for x is sufficiently cumbersome as to completely discourage us from finding δ exactly for a given ε. In such cases, we can still use the geometric approach in the sense that we can sketch f(x) with a sufficient degree of accuracy and employ the technique of diagram (b) to pin down the required interval pictorially, whereupon we merely measure what δ is.)

At any rate, we could proceed as follows:

We want $|f(x) - 6| < \varepsilon$.

Now $f(x) - 6 = (2 + x)(3 - x) - 6 = 6 + 3x - 2x - x^2 - 6$

$$= x - x^2$$

$\therefore |f(x) - 6| < \varepsilon$ means:

$$|x - x^2| < \varepsilon$$

Hence:

$$-\varepsilon < x - x^2 < \varepsilon \qquad (1)$$

Of course (1) is equivalent to the <u>two</u> equations:

$$\begin{cases} x - x^2 < \varepsilon & (2) \\ x - x^2 > -\varepsilon & (3) \end{cases}$$

Working on (2), we would want to solve $x - x^2 < \varepsilon$, which is equivalent to:

$$x^2 - x + \varepsilon > 0 \qquad (2')$$

A convenient device for handling (2') is to treat it as if it were an equality rather than an inequality. Namely, we solve

$$x^2 - x + \varepsilon = 0 \qquad (2")$$

to obtain by the quadratic formula:

$$\boxed{x = \frac{1 \pm \sqrt{1 - 4\varepsilon}}{2}} \qquad (4)$$

The "tough" thing is to interpret what (4) means. It is not hard to see that (4) yields the two values of x for which $f(x) = 6 + \varepsilon$. (As a check let us look at $x = \frac{1 + \sqrt{1-4\varepsilon}}{2}$ Then $2 + x = 2 + \frac{(1 + \sqrt{1-4\varepsilon})}{2}$

$$= \frac{5 + \sqrt{1-4\varepsilon}}{2}$$

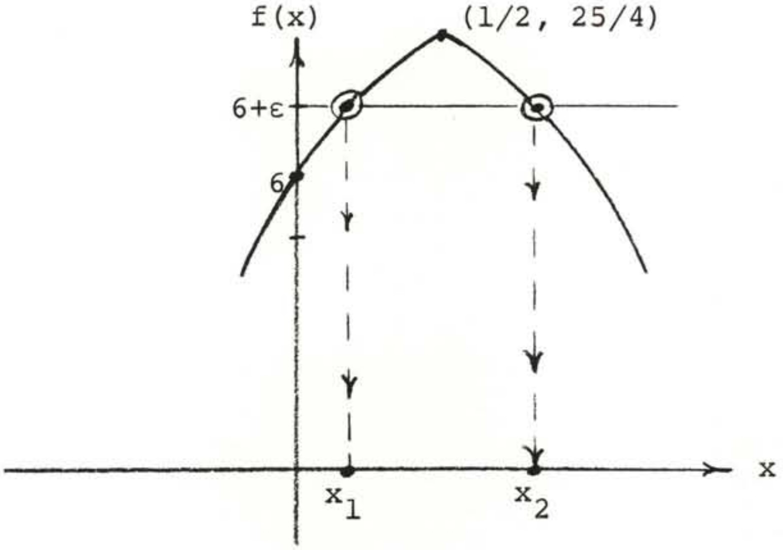
Similarly

$$3 - x = \frac{5 - \sqrt{1-4\varepsilon}}{2}$$

$$\therefore (2+x)(3-x) = \frac{(5 + \sqrt{1-4\varepsilon})}{2}\,\frac{(5 - \sqrt{1-4\varepsilon})}{2} = \frac{25 - (1-4\varepsilon)}{4} = 6 + \varepsilon$$

A similar procedure shows that if $x = \frac{1 - \sqrt{1-4\varepsilon}}{2}$, $f(x) = 6 + \varepsilon$.

Again, in terms of a graph:



(c)

Notice that since $x_1 < x_2$, x_1 corresponds to $\frac{1 - \sqrt{1-4\varepsilon}}{2}$ while $x_2 = \frac{1 + \sqrt{1-4\varepsilon}}{2}$.

(Recall we are interested in "small" values of ε. Hence we may assume that 4ε is also small $\therefore 0 < 1-4\varepsilon < 1$ $\therefore 0 < \sqrt{1-4\varepsilon} < 1$ $\therefore -\sqrt{1-4\varepsilon} < +\sqrt{1-4\varepsilon}$. Also note that if $\varepsilon > 1/4$, $1-4\varepsilon < 0$. This means $\sqrt{1-4\varepsilon}$ is imaginary. This merely verifies that $f(x) \leqslant 6 + 1/4$ for all x [see graph].)

At any rate it is now clear from our graph that since $f\left(\frac{1 - \sqrt{1-4\varepsilon}}{2}\right) = 6 + \varepsilon$,

$$\boxed{x < \frac{1 - \sqrt{1-4\varepsilon}}{2} \rightarrow f(x) < 6 + \varepsilon} \qquad \text{(A)}$$

Our job is now "half-done". That is, we have an UPPER BOUND on x. We also require a lower bound; and to this end, we return to (3); which is equivalent to

$$x^2 - x - \varepsilon < 0 \tag{3'}$$

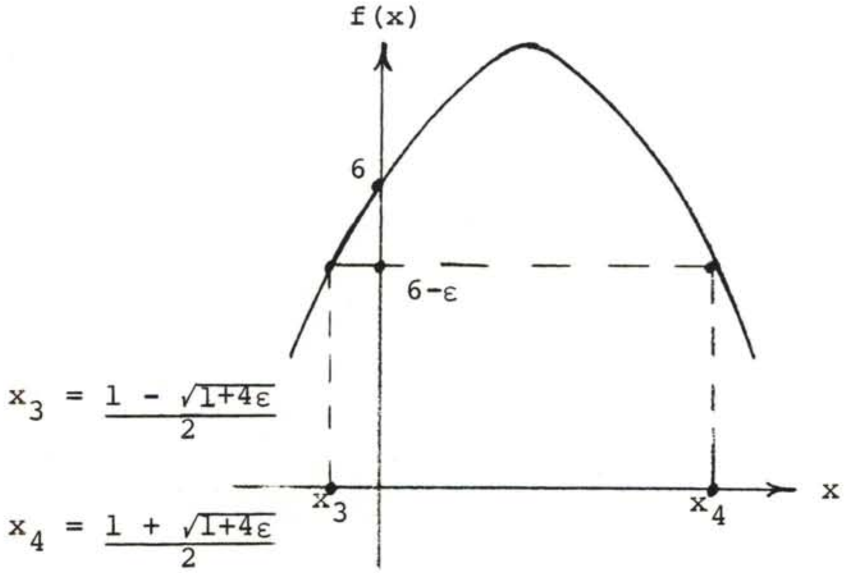
Again, we elect to look at:

$$x^2 - x - \varepsilon = 0 \tag{3"}$$

which leads to:

$$\boxed{x = \frac{1 \pm \sqrt{1+4\varepsilon}}{2}} \tag{5}$$

In a manner completely analogous to our previous treatment, it is readily verified that the values of x determined by (5) are precisely those for which $f(x) = 6 - \varepsilon$. $\frac{1 - \sqrt{1+4\varepsilon}}{2}$ corresponds to the smaller value while $\frac{1 + \sqrt{1+4\varepsilon}}{2}$ corresponds to the larger value. Again, in terms of a picture:


f(x)
6
6-ε
$x_3 = \frac{1 - \sqrt{1+4\varepsilon}}{2}$
$x_4 = \frac{1 + \sqrt{1+4\varepsilon}}{2}$
x_3
x_4
x

(d)

Thus a glance at (d) tells us that

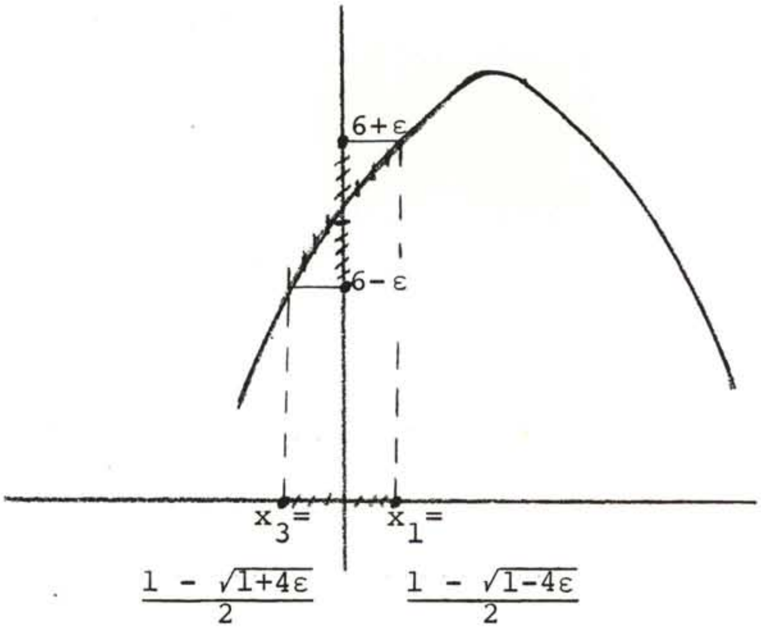
$$\boxed{x > \frac{1 - \sqrt{1+4\varepsilon}}{2} \rightarrow f(x) > 6 - \varepsilon} \qquad \text{(B)}$$

(Observe that in A and B we restrict our attention only to the neighborhood of 0 since we are not, <u>in this problem</u>, interested in what happens as x → 1.)

If we now combine (A) and (B), we obtain

$$\boxed{\begin{aligned} &\frac{1 - \sqrt{1+4\varepsilon}}{2} < x < \frac{1 - \sqrt{1-4\varepsilon}}{2} \rightarrow 6 - \varepsilon < f(x) < 6 + \varepsilon \\ &\qquad\qquad\qquad\qquad\qquad\qquad \rightarrow |f(x) - 6| < \varepsilon \end{aligned}} \qquad \text{(C)}$$

Pictorially, we may superimpose diagrams (c) and (d) to obtain the same result. Thus:



$6+\varepsilon$
$6-\varepsilon$
$x_3 =$
$\frac{1 - \sqrt{1+4\varepsilon}}{2}$
$x_1 =$
$\frac{1 - \sqrt{1-4\varepsilon}}{2}$

(e)

(By the way, a quick glance at (e) makes it clear that x_3 is negative while x_1 is positive.

Thus $|x_3| = -x_3 = \frac{\sqrt{1+4\varepsilon} - 1}{2}$ while $|x_1| = x_1 = \frac{1 - \sqrt{1-4\varepsilon}}{2}$

It need not be true that $|x_1| = |x_3|$. In any event the required δ is merely the smaller of $|x_1|$ and $|x_3|$. More symbolically

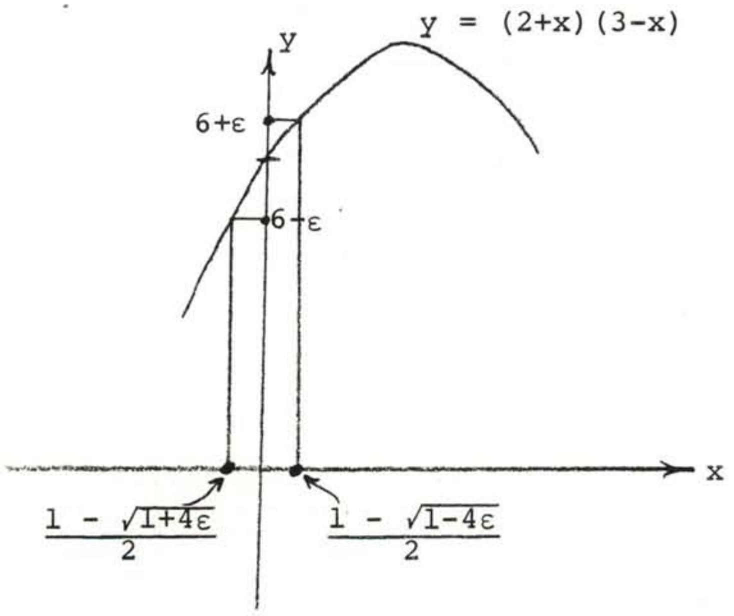
$$\delta = \min(|x_1|, |x_3|)$$

(You see, while $6 + \varepsilon$ and $6 - \varepsilon$ are symmetrically located around 6, the slope of the graph of $f(x)$ varies. Hence there is no reason for $|x_1| = |x_3|$ as would have been the case had the graph of $f(x)$ been a straight line [why?].)

In summary, if $f(x) = (2 + x)(3 - x)$ then $\lim\limits_{x \to 0} f(x) = 6$.

Moreover, given $\varepsilon > 0$;

$$\frac{1 - \sqrt{1+4\varepsilon}}{2} < x < \frac{1 - \sqrt{1-4\varepsilon}}{2} \rightarrow 6 - \varepsilon < f(x) < 6 + \varepsilon$$


y = (2+x)(3-x)
y
6+ε
6-ε
x
$\frac{1 - \sqrt{1+4\varepsilon}}{2}$
$\frac{1 - \sqrt{1-4\varepsilon}}{2}$

In the remainder of this course, the concept of limits will remain the underlying theme. For this reason it is crucial that we recognize both the geometric (intuitive) properties and the analytic properties of them. For obvious reasons, we will spell things out geometrically whenever we can. It must be observed, however, that there will be times when only analytic methods will be applicable. In particular, this will happen when we deal with functions of several variables. Our approach will be to capitalize on the geometric picture whenever it is available. We will then "translate" the picture into its equivalent analytic form (after all, this is what analytic geometry is all about) and we will finally extend this analytic form to those cases for which there exists no geometric interpretation.

Chapter V

MATHEMATICAL INDUCTION

A. Introduction

Mathematical induction is a rather powerful technique for proving certain types of theorems, *once we have a "hunch" as to what the correct answer is*. Let us illustrate what we mean through the use of a specific problem.

Recall that in our discussion of sets, we proved by means of circle-diagrams that for any three sets A, B, and C:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{1}$$

Suppose now that we have four sets A, B, C, and D. Can we somehow make use of (1) to derive an expression for $A \cap (B \cup C \cup D)$? To tackle this problem we proceed, as is so often the case in mathematics, by trying to reduce an unfamiliar problem to a more familiar one. Since we have already shown that union is associative, we may write that $B \cup C \cup D$ is equal to $(B \cup C) \cup D$, and we may rewrite $A \cap (B \cup C \cup D)$ as $A \cap [(B \cup C) \cup D]$. The advantage in this is that the bracketed expression is now the union of two sets, $(B \cup C)$ *and* D (that is, the union of two sets is *a* set, just as 3 + 2 is *a* number, namely 5, not two numbers). Thus by (1), we may write that

$$A \cap [(B \cup C) \cup D] = [A \cap (B \cup C)] \cup (A \cap D)^{*} \tag{2}$$

*If this seems a bit vague the following intermediate steps may prove helpful. Let $E = B \cup C$. Then by substitution, $A \cap [(B \cup C) \cup D] = A \cap (E \cup D)$, and by (1) this is in turn equal to $(A \cap E) \cup (A \cap D)$. Again replacing E by $B \cup C$, we obtain $[A \cap (B \cup C)] \cup (A \cap D)$.

Utilizing (1) in (2), we next obtain:

$$A \cap [(B \cup C) \cup D] = [(A \cap B) \cup (A \cap C)] \cup (A \cap D).$$

Finally recalling that $[(A \cap B) \cup (A \cap C)] \cup (A \quad D)$ is equal to $(A \cap B) \cup (A \cap C) \cup (A \cap D)$, we obtain the result that:

$$A \cap (B \cup C \cup D) = (A \cap B) \cup (A \cap C) \cup (A \cap D)^* \qquad (3)$$

Equation (3) tells us that intersection is distributive over the union of three sets, just as (1) told us that intersection was distributive over the union of two sets.

At this stage, we might have a "hunch" that intersection would be distributive over the union of any finite number of sets. To state our "hunch" more mathematically, we might say the following:

Let $A_1, A_2, \ldots, A_n$ be sets.** Then let P(n) denote a proposition concerning the n sets. Specifically, let P(n) be the proposition that

$$A_1 \cap (A_2 \cup A_3 \cup \ldots \cup A_n) = (A_1 \cap A_2) \cup (A_1 \cap A_3) \cup \ldots \cup (A_1 \cap A_n) \qquad (4)$$

*At this point one might wonder why we couldn't obtain (3) by circle-diagrams just as we did for (1). Circle-diagrams rapidly become unwieldy (even impossible) for large numbers of sets. Try for example to draw the proper circle diagram for four sets. Notice that for n sets the number of regions in our circle-diagram is 2^n. (For example if n = 1 there are two regions, A_1 and A_1'; if n = 2 there are four regions, $A_1 \cap A_2'$, $A_1' \cap A_2$, $A_1' \cap A_2'$, and $A_1 \cap A_2$, etc.) Thus if n = 4 we would have 16 regions. If n = 20 we would have over 1,000,000 regions since $2^{20} > 1,000,000$.

**Note that, as we switch from three or four sets to n sets, we are better off using subscripted symbols A_1, A_2, A_3, etc., rather than A, B, C, etc. Specifically, we do <u>not</u> want to use A, B, C ... N which denotes exactly 14 sets!

At this time it is crucial that we realize that $P(n)$ is only a conjecture. The mere fact that we have observed that $P(n)$ happens to be true when $n = 3$ and when $n = 4$, i.e. that $P(3)$ and $P(4)$ are true, does not in itself mean that $P(n)$ will be true for every positive integer, n.

For example, consider the following: Let n denote any positive integer, and consider the proposition, Q, defined by:

$$Q(n) \text{ means that } n^2 - n + 41 \text{ is a prime number*} \qquad (5)$$

Let us compute $Q(n)$ for consecutive integers starting with 1. We obtain:

$Q(1) = 1^2 - 1 + 41 = 41$, and 41 is a prime number. $Q(1)$ is true.

$Q(2) = 2^2 - 2 + 41 = 43$, and 43 is a prime number. $Q(2)$ is true.

$Q(3) = 3^2 - 3 + 41 = 47$, and 47 is a prime number. $Q(3)$ is true.

$Q(4) = 4^2 - 4 + 41 = 53$, and 53 is a prime number. $Q(4)$ is true.

$Q(5) = 5^2 - 5 + 41 = 61$, and 61 is a prime number. $Q(5)$ is true.

At this stage, we might be tempted to make the conjecture that $Q(n)$ will always be true. That is, for any positive integer, n, $n^2 - n + 41$ will denote a prime number. The interesting thing about (5) is that $Q(n)$ is true for all positive integers, n, from 1 through 40 inclusive, However, when $n = 41$, (5) becomes:

$$(41)^2 - 41 + 41 \text{ is a prime number,}$$

and this is false since $(41)^2 - 41 + 41 = 41^2 = 41 \times 41$ which certainly is not a prime number, thus our proposition is true for $n = 1,2,\ldots, 40$; but false for $n = 41$.

*An integer $n > 1$ is called a prime number if the only whole numbers which are divisors of n are n and 1 (1 is not called a prime number).

How then can we rigorously establish the truth of our conjecture (4) without worrying that things break down as in our last example. Hopefully, we would recognize a scheme similar to that of how we got from (1) to (3). For example, suppose that we know that (1) and (3) are true and we wanted to investigate $A \cap (B \cup C \cup D \cup E)$. We would write

$$A \cap (B \cup C \cup D \cup E) =$$
$$A \cap ([B \cup C \cup D] \cup E)^* =$$
$$(A \cap [B \cup C \cup D]) \cup (A \cap E) \text{ [by (1)] } =$$
$$[(A \cap B) \cup (A \cap C) \cup (A \cap D)] \cup (A \cap E) \text{ [by (3)] } =$$
$$(A \cap B) \cup (A \cap C) \cup (A \cap D) \cup (A \cap E)^*$$

If we now switch to the notation of (4), we have a new insight concerning $A_1 \cap (A_2 \cup \ldots \cup A_n) = (A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_n)$. Namely, our procedure indicates that once we knew that P(n) was true for n = 3, we <u>could prove</u> that it was true for n = 4. Once we knew it was true for n = 4 we could prove it was true for n = 5, and it begins to look as though we can continue this trend as long as we wish - but how can we be sure?

Well, suppose we could <u>prove</u> that whenever P(n) was true for n = k, it must also be true for n = k + 1. (Notice we are <u>not</u> saying that P(n) <u>is</u> true for n = k, but rather <u>if</u> P(n) is true for n = k.) Then if we <u>do</u> know P(n) is true for n = 3, it <u>must</u> be true for n = 4. But, if it is true for n = 4, it <u>must</u> then be true for n = 5, and the truth of P(5) implies the truth of P(6) etc.

*Here we are assuming that union is associative for four sets rather than only for three. If this is too much to accept on faith, let us merely agree that unless otherwise stated the union of any number of sets implies that the union is in the order in which the sets are written. That is, $B \cup C \cup D \cup E$ means $[(B \cup C) \cup D] \cup E$. With this convention $B \cup C \cup D \cup E = [(B \cup C) \cup D] \cup E = [B \cup C \cup D] \cup E$, etc.

To this end, assume that

$$A_1 \cap (A_2 \cup \ldots \cup A_k) = (A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_k) \quad (6)$$

(Notice, again, that we are not saying we know that P(k) is true.) We must show that this implies

$$A_1 \cap (A_2 \cup \ldots \cup A_{k+1}) = (A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_{k+1}) \quad (7)$$

Now:

$$\begin{aligned}
&A_1 \cap (A_2 \cup \ldots \cup A_{k+1}) = \\
&A_1 \cap (A_2 \cup \ldots \cup A_k \cup A_{k+1}) = \\
&A_1 \cap ([A_2 \cup \ldots \cup A_k] \cup A_{k+1}) = \\
&(A_1 \cap [A_2 \cup \ldots \cup A_k]) \cup (A_1 \cap A_{k+1}) \text{ [by (1)] } = \\
&[(A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_k)] \cup (A_1 \cap A_{k+1}) \text{ [by (6)] } = \\
&(A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_k) \cup (A_1 \cap A_{k+1})
\end{aligned}$$

We have just shown that the truth of (6) does imply the truth of (7).

Summed up, we now know about our proposition that:

(a) $P(n)$ is true when $n = 3$.

(b) If $P(n)$ is true for $n = k$, it is also true for $n = k+1$.

These two facts are all we need. For if we now let $k = 3$ in (3) (the one value for which we know $P(n)$ is true), we immediately learn that $P(n)$ is true for $k + 1 = 3 + 1 = 4$. That is, we now know that $P(4)$ is true. Then, taking this output and feeding it back into the input, we let $k = 4$. That makes $P(5)$ true, etc. Notice that this cannot fail for $P(41)$ or any other finite number and we have proved that (4) is true for all integers greater than or equal to 3, i.e. that $P(n)$ is true for $n \geqslant 3$. (In this case, notice that for $n < 3$, $P(n)$ doesn't make sense.)

The discussion of (4) can be generalized as follows, and is known as the principle of mathematical induction.

Let $P(n)$ denote a proposition that is defined for each positive integer n. Suppose that we know (a) that $P(n)$ is true for say $n = m$ and (b) that we can prove that whenever $P(n)$ is true for $n = k$, it is also true for $n = k + 1$. Then $P(n)$ is true for all integers which are greater than or equal to m. In our example, $P(n)$ was the proposition that for n sets, $A_1, A_2, \ldots A_n$,

$$A_1 \cap (A_2 \cup \ldots \cup A_n) = (A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_n)$$

and m was equal to 3; we proved that (4) was true for all $n \geqslant 3$.

While we shall talk more about this later, let us make a few warning remarks. Notice that to use mathematical induction, we must first be able to conjecture a proposition. In terms of our present example, notice that we did not use induction to find an equivalent expression for $A_1 \cap (A_2 \cup \ldots \cup A_n)$. Rather it was only after we were able to make the conjecture (4) that we used induction. In still other words, we must have an expression for both sides of the equation before we use induction. Secondly, notice that induction applies to the positive integers and not to the set of all real numbers. For example suppose we knew that $P(1/2)$ was true and that whenever $P(k)$ was true so also was $P(k+1)$. Then we could conclude that $P(1/2)$, $P(3/2)$, $P(5/2)$, etc. were all true. But we couldn't make any statement about other values of n.

The point is that if n is an integer then there are no other integers which exceed n but are less than $n + 1$. On the other hand, there are many (infinitely many) real numbers which lie between n and $n+1$.

We shall return to this discussion of limitations in Section C. For now we prefer to reinforce our definition of induction by applying the discussion of this section to another topic previously studied in our course - limits. It is our hope that applications to specific instances will help clarify the concept better than an abstract, philosophical discussion.

B. An Application to Limits

In the course of our studies about limits we proved that:

$$\lim_{x\to a} [f(x) + g(x)] = \lim_{x\to a} f(x) + \lim_{x\to a} g(x), \text{ provided } \lim_{x\to a} f(x)$$

$$\text{and } \lim_{x\to a} g(x) \text{ exist.}$$

To emphasize that we are dealing with the sum of two functions, perhaps a better notation would have been:

$$\boxed{\lim_{x\to a} [f_1(x) + f_2(x)] = \lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x)} \qquad (\underline{1})$$

Suppose now we were asked to find a similar formula for $\lim_{x\to a} [f_1(x) + f_2(x) + f_3(x)]$. For the sake of our illustration let us refrain from making any conjecture here and proceed instead by mathematical logic.

Since $f_1(x) + f_2(x) + f_3(x) = [f_1(x) + f_2(x)] + f_3(x)$ (why?) we may write:

$$\lim_{x\to a} (f_1(x) + f_2(x) + f_3(x)) =$$

$$\lim_{x\to a} ([f_1(x) + f_2(x)] + f_3(x))$$

But as we have already seen in our study of functions, $f_1 + f_2$ is also <u>a</u> function.

So by (1) we have

$$\lim_{x\to a} ([f_1(x) + f_2(x)] + f_3(x)) = \lim_{x\to a} [f_1(x) + f_2(x)]^* + \lim_{x\to a} f_3(x)$$

But by (1) $\lim_{x\to a} [f_1(x) + f_2(x)] = \lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x)$

In summary:

$$\lim_{x\to a} (f_1(x) + f_2(x) + f_3(x)) =$$
$$\lim_{x\to a} ([f_1(x) + f_2(x)] + f_3(x)) =$$
$$\lim_{x\to a} ((f_1(x) + f_2(x)) + \lim_{x\to a} f_3(x) =$$
$$[\lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x)] + \lim_{x\to a} f_3(x) =$$
$$\lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x) + \lim_{x\to a} f_3(x)$$

Knowing that

$$\lim_{x\to a} [f_1(x) + f_2(x) + f_3(x)] = \lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x) + \lim_{x\to a} f_3(x) \qquad (\underline{2})$$

---

*Recall the intermediate step that we used in the previous section:

Let $h(x) = f_1(x) + f_2(x)$

Then $\lim_{x\to a} ([f_1(x) + f_2(x)] + f_3(x)) = \lim_{x\to a} (h(x) + f_3(x))$

By (1) $\lim_{x\to a} (h(x) + f_3(x)) = \lim_{x\to a} h(x) + \lim_{x\to a} f_3(x)$

but by definition of $h(x)$, $\lim_{x\to a} h(x) = \lim_{x\to a} [f_1(x) + f_2(x)]$

$\therefore \lim_{x\to a} ([f_1(x) + f_2(x)] + f_3(x)) = \lim_{x\to a} [f_1(x) + f_2(x)] + \lim_{x\to a} f_3(x)$

we can show that:

$$\lim_{x\to a} [f_1(x) + f_2(x) + f_3(x) + f_4(x)] = \lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x) + \lim_{x\to a} f_3(x) + \lim_{x\to a} f_4(x)$$

Namely:

$$\begin{aligned}
&\lim_{x\to a} [f_1(x) + f_2(x) + f_3(x) + f_4(x)] = \\
&\lim_{x\to a} [\{f_1(x) + f_2(x) + f_3(x)\} + f_4(x)] = \\
&\lim_{x\to a} \{f_1(x) + f_2(x) + f_3(x)\} + \lim_{x\to a} f_4(x) \quad (\text{by } (\underline{1})) = \\
&\{\lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x) + \lim_{x\to a} f_3(x)\} + \lim_{x\to a} f_4(x) \quad (\text{by } (\underline{2}))
\end{aligned}$$

and the result follows.

This procedure practically begs the mathematical induction approach. Specifically if we let P(n) denote the proposition that

$$\lim_{x\to a} [f_1(x) + \ldots + f_n(x)] = \lim_{x\to a} f_1(x) + \ldots + \lim_{x\to a} f_n(x)$$

What we have already shown is that

P(2), P(3), and P(4) are true.

More importantly our procedure in going from n = 2 to n = 3 and from n = 3 to n = 4 seems to dictate a way of reducing P(k+1) to P(k). Namely:

$$\begin{aligned}
&\lim_{x\to a} [f_1(x) + \ldots + f_k(x) + f_{k+1}(x)] = \\
&\lim_{x\to a} [\{f_1(x) + \ldots + f_k(x)\} + f_{k+1}(x)] = \\
&\lim_{x\to a} \{f_1(x) + \ldots + f_k(x)\} + \lim_{x\to a} f_{k+1}(x) \quad (\text{by } (1)) \qquad (\underline{3})
\end{aligned}$$

From (3) it should be clear that all we need now is that $\lim_{x\to a} [f_1(x) + \ldots + f_k(x)] = \lim_{x\to a} f_1(x) + \ldots + \lim_{x\to a} f_k(x)$ and this outlines the entire approach. In retrospect P(2) is true. We now assume P(k) is true [that is, $\lim_{x\to a} [f_1(x) + \ldots + f_k(x)\ ] = \lim_{x\to a} f_1(x) + \ldots + \lim_{x\to a} f_k(x)$] and we must show that this implies the truth of P(k+1) [that is, $\lim_{x\to a} [f_1(x) + \ldots + f_k(x) + f_{k+1}(x)] = \lim_{x\to a} f_1(x) + \ldots + \lim_{x\to a} f_{k+1}(x)$]. We then mimic what we did above. Thus: $\lim_{x\to a} [f_1(x) + \ldots + f_{k+1}(x)] = \ldots = \lim_{x\to a} \{f_1(x) + \ldots + f_k(x)\} + \lim_{x\to a} f_{k+1}(x)$ and by the assumption that P(k) is true, the result follows.

C. Limitations of Induction.

There are two major limitations to the use of induction. For one thing, it is possible that some proposition P(n) is indeed true for all n, but that structurally the truth of P(k+1) in no way depends on the truth of P(k). In still other words, both P(k) and P(k+1) can be true but for completely independent reasons. As a mild example, let P(n) denote the statement that any positive integer $n > 1$ can be factored uniquely as a product of primes (this is often called the fundamental theorem of arithmetic). While we won't prove this theorem, it is not difficult to conjecture its truth. Yet a glance at how the various integers factor into primes also makes it plausible to believe that there is no structural pattern whereby we can factor n+1 just by knowing how to factor n. For example:

$$
\begin{aligned}
2 &= 2\\
3 &= 3\\
4 &= 2 \times 2\\
5 &= 5\\
6 &= 2 \times 3\\
7 &= 7\\
8 &= 2 \times 2 \times 2\\
9 &= 3 \times 3\\
10 &= 2 \times 5\\
11 &= 11\\
12 &= 2 \times 2 \times 3
\end{aligned}
$$

In fact in trying to determine whether a given number is a prime we must use trial-and-error techniques and cannot tell too much by looking at the numbers that come before it.

In summary, then, one limitation to mathematical induction is that it is possible that P(n) can be true for all n but that the result cannot be shown by induction.

A second weakness with induction is that it requires that we have a conjecture. Notice that in both of the examples in the previous section, we first speculated on what we felt was a "sensible" conjecture and then tried the induction technique. There are, however, situations that do not lend themselves that easily to conjecture. By way of example, the following is a problem that seems to appear in virtually every textbook that describes induction:

Prove that for each positive integer n,

$$1 + 2 + \ldots. + n = \frac{n(n+1)}{2} \qquad (1)$$

Now once (1) is given, we certainly have a conjecture. Indeed we can test (1) for various values of n and find that: $1 = 1x2/2$, $1 + 2 = 2x3/2$, $1 + 2 + 3 = 3x4/2$, $1 + 2 + 3 + 4 = 4x5/2$ are true statements. (To reinforce the notation of P(n), what this last statement says is that if we let P(n) denote the proposition that $1 + 2 + \ldots. + n = \frac{n(n+1)}{2}$, then P(1), P(2), P(3), P(4) are all true and at least we see that (1) is a plausible conjecture.) It is left as an exercise to show the truth of this conjecture by mathematical induction.

The issue we wish to raise by (1), though, is: How likely is it that if we had not been given the recipe in (1) we would have discovered it by trial-and-error? In other words, suppose instead of being given (1), we were asked to find a "convenient" expression for computing $1 + 2 + 3 + \ldots. + n$.

Certainly we could have looked at the sum of the first n integers and found for n = 1,2,3,4,5, etc. that these sums were:

1, 3, 6, 10, 15, 21, 28, etc.

But is it likely that even with a large list of computed values we would have hit on the hunch that the nth sum was simply n(n+1)/2? It's possible but, for most of us, not likely. (For this reason, such a textbook illustration is at best contrived.)

We can use this same problem to make another point. There is an anecdote connected with the famous mathematician, Gauss. The story is told that as a punishment young Gauss was told to compute the sum 1 + 2 + 3 + 4 + ... + 197 + 198 + 199 + 200. He observed that the first and the last terms formed a sum of 201, as did the second and next to the last, and in this way he noted that there were 100 pairs each of sum 201. Thus, he quickly concluded that the required sum was 20,100.

This result is readily generalized (and you may recall seeing it in high school algebra under the topic of arithmetic progressions) as follows:

To compute the sum 1 + 2 + 3 + ... + (n-2) + (n-1) + n, write the sum twice, but once in the reverse order. That is:

$$\begin{array}{lcccccccccccc}
\text{Let } S = & 1 & + & 2 & + & 3 & + & (n-2) & + & (n-1) & + & n \\
 & \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\
\text{then } S = & n & + & (n-1) & + & (n-2) & + & 3 & + & 2 & + & 1.
\end{array}$$

Upon adding these two rows, we obtain:

$$2S = (n+1) + (n+1) + \dots + (n+1)$$
$$= n(n+1) \text{ [since there are n terms being added each of which yields n+1 as a sum]}$$

Dividing through by 2, we obtain:

$$S = n(n+1)/2,$$

which is not only the correct and desired result but it in no way depends on induction. That is, we can determine the sum of the first 600 integers in this way without first having to know the sum of the first 599 positive integers.

In summary, in certain problems the proper conjecture is difficult to come by (and certainly this is a subjective problem in the sense that it might be difficult for one person but not for another). Moreover, in those cases where we cannot sense the proper conjecture the fact of life is that either we will construct a more direct proof for determining the correct result as did Gauss or else the problem will remain unsolved. By way of review, the point we are making is that if we were given the problem of forming the sum of the first n integers, we might never have stumbled on the proper conjecture; and that perhaps the original proof of (1) was not by induction at all but rather by a method similar to Gauss'.

At the same time that we point out this latter weakness, however, let us also point out that in the two examples of the previous sections, it seemed "almost natural" to guess not only the conjecture but how one would proceed from $P(k)$ to $P(k+1)$. The point is that in those real-life situations where induction has the most significant use we find that it is precisely when the conjecture and the construction of the inductive proof practically dictate themselves. We shall see several examples of this throughout the rest of this course; but for now we are content to let the subject drop.

Chapter VI
INFINITESIMALS AND DIFFERENTIALS

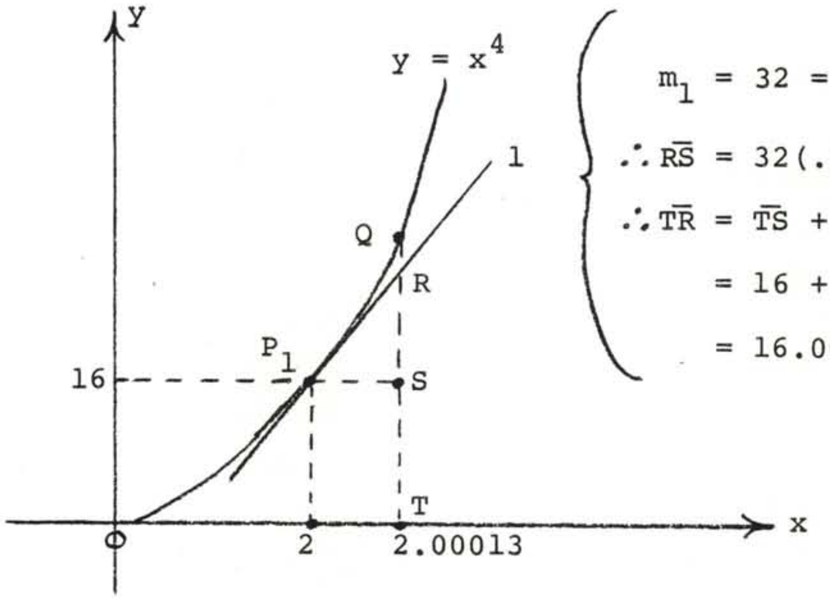
A. Introduction

It may be more than coincidence that one refers to differential calculus rather than to derivative calculus even though that branch of calculus deals with the concept of derivatives. The point we are trying to make is that the notion of a differential lies at the very foundation of calculus.

While the concept of a differential is crucial to a proper understanding of calculus, the fact remains that it is an extremely subtle concept, and unless great care is taken the true significance of a differential can be missed. In this chapter our aim is to develop the notion of a differential slowly and meaningfully in the hope that such a development will focus attention on the makeup of calculus and what the real problem is in dealing with 0/0.

We shall start with a rather simple problem. Suppose we wish to evaluate, say, $(2.00013)^4$. (Certainly, a less cumbersome expression could have been used, but we want to emphasize a certain amount of computational work that can be involved.) The most direct attack, of course, is actually to compute 2.00013 x 2.00013 x 2.00013 x 2.00013 and obtain the result 16.0041604056175762856l. Our claim is that by use of derivatives we can obtain 16.00416 as a very quick approximation. To this end, notice that from a geometric point of view, we are trying to locate the y-coordinate of the point on the curve $y = x^4$ whose x-coordinate is 2.00013.

What is known to us at little more than a glance is that (1) the point $P_1(2,16)$ is a rather convenient (i.e., easily located) "nearby" point to the one we seek, and (2) the slope of the line tangent to $y = x^4$ at (2,16) is 32 (i.e. the slope is precisely $\frac{dy}{dx}$, which in this case is $4x^3$, and at $x = 2$ this is 32). We thus have the following picture (which is deliberately drawn with an exaggerated scale so that we can see "what's going on").



$$\left\{\begin{aligned} m_1 &= 32 = \frac{\overline{RS}}{\overline{P_1S}} = \frac{\overline{RS}}{.00013} \\ \therefore \overline{RS} &= 32(.00013) = 0.00416 \\ \therefore \overline{TR} &= \overline{TS} + \overline{RS} \\ &= 16 + 0.00416 \\ &= 16.00416 \end{aligned}\right.$$

(Figure 1)

In terms of Figure 1, we are trying to find the coordinates of Q. Since $\overline{OT} = 2.00013$ we know the x-coordinate of Q. The y-coordinate of Q is $\overline{TQ}$ (which we previously computed as 16.0041604056175762856l but which we now pretend we haven't).

What we are <u>sure</u> of is that $\overline{RT}$ is <u>exactly</u> 16.00416.

Now the key idea here is that while $\overline{RT}$ is exactly 16.00416, it is also approximately equal to $\overline{TQ}$. That is, we suspect since Q is "close to" P that the length of $\overline{TQ}$ is approximately the length of $\overline{RT}$ which, in turn, is exactly 16.00416.

In any event, this technique yields the result that:

$$(2.00013)^4 \approx 16.00416 \qquad (1)$$

Again in terms of our picture we are saying that (2.00013, 16.00416) precisely names the point R on the line l [in fact the equation of l is $y = 32x - 48$ (why?)] and R is being used to approximate the location of Q.

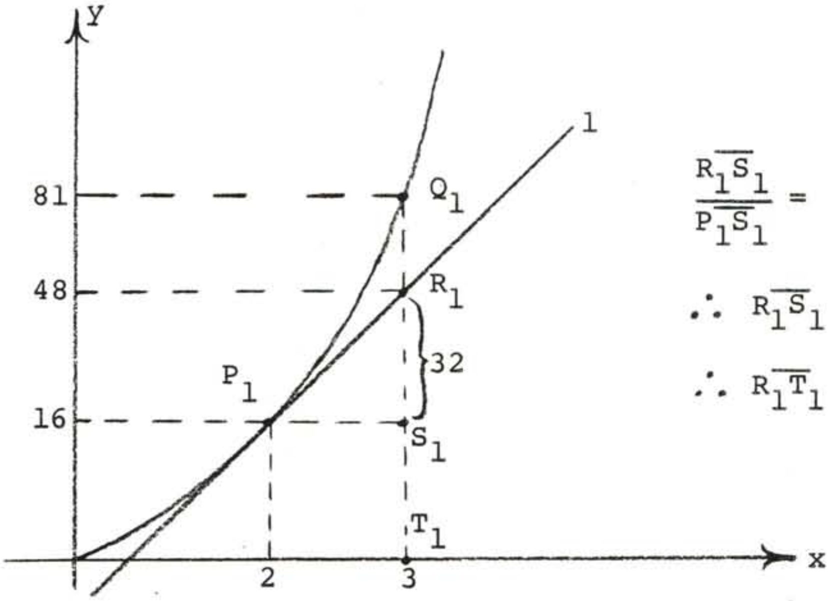
To see how accurate (1) is, recall that $(2.00013)^4 =$ 16.0041604056175762856l, so that the total error in our approximation is:

$$\underline{0.0000004056175762856l}$$

Of course, total errors are misleading, so we instead estimate the percentage error by observing that our total error is about 0.00000041 parts in about 16, hence the percentage error is about

$$\frac{(0.00000041)(100)}{16} < 0.000003\% \qquad (2)$$

We also recognize that the great degree of accuracy reflected in (2) depends on how close we are to the point of tangency. For example if we tried to use Figure 1 to find an approximate value for 3^4 we would obtain



$$\frac{\overline{R_1S_1}}{\overline{P_1S_1}} = \overline{R_1S_1} = \text{slope of } l = 32$$

$$\therefore \overline{R_1S_1} = 32$$

$$\therefore \overline{R_1T_1} = 48 \quad \therefore 3^4 \approx 48$$

(Figure 2)

In this case $R_1(3,48)$ is still a point *exactly* on l but now R_1 is *not* a "good" approximation to the location of Q_1.

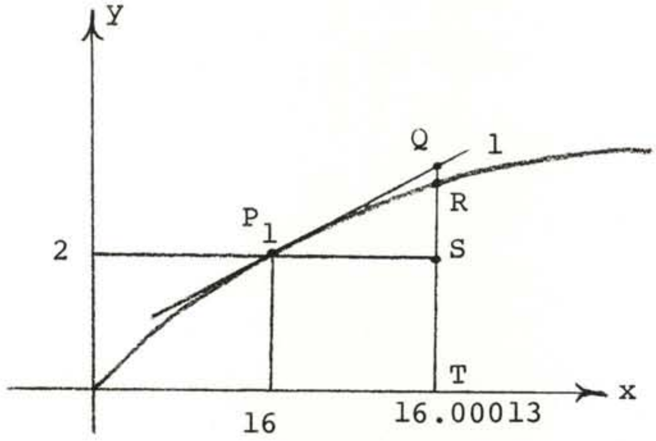
Thus one problem which confronts us is that of determining a fairly objective way for defining "reasonable" approximations.

Secondly, to find a percentage error it is required that we know the *exact* answer and this isn't always easy or even possible (if it were, the chances are we wouldn't be making approximations in such cases). In fact we chose $(2.00013)^4$ rather than, say, $\sqrt[4]{16.00013}$ because cumbersome as it might be we can compute $(2.00013)^4$ *exactly* but, at least in decimal form, we can only estimate $\sqrt[4]{16.00013}$.

Notice, of course, that our analytic process can still be used in such a case. For example, letting $y = x^{1/4}$ we see that $\frac{dy}{dx} = \frac{1}{4} x^{-\frac{3}{4}}$

$$\therefore \quad \left(\frac{dy}{dx}\right)_{x=16} = \frac{1}{4}(16)^{-\frac{3}{4}} = \frac{1}{32}$$

Thus:



$$\frac{\overline{QS}}{\overline{P_1S}} = m_1 = \frac{1}{32}$$

$$\overline{QS} = \frac{\overline{P_1S}}{32} = \frac{.00013}{32}$$

$$\approx .000004$$

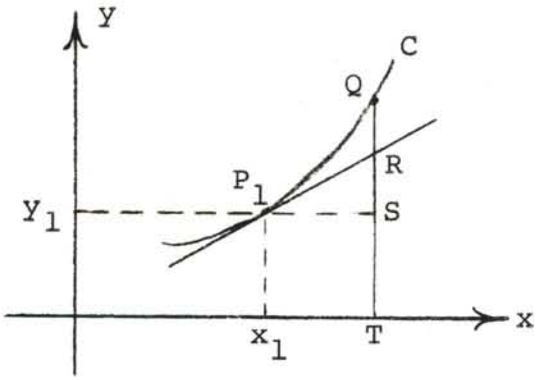
$$\therefore \overline{QT} \approx 2.000004$$

$$\therefore \overline{RT} \approx 2.000004$$

$$\therefore \sqrt[4]{16.00013} \approx 2.000004$$

It is not our aim to focus attention on finding approximations. Basically, our aim was to help set the scene for some more incisive observations that will be made in the next section.

For now, let us sum up our results in as general a way as we can. We assume that we have a curve C which is smooth in a neighborhood of $P_1(x_1,y_1)$ and we let Q denote another point on C which exists in this neighborhood. We let R denote the point at which the tangent line to C at P_1 (the existence of the tangent line is implied by the meaning of "smooth") meets the line TQ (see Figure 4):



(Figure 4)

Then $\overline{TR}$ and $\overline{TQ}$ are nearly equal in length.

Stated more mathematically, we may let Δy denote the length of $\overline{SQ}$ and Δy_{tan} denote the length of $\overline{SR}$. We are then saying that if Δx is small, the lengths of Δy and Δy_{tan} are approximately equal.

It should also be noted that we can make the same observations without reference to a graph and talk strictly in terms of functions. Namely, suppose f is defined in some neighborhood N of x_1 and that $f'(x_1)$ exists. Then if $x_1 + \Delta x$ denotes a number in N we may approximate $f(x_1+\Delta x)-f(x_1)$ by $f'(x_1)\Delta x$, and the approximation improves as $\Delta x \to 0$.

What we want to do next is to show just how rapidly $f'(x_1)\Delta x$ improves as an approximation for $f(x_1+\Delta x)-f(x_1)$.

B. Infinitesimals

As we have previously mentioned, the study of calculus is a refinement of the study of 0/0. That is, we are often interested in quotients of the form m/n where m and n both approach zero as a limit. A variable which approaches zero as a limit is called an infinitesimal. Thus, the study of differential calculus involves the quotient of two infinitesimals.

Moreover, since the quotient of two small numbers is rather unpredictable, we must often require that the numerator be a higher order infinitesimal than the denominator. That is, if the numerator grows too rapidly compared with the growth of the denominator the limit of the quotient may increase without bound.

Our claim is that this problem was already present in our discussion of the last section but it was not obvious from the pictorial point of view. What we intend to do now is revisit the material of the last section but from a more quantitative point of view. First of all let us study the difference between $\frac{\Delta y}{\Delta x}$ and $(\frac{dy}{dx})_{x=x_1}$ by letting, say, k, denote this difference (most books use ε^1[epsilon] to denote the difference but our feeling is that such a symbol might inadvertently make us think of the same epsilon that we used in our discussion of limits. The point is that these two epsilons represent different concepts).

At any rate, we have:

$$\frac{\Delta y}{\Delta x} - \left(\frac{dy}{dx}\right)_{x=x_1} = k \qquad (1)$$

As a word of caution, observe that k is a variable and its value usually will depend on the value of Δx. To see why, observe that once x_1 is fixed so is $(\frac{dy}{dx})_{x=x_1}$ but $\frac{\Delta y}{\Delta x}$ is the slope of the line which joins P_1 and Q, and the location of Q certainly depends on Δx.

We may now go one step further and establish the fact that k is actually an infinitesimal. For if we take the limit in (1) as we let Δx approach zero, we find that:

$$\lim_{\Delta x \to 0} \left(\frac{\Delta y}{\Delta x} - \left(\frac{dy}{dx}\right)_{x=x_1} \right) = \lim_{\Delta x \to 0} k \qquad (2)$$

If we now use the fact that the limit of a sum [difference] is the sum [difference] of the limits (and as an important aside, observe here that we are more interested in this property of a limit than we are in finding an appropriate delta for a given epsilon; in other words, we often use epsilon-delta techniques to prove various theorems concerning limits but once the theorems are established we usually use them directly without reference to how they were derived), we see that (2) becomes:

$$\lim_{\Delta x\to 0} \frac{\Delta y}{\Delta x} - \lim_{\Delta x\to 0} \left[\left(\frac{dy}{dx}\right)_{x=x_1} \right] = \lim_{\Delta x\to 0} k \qquad (3)$$

Now since $\left(\frac{dy}{dx}\right)_{x=x_1}$ is a constant with respect to Δx, it follows that the limit of $\left(\frac{dy}{dx}\right)_{x=x_1}$ is precisely $\left(\frac{dy}{dx}\right)_{x=x_1}$. Also by definition of derivative,

$$\lim_{\Delta x\to 0} \left(\frac{\Delta y}{\Delta x}\right)\Bigg|_{x=x_1} = \left(\frac{dy}{dx}\right)_{x=x_1}$$

Putting this into (3) we obtain:

$$\left(\frac{dy}{dx}\right)_{x=x_1} - \left(\frac{dy}{dx}\right)_{x=x_1} = \lim_{\Delta x\to 0} k$$

$\therefore \lim_{\Delta x\to 0} k = 0$ and so k is an infinitesimal, as claimed.

With this information we can return to (1) and conclude:

$$\boxed{\Delta y = \left(\frac{dy}{dx}\right)_{x=x_1} \Delta x + k\Delta x, \text{where } \lim_{\Delta x\to 0} k = 0} \qquad (4)$$

If we now recall that $\left(\frac{dy}{dx}\right)_{x=x_1} \Delta x = \Delta y_{\tan}$, (4) becomes:

$$\boxed{\Delta y = \Delta y_{\tan} + k\Delta x, \text{ where } \lim_{\Delta x \to 0} k = 0} \qquad (5)$$

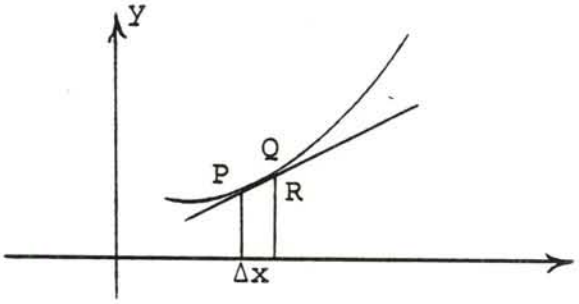
At first glance (5) doesn't seem to do more than confirm what we already sensed more intuitively in the previous section - that $\Delta y_{\tan}$ is a good approximation for Δy, if Δx is sufficiently small. However, (5) tells us much more than that. If we read it correctly, it tells us that the error in the approximation ($k\Delta x$) goes to zero "faster" than Δx goes to zero. That is, for small values of Δx, both Δx and k are small; hence, $k\Delta x$ is even smaller.

It is the fact that $\lim_{\Delta x \to 0} k = 0$ that is essential to the study of calculus. That is, even if $\lim_{\Delta x \to 0} k \neq 0$, we would have that $\lim_{\Delta x \to 0} (k\Delta x) = 0$ as long as $\lim_{\Delta x \to 0} k$ was finite. In other words, regardless of whether k was an infinitesimal, it would be true that Δy and $\Delta y_{\tan}$ were nearly equal for small values of Δx.

The key point lies in the idea that it is not enough in most applications of the limit process to know that $k\Delta x$ approaches zero, for in most cases we will be dividing $k\Delta x$ by another infinitesimal and we must be sure that the quotient, not just the numerator, is also an infinitesimal.

We shall begin to illustrate this idea more in the next section. For now we conclude this section with two remarks: (1) The notion of $\left(\frac{dy}{dx}\right)_{x=x_1} \Delta x$ corresponds to the pre-calculus notion of distance equals rate times time. Such a notion utilizes constant speed and clearly $\left(\frac{dy}{dx}\right)_{x=x_1}$ is constant. In this context $k\Delta x$ is the "correction" factor to adjust for the fact that speed was not constant.

(2) The Geometric significance of $\lim_{\Delta x \to 0} k = 0$ is that Δy and Δy_{tan} seem to coincide long before Δx "looks like" zero. That is:



Notice that the size of Δx is quite noticeable, but QR, which denotes $\Delta y - \Delta y_{tan}$, seems like almost one "thick" point. (Again, QR is not k rather $k = QR/\Delta x$.)

C. The Chain Rule

In this section we shall apply the idea of differentials to arrive at the very important chain rule. On a non-calculus level the chain rule is particularly easy to explain. Namely if a first variable (y) can be expressed in terms of a second (x), and the second can be expressed in terms of a third (t), then the first can be expressed in terms of the third. Certainly there is nothing strange about this result as it involves no more than an application of the idea of substitution.

The chain rule takes on more meaning in terms of calculus. In this case, we add the condition that the first variable (y) is a <u>differentiable</u> function of the second (x), which means that $\frac{dy}{dx}$ exists. In a similar way, we also assume that the second is a differentiable function of the third (t) - that is, we also assume that $\frac{dx}{dt}$ exists. Under these conditions it not only follows that y is a function of t (since that much we saw was true regardless of differentiability) but that y is a <u>differentiable</u> function of t. Of even more importance, $\frac{dy}{dt}$ is very strongly related to $\frac{dy}{dx}$ and $\frac{dx}{dt}$. Namely:

$$\frac{dy}{dt} = \left(\frac{dy}{dx}\right)\left(\frac{dx}{dt}\right) \tag{1}$$

Equation (1) may seem "self-evident" since we merely seemed to cancel a common factor from numerator and denominator. It is important to understand, however, that at least for now $(\frac{dy}{dx})$ is not a quotient of two numbers but merely <u>one</u> symbol. That is, $\frac{dy}{dx}$ is defined as a number obtained by taking a limit.

$$\frac{dy}{dx} = \lim_{\Delta x \to 0}\left[\frac{\Delta y}{\Delta x}\right] \quad \text{(and, recall, this is } \underline{\text{not}} \text{ 0/0)}$$

Of course, had (1) turned out not to be true, it is quite likely that we would not have invented the notation $\frac{dy}{dx}$ (we would have stuck with f') since this would tend to make us misinterpret the symbol with properties possessed by common fractions (such as cancellation).

How, then, do we establish the proof of (1)? To make the result seem correct observe that $\frac{dy}{dt} = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t}$. Moreover, we can write:

$$\frac{\Delta y}{\Delta t} = \left(\frac{\Delta y}{\Delta x}\right)\left(\frac{\Delta x}{\Delta t}\right) \tag{2}$$

In (2) Δx, Δy, and Δt are <u>numbers</u> which can be cancelled. Thus, it appears that:

$$\text{(a)} \quad \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \lim_{\Delta t \to 0} \left[\left(\frac{\Delta y}{\Delta x}\right)\right] \left[\left(\frac{\Delta x}{\Delta t}\right)\right] \quad \text{or:} \quad \frac{dy}{dt} = \left(\frac{dy}{dx}\right)\left(\frac{dx}{dt}\right)$$

While this makes (1) seem plausible, we should observe that our derivation was not quite "legal." For one thing, it is possible that for a given Δt, Δx might be zero. If this happens then the factor $\frac{\Delta y}{\Delta x}$ is undefined since we are never allowed to divide by zero.*

While our intuition might support the truth of (1), we have experienced enough trouble in trying to use our intuition for such expressions as 0/0 that probably we should require a more rigorous form of argument before accepting the truth of (1). Before doing this, however, perhaps a few remarks concerning notation are in order. When we talk about $f'(x)$ we are talking about a function of x. That is, we may think of the f'-machine and see that for different x's there can be different outputs. Thus to determine $f'(x)$ as a specific number, we must think of a specific value for x. This is why we often write $f'(x_1)$ which is an abbreviation for $[f'(x)]_{x=x_1}$. At any rate, when we talk about y being a function of x, we think of a specific value of y for a specific value of x. In this vein, notice in the last section we wrote things like:

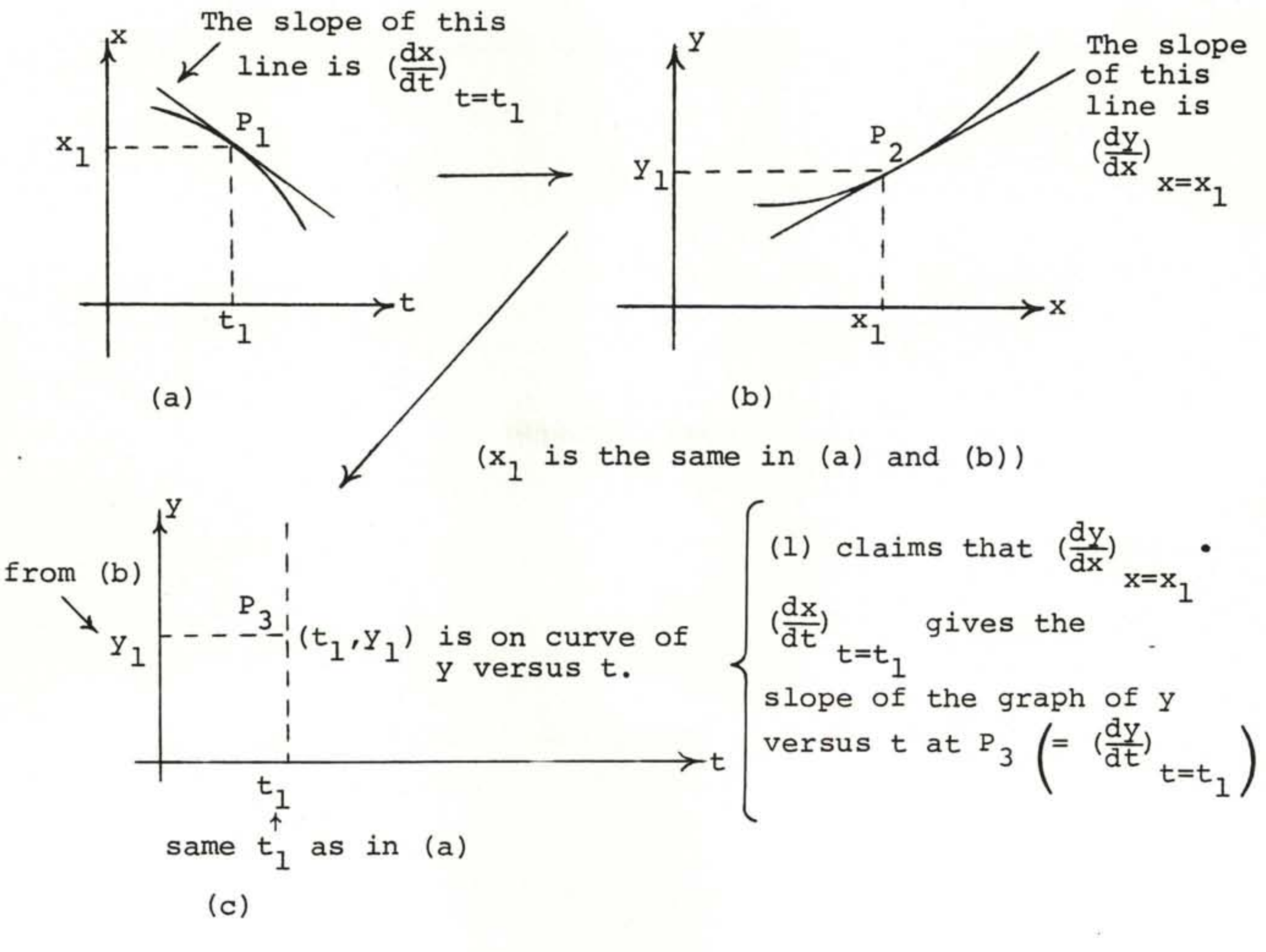
$$y = \left(\frac{dy}{dx}\right)_{x=x_1} \Delta x + k\Delta x$$

to indicate that we were referring to the tangent to the curve at a specific point.

Perhaps the following geometric interpretation will help illustrate our point. Suppose we have two graphs, one of which plots y versus x and the other of which plots x versus t.

*Recall that in our previous discussion we talked about 0/0. In the event that we have b/0 where $b \neq 0$, we find that there can be no answer. For example, 3/0 would mean the number which when multiplied by 0 equaled 3. There is no such number since any number times 0 is 0. Thus 0/0 is indeterminate while 3/0 is undefined. In either event we exclude division by 0. In terms of dividing by "small" numbers, we may think of 3/0 as meaning the limit of 3/x as x approaches 0. In this event 3/x increases without bound. Perhaps it is for this reason that one often finds the notation $3/0 = \infty$.

We can then match a value of t with a value of y as follows: Given $t = t_1$ we look at the graph of x versus t to find the corresponding value of x which we label x_1. We then go to the graph of y versus x and find the value of y which corresponds to x_1. We call this y_1. (By the way, notice how we make use of the concept of single-valuedness. We tacitly assume that a value of t yields a unique value of y, etc.) We next plot y as a function of t by locating the point (t_1, y_1) in the y-t plane. Thus:



The slope of this line is $\left(\frac{dx}{dt}\right)_{t=t_1}$
x
x_1
P_1
t
t_1
The slope of this line is $\left(\frac{dy}{dx}\right)_{x=x_1}$
y
y_1
P_2
x
x_1
from (b)
y
y_1
P_3
(t_1, y_1) is on curve of y versus t.
t
t_1
same t_1 as in (a)

(a)

(b)

(x_1 is the same in (a) and (b))

(c)

(1) claims that $\left(\frac{dy}{dx}\right)_{x=x_1} \cdot \left(\frac{dx}{dt}\right)_{t=t_1}$ gives the slope of the graph of y versus t at P_3 $\left(= \left(\frac{dy}{dt}\right)_{t=t_1}\right)$

Thus when we talk about, say, $\frac{dy}{dx}$ we mean $\left(\frac{dy}{dx}\right)_{x=x_1}$. At any rate, then, we may think of the problem described in the chain rule as follows.

We know that $\left(\frac{dy}{dx}\right)_{x=x_1}$ exists as does $\left(\frac{dx}{dt}\right)_{t=t_1}$. We wish to prove that:

$$\left(\frac{dy}{dt}\right)_{t=t_1} \text{ exists and equals } \left(\frac{dy}{dx}\right)_{x=x_1} \left(\frac{dx}{dt}\right)_{t=t_1}$$

To this end we invoke the result of the last section to obtain:

$$\boxed{\Delta y = \left(\frac{dy}{dx}\right)_{x=x_1} \Delta x + k\Delta x, \quad \lim_{\Delta x\to 0} k = 0} \qquad (3)$$

From (3), we obtain, if $\Delta t \neq 0$,

$$\boxed{\frac{\Delta y}{\Delta t} = \left(\frac{dy}{dx}\right)_{x=x_1} \frac{\Delta x}{\Delta t} + k\frac{\Delta x}{\Delta t}; \quad \lim_{\Delta x\to 0} k = 0} \qquad (4)$$

Recalling that $\frac{dy}{dt}$ means $\lim_{\Delta t\to 0} \frac{\Delta y}{\Delta t}$ and using our limit theorems, we see from (4) that:

$$\frac{dy}{dt} = \lim_{\Delta t\to 0} \frac{\Delta y}{\Delta t} = \lim_{\Delta t\to 0} \left[\left(\frac{dy}{dx}\right)_{x=x_1} \frac{\Delta x}{\Delta t} + k\frac{\Delta x}{\Delta t}\right]$$

$$= \lim_{\Delta t\to 0} \left[\left(\frac{dy}{dx}\right)_{x=x_1} \frac{\Delta x}{\Delta t}\right] + \lim_{\Delta t\to 0} \left[k\frac{\Delta x}{\Delta t}\right]$$

$$= \left\{\left[\lim_{\Delta t\to 0} \left(\frac{dy}{dx}\right)_{x=x_1}\right] \left[\lim_{\Delta t\to 0} \frac{\Delta x}{\Delta t}\right]\right\} + \left\{\left[\lim_{\Delta t\to 0} k\right] \left[\lim_{\Delta t\to 0} \frac{\Delta x}{\Delta t}\right]\right\} \qquad (5)$$

Since $\left(\frac{dy}{dx}\right)_{x=x_1}$ is a <u>fixed</u> number, it remains unchanged when we let $\Delta t \to 0$. Also the definition of x being a differentiable function of t means that $\lim\limits_{\Delta t\to 0} \frac{\Delta x}{\Delta t} \equiv \frac{dx}{dt}$. Thus (5) becomes:

$$\frac{dy}{dt} = \left(\frac{dy}{dx}\right)_{x=x_1} \frac{dx}{dt} + \left[\lim_{\Delta t\to 0} k\right] \frac{dx}{dt}$$

or:

$$\boxed{\left(\frac{dy}{dt}\right)_{t=t_1} = \left(\frac{dy}{dx}\right)_{x=x_1} \left(\frac{dx}{dt}\right)_{t=t_1} + \left[\lim_{\Delta t\to 0} k\right] \left(\frac{dx}{dt}\right)_{t=t_1}} \quad (6)$$

If we concentrate on $\lim\limits_{\Delta t\to 0} k$ we see that it is zero. For since x is a differentiable function of t, $\Delta x \to 0$ as $\Delta t \to 0$. Therefore $\lim\limits_{\Delta t\to 0} k = \lim\limits_{\Delta x\to 0} k$* and this, by (3) is 0.

*For a more rigorous derivation of this result we must show that given $\varepsilon > 0$ there exists $\delta > 0$ such that $0 < |\Delta t| < \delta \to |k| < \varepsilon$. Since we are given that $\lim\limits_{\Delta x\to 0} k = 0$ we know that for the given ε we can find $\delta_1 > 0$ such that:

$$0 < |\Delta x| < \delta_1 \to |k| < \varepsilon \quad \text{(i)}$$

But $\lim\limits_{\Delta t\to 0} \Delta x = 0$ (since $\Delta x = \frac{\Delta x}{\Delta t}\Delta t$, hence $\lim\limits_{\Delta t\to 0} \Delta x = \lim\limits_{\Delta t\to 0} \frac{\Delta x}{\Delta t} \lim\limits_{\Delta t\to 0} \Delta t$

$\therefore \lim\limits_{\Delta t\to 0} \Delta x = \frac{dx}{dt}\, 0 = 0$) [Notice here that we used $\frac{dx}{dt} = \lim\limits_{\Delta t\to 0} \frac{\Delta x}{\Delta t}$. If x were not a differentiable function of t, $\lim\limits_{\Delta t\to 0} \frac{\Delta x}{\Delta t}$ might not exist.

$\therefore$ Given $\delta_1 > 0$, we can find $\delta > 0$ such that

$$0 < |\Delta t| < \delta \to |\Delta x| < \delta_1 \quad \text{(ii)}$$

Combining (i) and (ii), we have:

$$0 < |\Delta t| < \delta \to |k| < \varepsilon$$

which establishes the desired result: $\lim\limits_{\Delta t\to 0} k = 0$.

Hence (6) becomes

$$\left(\frac{dy}{dt}\right)_{t=t_1} = \left(\frac{dy}{dx}\right)_{x=x_1} \left(\frac{dx}{dt}\right)_{t=t_1} + 0 \overset{= \text{ some number}}{\left(\frac{dx}{dt}\right)_{t=t_1}}$$

But since any number times 0 is 0, we have:

$$\left(\frac{dy}{dt}\right)_{t=t_1} = \left(\frac{dy}{dx}\right)_{x=x_1} \left(\frac{dx}{dt}\right)_{t=t_1}$$

which confirms our intuitive belief in (1).

The crucial point to observe is that $k \frac{\Delta x}{\Delta t}$ approached zero <u>not</u> because Δx became small (i.e. don't replace Δx by 0 and say the result is 0, for when $\Delta x = 0$ so is Δt and we arrive at 0/0 and in fact $\lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt}$) but because k becomes small as well. In other words if $\lim_{\Delta t \to 0} k \neq 0$ then $k \frac{\Delta x}{\Delta t}$ need not approach zero as Δt approaches zero.

This is precisely why it was crucial that $k\Delta x$ be a <u>higher-order</u> infinitesimal.

We conclude this section on the chain rule with the observation that the chain rule is directly connected with the composition of functions. That is, when we talked about $y = f(x)$ and $x = g(t)$, we were really saying that $y = f(g(t))$. In other words, if we let $h = f \circ g$ then $y = h(t)$ and what the chain rule says is that we may compute $h'(t_o)$ by taking the product of $f'(x_o)$ and $g'(t_o)$ where $x_o = g(t_o)$.

The point is that in many applications of the chain rule, we are not given the composition of two functions explicitly. For example, consider the problem of finding $\frac{dy}{dt}$ if we are given that $y = (t^3 + 1)^2$. The chain rule comes into play here if we make the substitution $x = t^3 + 1$. We then obtain that $y = x^2$ and $x = t^3 + 1$, whereupon the chain rule seems to present itself. More explicitly in terms of composition of functions, we may define f by $f(u) = u^2$ and g by $g(v) = v^3 + 1$. Then $y = (t^3 + 1)^2$ can be written as $y = f(g(t))$. In any event the chain rule now tells us that

$$\frac{dy}{dt} = \left(\frac{dy}{dx}\right)\left(\frac{dx}{dt}\right) = (2x)\ (3t^2) = 6xt^2{}^{*} \tag{7}$$

As a check notice that we could, in this case, solve for y explicitly as a function of t, and obtain:

$$y = (t^3 + 1)^2 = t^6 + 2t^3 + 1$$

whereupon:

$$\frac{dy}{dt} = 6t^5 + 6t^2 \tag{8}$$

The equivalence of (7) and (8) follows from the fact that:

$$6t^5 + 6t^2 + 6t^2(t^3 + 1) \text{ and } x = t^3 + 1,$$

therefore

$$6t^5 + 6t^2 = 6t^2x = 6xt^2.$$

What was important to note was that when we want to differentiate $(t^3 + 1)^2$ with respect to t we must do more than bring down the exponent and replace it by one less. That is, the recipe that $\frac{d(u^2)}{du} = 2u$ implies that we are differentiating with respect to the SAME variable that is

*In this case we explicitly can express x in terms of t. That is, $x = t^3 + 1$. Hence we could have written $\frac{dy}{dt} = 6(t^2 + 1)t^2$. However in some cases x is at best an implicit function of t in which case our explicit expression for $\frac{dy}{dt}$ will contain both x's and t's.

being raised to the power. Using brackets to represent the variable, we are saying that $\frac{d([\ ]^2)}{d([\ ])} = 2[\ ]$. In particular, replacing the brackets by $t^3 + 1$, we see that:

$$\frac{d(t^3 + 1)^2}{d(t^3 + 1)} = 2(t^3 + 1)$$

What the chain rule tells us is that:

$$\frac{d(t^3 + 1)^2}{dt} = \left[\frac{d(t^3 + 1)^2}{d(t^3 + 1)}\right]\left[\frac{d(t^3 + 1)}{dt}\right] =$$

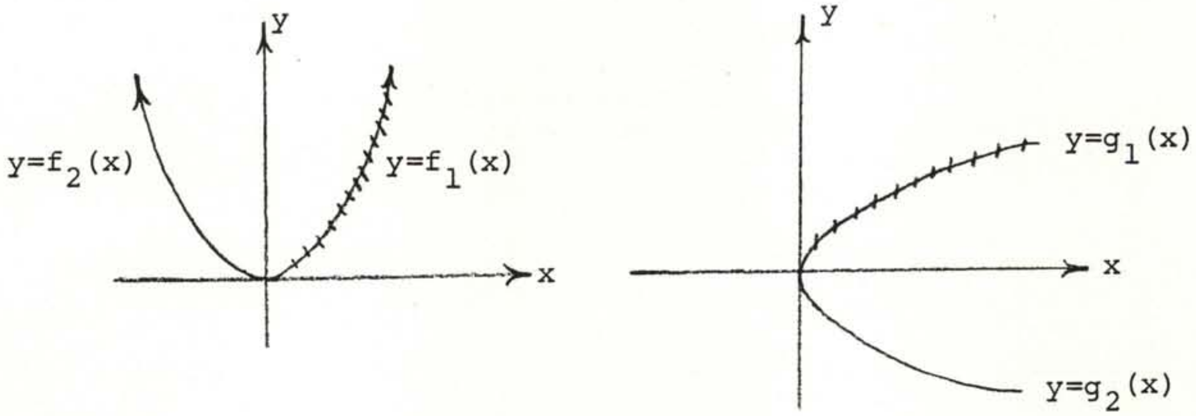
$$\left[2(t^3 + 1)\right]\ \left[3t^2\right] = 6t^2(t^3 + 1)$$

D. A Note About Inverse Functions

In the statement of the chain rule, when we say that if y is a differentiable function of x and x is a differentiable function of t then y is a differentiable function of t and $\frac{dy}{dt} = (\frac{dy}{dx})\ (\frac{dx}{dt})$, nothing excludes the possibility that t might equal y. In this event, the chain rule says that if y is a differentiable function of x and if x is also a differentiable function of y then $\frac{dy}{dy} = (\frac{dy}{dx})\ (\frac{dx}{dy})$. Since $\frac{dy}{dy} = 1$, it follows that $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are reciprocals of one another - just as the fractional notation seems to indicate.

To be sure, it is not always true that if y is a function of x then x is a function of y. One major reason for this problem is our insistence that functions be single-valued. For example, if we let $y = f(x) = x^2$ then y is certainly a differentiable function of x. In fact, $\frac{dy}{dx} = 2x$. On the other hand, if we "invert" the roles of x and y, we obtain $x = \pm\sqrt{y}$. In other words, if we define f by $f(x) = x^2$ and the domain of f is the set of all real numbers, then f is single-valued but it is not 1-1 since both x and -x have the same image with respect to f. In this case it is not difficult to see that we can "partition" f into the union of two other single-valued functions, say f_1 and f_2, which "look like" f except that the domains are different. Namely, we define f_1 by $f_1(x) = x^2$ where dom f_1 is the set of non-negative real numbers; and we define f_2 by $f_2(x) = x^2$ where the domain of f_2 is the set of negative numbers.

In terms of a picture:



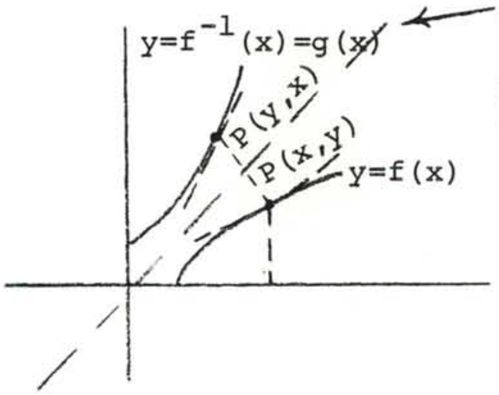
In other words, g_1 is the inverse of f_1 and g_2 is the inverse of f_2, where g_1 and g_2 are defined by:

$$g_1(y) = \sqrt{y} \text{ and } g_2(y) = -\sqrt{y}$$

and the domain of both g_1 and g_2 is the set of non-negative real numbers.

Of course, in this particular problem we were a bit fortunate in that we could solve explicitly for x in terms of y. There are many times when we cannot do this, or if we can it is not very convenient. For example, if we were given that $y = x^7 + 9x^3 + 6x + 1$ it would take a great deal of arithmetic if we desired to specify the inverse relation explicitly.

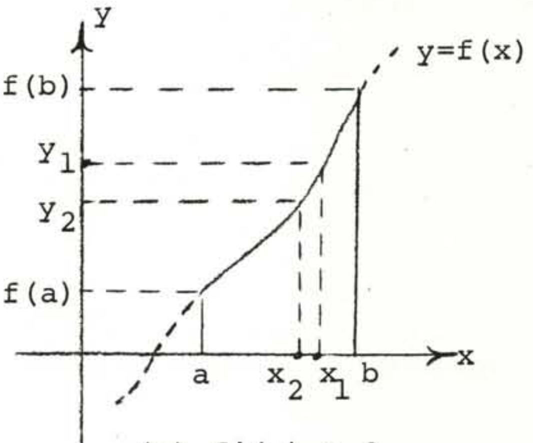
At any rate, let us suppose we have avoided this problem by having $y = f(x)$ where f is 1-1 in its domain of definition. The next problem is that we have not yet shown that if f is differentiable so also is f^{-1}. From an intuitive point of view, this is not too difficult a result to accept. If we recall that the graphs of $y = f(x)$ and $y = f^{-1}(x)$ are symmetric with respect to the line $y = x$ then it is easy to surmise that if one curve is smooth so is its reflected image. In terms of this idea, it is probably easy to see also why $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are reciprocals. That is:



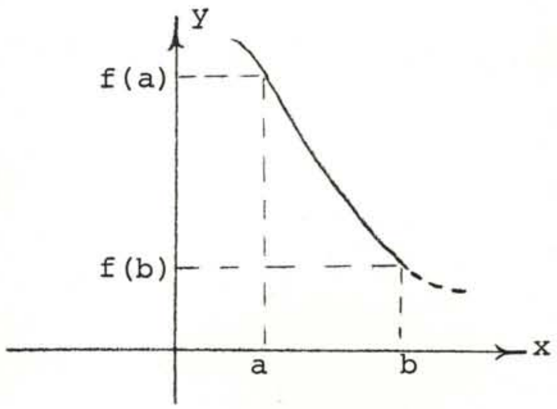
Curves are mirror images with respect to $y = x$. Hence $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are obtained merely by reversing the roles of x and y.

There are two problems with using the picture. In the first place, we would like to feel that analytic concepts can be proved analytically. Thus, while a picture may be reassuring or even helpful in getting us to visualize what is happening, we do not wish to feel beholden to the picture. Of even more importance, especially when we deal later with functions of several variables, it will be impossible to draw pictures and we will have only the analysis upon which to rely - so we might just as well get used to it now!

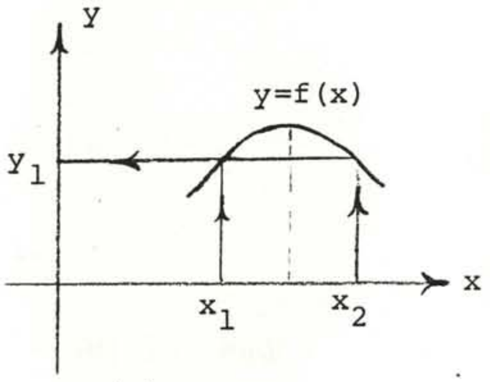
To show what we mean by our last remarks, let us assume that $y = f(x)$ where dom $f = [a,b]$ and f is both differentiable and 1-1 in this interval. Our picture then indicates:



(a) $f'(x) \geqslant 0$



(b) $f'(x) \leqslant 0$



(c)

If $f'(x)$ changes sign then f is not 1-1. For example, $x_1 \neq x_2$ but $f(x_1) = f(x_2)$.

For the sake of our demonstration we shall assume that the case $f'(x) \geqslant 0$ prevails. (The first thing we would have to do is to show that we can substantiate analytically the result that if f is differentiable in the interval [a,b] and f^{-1} exists, then $f'(x)$ is never equal to zero in this interval. This is done in virtually every calculus textbook and the analytic proof is particularly easy to follow if we keep the picture in mind.)

We would then turn to the definition of a derivative and if g denotes f^{-1}, we would write for any y_1 in $[f(a),f(b)]$

$$g'(y_1) = \lim_{y_2 \to y_1}\left[\frac{g(y_2) - g(y_1)}{y_2 - y_1}\right] \tag{1}$$

Since f and g are inverse functions, we have that for a given y in $[f(a),f(b)]$ there exist one and only one x in $[a,b]$ for which $f(x) = y$. Thus, there exist unique numbers x_1 and x_2 in $[a,b]$ for which $f(x_1) = y_1$ and $f(x_2) = y_2$.

Putting this into (1) we obtain:

$$g'(y_1) = \lim_{y_2 \to y_1}\left[\frac{x_2 - x_1}{f(x_2) - f(x_1)}\right] \tag{2}$$

and since $x_2 \to x_1$ if and only if $y_2 \to y_1$, (2) can be written as:

$$g'(y_1) = \lim_{x_2 \to x_1}\left[\frac{x_2 - x_1}{f(x_2) - f(x_1)}\right] \tag{3}$$

and since we know that $f'(x_1)$ exists, (3) together with the definition of f' tells us that:

$$g'(y_1) = \frac{1}{f'(x_1)} \tag{4}$$

Notice that (4) says in functional notation the same thing as our conjecture that dy/dx and dx/dy are reciprocals. The only problem with using the chain rule by itself was that all the chain rule told us was that if y was a differentiable function of x and x was a differentiable function of y then dy/dx and dx/dy were reciprocals. It did not tell us how we could tell that the inverse function existed nor did it tell us that even if f and f^{-1} existed the differentiability of f implied the differentiability of f^{-1}.

To put all our theory into action, let us find dy/dx if we are given that $x = y^5 - y^3$. What we know from "way back" is that $dx/dy = 5y^4 - 3y^2$. What we must now assume is that there exists a 1-1 function say $y = f(x)$ on a suitable domain whose inverse is given by $f^{-1}(y) = y^5 - y^3$. (The conditions which guarantee the existence of such a function require some knowledge of calculus of several variables and that is why the concept of implicit differentiation must be taken on faith until later in the course. That is, a major problem is in showing the existence of single valued branches of multi-valued functions.) In any event once we assume that the desired f exists, we may invoke the result that dy/dx and dx/dy are reciprocals and that, therefore, dy/dx is equal to $1/(5y^4 - 3y^2)$. In the language of functions, we are saying that if $x = g(y) = y^5 - y^3$ then there exists a domain [a,b] on which $y = f(x)$ and f is 1-1. Moreover in this case the inverse of f, f^{-1}, is a suitable single-valued branch of g, say g_1 where $g_1(y) = y^5 - y^3$ and dom $g_1 = [f(a), f(b)]$ (or [f(b), f(a)] if f'(x) is always negative since then the graph is always falling). Then for any x_1 in [a,b], $f'(x_1)$ exists and:

$$f'(x_1) = \frac{1}{g'(y_1)} = 1/(5y^4 - 3y^2)$$

where $y_1 = f(x_1)$ or equivalently, $x_1 = g(y_1)$.

E. Differentials

As we mentioned earlier, such notations as $\frac{dy}{dx}$ could have been most unfortunate had results like $\frac{dy}{dx} = (\frac{dy}{dt})\ (\frac{dt}{dx})$ not been true. That is, unless the derivative had properties similar to those of fractions we would have stuck with such notation of f'(x) rather than to introduce a misleading notation.

The fact that derivatives have such nice "fraction-like" properties tempts us to try to define dy and dx as separate entities in such a way that when we divide dy by dx we obtain the same result as if we had found the derivative of y with respect to x. More symbolically, we would like the expressions dy, dx and $(\frac{dy}{dx})$ to be related by:

$$dy = (\frac{dy}{dx})\,dx \qquad (1)$$

Recall that we have already defined Δy_{tan} by:

$$\Delta y_{tan} = (\frac{dy}{dx})\,\Delta x \qquad (2)$$

If we now compare (1) and (2) it becomes clear that one way of accomplishing our goal is to let dy mean Δy_{tan} and dx merely mean Δx. Then (1) and (2) are synonymous. Among other things this would mean that we could interpret $\frac{dy}{dx}$ either as a derivative or as a quotient and get the same answer in either case. This, in turn, means that in any situation in which $\frac{dy}{dx}$ is involved we can treat it as a derivative or as a quotient depending on which interpretation better serves our needs.

From the point of view of approximations, dy is nothing more than another name for Δy_{tan}. Consequently, other than for a new notation, we gain nothing by replacing Δy_{tan} by dy.

From an analytical point of view, however, the fractional notation is very important in that it is more suggestive than the usual functional notation. For example, in terms of an illustration from the last section, it seems easier to interpret that dy/dx and dx/dy are reciprocals than it is to see that if $g = f^{-1}$ then $g'(y_1) = 1/(f'(x_1)$ where $y_1 = f(x_1)$.

Without worrying about the semantic difference between dy and Δy_{tan}, let us note that the key computational device is that the error in replacing Δy by dy is a higher order infinitesimal and as a result in any limit problem we get the same answer using dy and dx as we would had we used the derivative $(\frac{dy}{dx})$.

With respect to this course, our main exploitation of this notation is that we will often elect to write such expressions as $\frac{dy}{dx} = x^2$ in the form $dy = x^2dx$. Notice that this transformation may be viewed either as an algebraic cross-multiplication or as an application of $\Delta y_{tan} = (\frac{dy}{dx})\ \Delta x$.

However, we shall say more about this later in the course. For now we will leave things as they stand with additional reinforcement coming from the exercises.

Chapter VII
THE TRIGONOMETRIC FUNCTIONS

A. Introduction

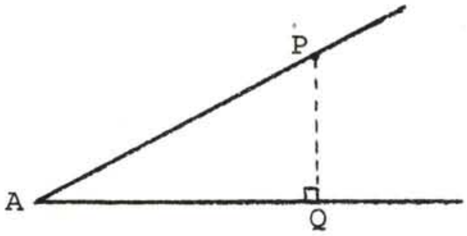
Judging from the usual way in which fractions are taught there is ample reason to wonder whether fractions would have been invented had not pies been invented! On a much more subtle level it seems that the traditional teaching of trigonometry leads one to believe that trigonometry would not have been invented had not triangles been invented. Yet the fact is that the concept of the trigonometric functions is far more important than its role in numerical geometry.

In this chapter it shall be our purpose to liberate the full flavor of trigonometry and, at the same time, to present a brief revisit with classical trigonometry.

Let us recall that, historically, the subject called trigonometry was an outgrowth of geometry. In fact, the very name "geometry" suggests the course which was traditionally called trigonometry. That is, "geo" is derived from the Greek word for "earth" and "metry" is a derivative of "measure." Thus, geometry was the science of measuring the earth. It seems, to put it in other words, that what geometry did qualitatively, trigonometry did quantitively. For example, in plane geometry one demonstrates that a triangle is completely determined up to its position in space once we know the measure of two sides and the included angle (this is precisely what is meant when one says that two triangles are congruent if two sides and the included angle of one are respectively equal to sides and the included angle of the other). On the other hand, in trigonometry, by such devices as the Law of

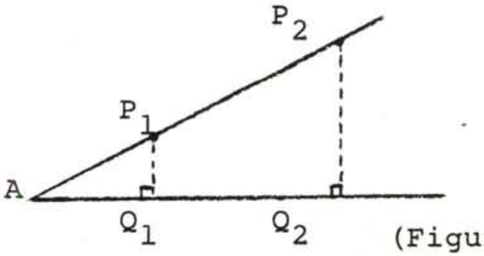
Cosines, one determines the measurements of the remaining parts of a triangle once the measurements of two sides and the included angle are known. It is in this sense that we mean that classical trigonometry is numerical geometry.

We may also assume that the traditional study of the trigonometric functions was based on the knowledge of properties of right triangles. Among other things, trigonometry was interested in showing how one could reduce the study of angles to a study of lengths. Specifically, given any acute angle, we could "imbed" it in a right triangle by dropping a perpendicular from one side of the angle to the other side. Thus:



(Figure 1)

Of course, the above construction is subjective since two different observers might choose different points from which to drop the perpendicular. Again, pictorially:



$$\begin{cases} \overline{P_1Q_1} \neq \overline{P_2Q_2},\ \overline{AQ_1} \neq \overline{AQ_2} \\ \text{But } \dfrac{\overline{P_1Q_1}}{\overline{AQ_1}} = \dfrac{\overline{P_2Q_2}}{\overline{AQ_2}} \qquad \text{etc.} \end{cases}$$

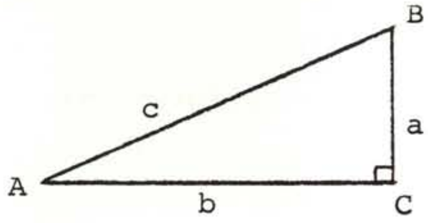
(Figure 2)

From their knowledge of plane geometry, however, the ancient Greeks knew that while the lengths of the sides of the right triangle depended on the choice of the point P (see Figure 2), the ratio between the lengths of any pair of sides did not. This, of course, was a consequence of similar triangles. In any event, then, by studying the ratios between the lengths of each pair of sides in the "imbedding" triangle, the study of trigonometry was born.

We should also point out that from an informal point of view, it is fair to say that trigonometry was born much earlier. It was born as soon as man realized that he could, by the principles of similar triangles, draw things to scale. In other words, given two sides and the included angle, he did not have to wait for the Law of Cosines to be invented in order to measure the remaining parts. All he had to do was use a ruler and protractor to draw the triangle to scale and then directly measure the other parts.

Be this as it may, it is correct to say that initially the trigonometric functions were functions whose domain was the set of acute angles and whose range consisted of real numbers. More specifically, given an acute angle, A, the angle was imbedded in a right triangle (and this is why it was crucial for the angle to be acute, otherwise it couldn't be imbedded in a right triangle), say, ACB; whereupon sin A, cos A, tan A, cot A, sec A, and csc A were defined by:

$$\sin A = \frac{a}{c} \qquad \cos A = \frac{b}{c} \qquad \tan A = \frac{a}{b}$$



$$\csc A = \frac{c}{a} \qquad \sec A = \frac{c}{b} \qquad \cot A = \frac{b}{a}$$

(Figure 3)

To be sure, our knowledge of arithmetic and geometry showed us that our definitions were well-defined (that is, the definition of the trigonometric function of the angle depended only on the angle and not on the triangle in which it was imbedded) and that there were certain relations that existed between the various functions. For example, the Pythagorean theorem led to such results as:

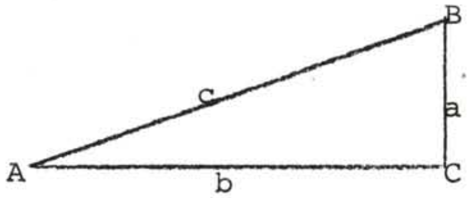
$$\sin^2 A + \cos^2 A = 1$$

$$\sec^2 A - \tan^2 A = 1$$

$$\csc^2 A - \cot^2 A = 1$$

These results do not have to be memorized. Each is an almost immediate consequence of the fact that $a^2 + b^2 = c^2$. The three relations given above can be obtained from this equation by dividing both sides of it by c^2, b^2, or a^2 respectively. Thus:

$$\left.\begin{array}{ll} a^2+b^2=c^2 \rightarrow (1) & (\frac{a}{c})^2+(\frac{b}{c})^2=1 \text{ or } \sin^2+\cos^2 A=1 \\ \qquad\text{or } (2) & (\frac{a}{b})^2+1=(\frac{c}{b})^2 \text{ or } \tan^2 A+1=\sec^2 A \\ \qquad\text{or } (3) & 1+(\frac{b}{a})^2=(\frac{c}{a})^2 \text{ or } 1+\cot^2 A=\csc^2 A \end{array}\right\} \begin{array}{l}\text{from the} \\ \text{definitions} \\ \text{in figure (3)}\end{array}$$



A second type of relation came about from the observation that in the right triangle ACB, the side that was opposite A (that is, *a*) was adjacent to B and vice versa, while the hypotenuse was independent of which acute angle we studied. Thus if we think of the sine of an angle as being the ratio of the length of the side opposite and the hypotenuse (rather than relying too heavily on the labels a, b, c) and of the cosine as being the ratio of the side adjacent and the hypotenuse, we see that:
$\sin A = \cos(90° - A)$, $\sec A = \csc(90° - A)$, $\tan A = \cot(90° - A)$.
More specifically,

$$\sin A = \frac{\text{length of side opposite A}}{\text{length of hypotenuse}} = \frac{\text{length of side adjacent to B}}{\text{length of hypotenuse}}$$

$$= \cos B = \cos(90° - A), \text{ since } A+B=90°$$

In other words, if f denotes any trigonometric function, $f(A) = \text{co-}f\ (90° - A)$.

Rather than belabor this part of review in more detail, let us also note that other nice relations come directly from the role of fractions. For example, referring to the definitions given in Figure (3), we have sin A = a/c while cos A = b/c. Hence sin A ÷ cos A = a/c ÷ b/c = a/b and this is precisely the definition of tan A. Thus, we have shown:

$$\frac{\sin A}{\cos A} = \tan A$$

In a similar way we can show that $\sec A = \frac{1}{\cos A}$, $\csc A = \frac{1}{\sin A}$ and $\cot A = \frac{1}{\tan A}$.

Even the concept of limit was, at least subconsciously, introduced into this study when one wished to talk about such things as sin 0° or cos 0° or sin 90°, etc. That is, neither 0° nor 90° can be viewed as acute angles in a right triangle, yet we may think of studying what happens to sin A as A is allowed to take on values *arbitrarily close* to 0° (or 90°) but never equal 0° (or 90°) - since then the angle A could not be imbedded in a right triangle. Thus, in terms of the language of our present course, one defined $\sin 0° = \lim\limits_{A \to 0°^+} \sin A$, $\cos 90° = \lim\limits_{A \to 90°^-} \cos A$, etc.

In this way, one obtained the additional results: sin 0° = cos 90° = 0 (notice here that it is the *number* 0 not the angle 0°) while cos 0° = sin 90° = 1 (again this is the number 1 not the angle 1°). Pictorially:

as A→0, a→0 and c→b

$\therefore \frac{a}{c} \to \frac{0}{b} = 0$

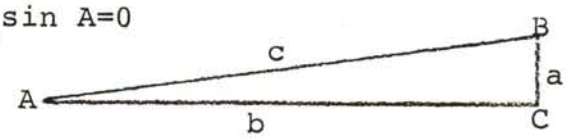
$\therefore \sin A \to 0$

$\therefore \sin 0° = \lim\limits_{A \to 0^+} \sin A = 0$

and B→90° so $\sin B = \frac{b}{c} \to \frac{b}{b} = 1$

$\therefore \sin 90° = \lim\limits_{B \to 90^-} \sin B = 1$

etc.



(Figure 4)

Once these results were obtained, it was then easy in terms of the existing relations to define the remaining trigonometric functions evaluated at 0. Namely:

$$\tan 0^\circ = \frac{\sin 0^\circ}{\cos 0^\circ} = \frac{0}{1} = 0 = \cot 90^\circ$$

$$\cot 0^\circ = \frac{\cos 0^\circ}{\sin 0^\circ} = \frac{1}{0} = \infty = \tan 90^\circ$$

$$\sec 0^\circ = \frac{1}{\cos 0^\circ} = \frac{1}{1} = 1 = \csc 90^\circ$$

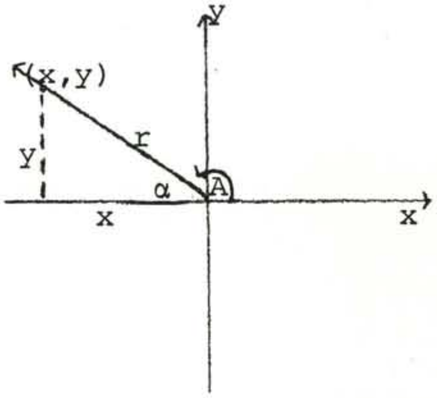
$$\csc 0^\circ = \frac{1}{\sin 0^\circ} = \frac{1}{0} = \infty = \sec 90^\circ$$

Here ∞ means $\lim\limits_{x \to 0} \frac{1}{x}$ Technically, then, cot 0° and tan 90° are not defined.

These extended definitions agreed with our other "recipes" such as: $\sin^2 A + \cos^2 A = 1$, since:

$$\sin^2 0^\circ + \cos^2 0^\circ = (0)^2 + (1)^2 = 1.$$

Still later, the results of trigonometry were extended to include all angles, not just acute angles. This idea utilized coordinate geometry. In terms of a brief review, recall that we place the angle with its vertex at the origin and its so-called initial side in the direction of the positive x-axis. We then see along what line the angle terminates, measuring counterclockwise if the angle is positive and clockwise if the angle is negative. We pick any point (x, y) on the terminal side of the angle and define the trigonometric functions of A by: sin A = y/r, cos A = x/r, tan A = y/x, etc., where r is the (positive) distance from the origin to the point. In this way we observe that our new definitions agree with the old in the case that A is a first-quadrant angle. In the other quadrants, the signs of x and y come into play.



In this Figure A terminates in the second quadrant. Hence:

$$\sin A = \frac{y}{r} = \frac{+}{+} = +$$

$$\cos A = \frac{x}{r} = \frac{-}{+} = -$$

$$\tan A = \frac{y}{x} = \frac{+}{-} = -$$

etc.

(Figure 5)

In other words, given any angle we can study the acute angle that its terminal side makes with the x-axis and then apply the acute-angle-trigonometry, subject only to the adjustments for the signs of x and y.

The important point is that in this way we can bring into play all the things we know about coordinate (analytic) geometry to extend our knowledge of trigonometry. These things are standard parts of textbooks and consequently we leave a further discussion of this idea to these textbooks for the interested reader to pursue on his own.

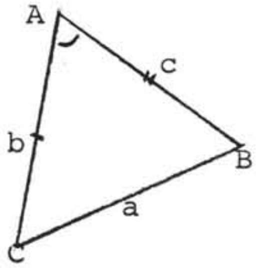
At this stage in our review, it might seem to make more sense if we found practical reasons for introducing further trigonometric relations. However, in order that we better pave the way for what follows, we prefer to leave this part of the discussion to Section C of this chapter, and for now to mention what aspects of trigonometry will be most important to our later needs. Moreover, whenever proofs are supplied we shall be traditional and use plane rather than analytic geometry. That is, we shall prove all our results in the special case that we are dealing with acute angles and leave the analytic geometry type of generalization to other angles as an exercise for those who wish to pursue it.

As a "starter," let us prove the Law of Cosines, if for no other reason than the fact that we have already talked about it.

The Law of Cosines states that if two sides (b and c) and the included angle (A) are known, then the length of the third side (a) is given by:

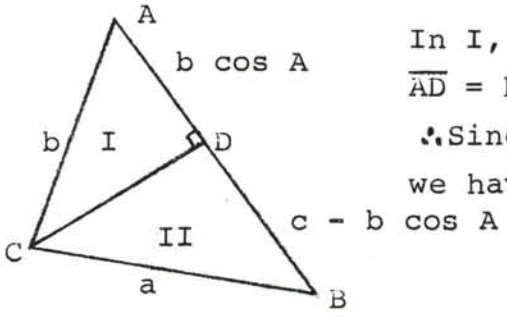
$$a^2 = b^2 + c^2 - 2bc(\cos A)$$

where we are restricting A to being an acute angle. Pictorially:



(Figure 6)

The approach is to find a way of utilizing right triangles, since we do have some specific knowledge about properties of right triangles. With this in mind, we "decompose" ΔABC into two right triangles either by drawing the altitude from B or from C (see Figure 7, where we elected to drop the perpendicular from C). We do not want to drop the perpendicular from A since A is the only "known" angle in this problem and we do not wish to "destroy" this information. In any event:



In I, since $\cos A = \frac{\overline{AD}}{b}$, we have $\overline{AD} = b \cos A$

$\therefore$Since $\overline{AB} = c$ and $\overline{BD} = \overline{AB} - \overline{AD}$, we have $\overline{BD} = c - b \cos A$

(Figure 7)

Referring still to Figure 7, we may apply the Pythagorean Theorem to triangles I and II to obtain:

$$\overline{DC}^2 = \overline{AC}^2 - \overline{AD}^2 \text{ (from I), and}$$
$$\overline{DC}^2 = \overline{BC}^2 - \overline{BD}^2 \text{ (from II)}$$

Hence:

$$\overline{AC}^2 - \overline{AD}^2 = \overline{BC}^2 - \overline{BD}^2$$

This, in turn, leads to:

$$b^2 - (b\cos A)^2 = a^2 - (c - b\cos A)^2, \text{ or:}$$
$$b^2 - b^2\cos^2 A = a^2 - c^2 + (2bc)\cos A - b^2\cos^2 A, \text{ or:}$$
$$\underline{a^2 = b^2 + c^2 - 2bc(\cos A)} \qquad (1)$$

It is equation (1) that is known as the Law of Cosines, or in some quarters, as the Extended Pythagorean Theorem - this name coming from the fact that if A = 90°, cos A = 0 and Equation (1) then reduces to the usual Pythagorean Theorem. As an interesting result concerning the idea of circular reasoning, notice it would be wrong to say that we can prove the Pythagorean Theorem from the Law of Cosines. For while the Pythagorean Theorem follows from Equation (1), our proof of Equation (1) required that we <u>already knew</u> the Pythagorean Theorem. Nonetheless, the Pythagorean Theorem is still a good check for Equation (1) (that is, if Equation (1) doesn't reduce to the Pythagorean Theorem when A = 90° there is something wrong). Moreover, if we could find a way of deriving Equation (1) without having to use the Pythagorean Theorem then we would not have circular reasoning.

Another fundamental "recipe" involves the sine of the sum of two angles. Again we will limit our discussion to the case in which A, B and A + B are acute angles. We show that:

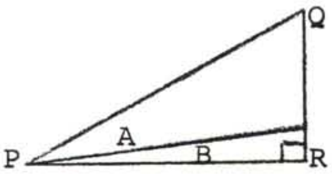
$$\sin (A + B) = \sin A \cos B + \cos A \sin B \qquad (2)$$

Equation (2) might not seem very "natural" but it turns out to be correct. While one might prefer to have had:

$$\sin(A + B) = \sin A + \sin B^*$$

this recipe would be incorrect. Among other reasons, we need only consider $A = B = 90°$ to see that this would yield the incorrect result that $\sin(90° + 90°) = \sin 90° + \sin 90°$ or $\sin 180° = 1 + 1 = 2$. Not only is it false that sin 180° equals 2, but for any angle, the sine is restricted to values not in excess of 1 since the side opposite any angle cannot exceed the length of the hypotenuse of the right triangle.

In any event we have:

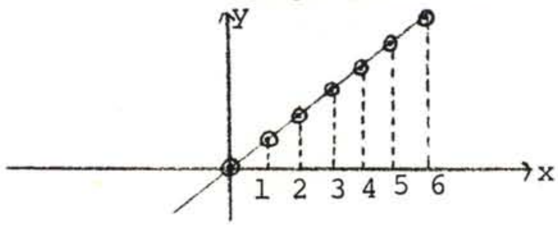


$$\sin(A + B) = \sin\sphericalangle QPR = \frac{\overline{QR}}{\overline{PQ}}$$

(Figure 8)

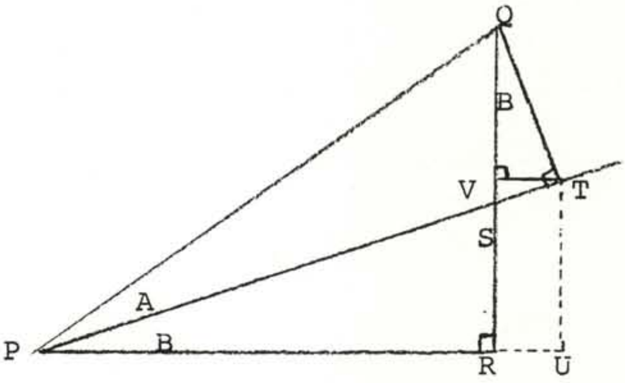
So, somehow or other, we would like to be able to form some right triangles that would have A and/or B as acute angles and so that we could still utitize PQ as a hypotenuse.

*It is indeed a very special type of function f for which $f(x + y) = f(x) + f(y)$. For if f has this property then letting x and y both equal 0, we obtain $f(0 + 0) = f(0) + f(0)$ or $f(0) = 2f(0)$, from which it follows that $f(0) = 0$. Next observe that $f(1 + 1) = f(1) + f(1)$, or $f(2) = 2f(1)$. Quite in general for any whole number n, $f(n) = f(1+...+1) = f(1) + f(1) +.... + f(1)$. In other words, we can show by induction that $f(n) = nf(1)$. That is, it appears that the graph of f would be the straight line which passes through the origin and has its slope equal to f(1). We shall have more to say about this much later in the course under the heading of <u>linear functions</u>. With respect to our present topic all we are saying is that if we define f by $f(x) = \sin x$, this f does not have the property of being a linear function.



If $f(a + b) = f(a) + f(b)$ the circled points all belong to the graph $y = f(x)$.

One way of doing this is to drop a perpendicular from Q to the extension of PS, meeting this extension at T. Then at T we would drop a perpendicular to QR, meeting QR at V. Thus:



$$\begin{cases} \overline{QV} = \overline{QT} \cos B \\ \overline{TU} = \overline{PT} \sin B \\ \dfrac{\overline{QT}}{\overline{PQ}} = \sin A \\ \dfrac{\overline{PT}}{\overline{PQ}} = \cos A \end{cases}$$

(Figure 9)

$$\sin(A + B) = \frac{\overline{QR}}{\overline{PQ}} = \frac{\overline{QV} + \overline{RV}}{\overline{PQ}} = \frac{\overline{QV} + \overline{TU}}{\overline{PQ}}$$

$$= \frac{\overline{QT} \cos B + \overline{PT} \sin B}{\overline{PQ}}$$

$$= \left(\frac{\overline{QT}}{\overline{PQ}}\right) \cos B + \left(\frac{\overline{PT}}{\overline{PQ}}\right) \sin B$$

$$= \sin A \cos B + \cos A \sin B$$

In similar ways, one can also derive the results:

$$\cos(A + B) = \cos A \cos B - \sin A \sin B \qquad (3)$$

$$\sin(A - B) = \sin A \cos B - \sin B \cos A \qquad (4)$$

$$\cos(A - B) = \cos A \cos B + \sin A \sin B \qquad (5)$$

(These equations can also be derived from (2) together with some of our previous results. For example, another way to derive (5) is:

$$\begin{aligned} \cos(A - B) &= \sin(90^\circ - [A - B]) \\ &= \sin([90^\circ - A] + B), \text{ and by (2)} \\ &= \sin(90^\circ - A) \cos B + \sin B \cos(90^\circ - A) \\ &= \cos A \cos B + \sin B \sin A \\ &= \cos A \cos B + \sin A \sin B \end{aligned}$$

Equations (2) and (3) can be used to obtain some important results known as the double angle formulas. Namely we need only let A = B in both (2) and (3) to obtain:

$$\sin(A + A) = \sin A \cos A + \cos A \sin A, \text{ or}$$
$$\sin 2A = 2 \sin A \cos A \qquad (6)$$

$$\cos(A + A) = \cos A \cos A - \sin A \sin A, \text{ or}$$
$$\cos 2 A = \cos^2 A - \sin^2 A \qquad (7)$$

Recalling also that $\sin^2 A + \cos^2 A = 1$, we have that $\cos^2 A = 1 - \sin^2 A$ and $\sin^2 A = 1 - \cos^2 A$. We can combine these results with Equation (7) to obtain the equally important half-angle formulas:

$$\cos 2 A = (1 - \sin^2 A) - \sin^2 A, \text{ or}$$
$$\sin^2 A = \frac{1 - \cos 2 A}{2} \qquad (8)$$

and

$$\cos 2 A = \cos^2 A - (1 - \cos^2 A), \text{ or}$$
$$\cos^2 A = \frac{1 + \cos 2 A}{2} \qquad (9)$$

Notice also that by adding, for example, Equations (3) and (5), we obtain:

$\cos(A + B) + \cos(A - B) = 2\cos A \cos B$. If we let $X = A + B$ and $Y = A - B$, we obtain:

$$\cos X + \cos Y = 2\cos\left(\frac{X + Y}{2}\right)\cos\left(\frac{X - Y}{2}\right) \qquad (10)$$

Recipes like (10) allow us to replace sums of functions by products, and in many equations the factored form is more advantageous (i.e., $xy = 0$ tells us more about the explicit value of x or y than does $x + y = 0$). At any rate, it is not our purpose to rederive completely a course in traditional trigonometry.

The above remarks then complete our revisit to traditional trigonometry. What we would like to do next (and we shall in the next section) is to free the trigonometric functions from their dependence on angles. More specifically we shall try to define the trigonometric functions in such a way that their domain will be the set of real numbers. In this way we can think of the trigonometric functions as being in the category of functions of a real variable, which is after all, the main theme of this course.

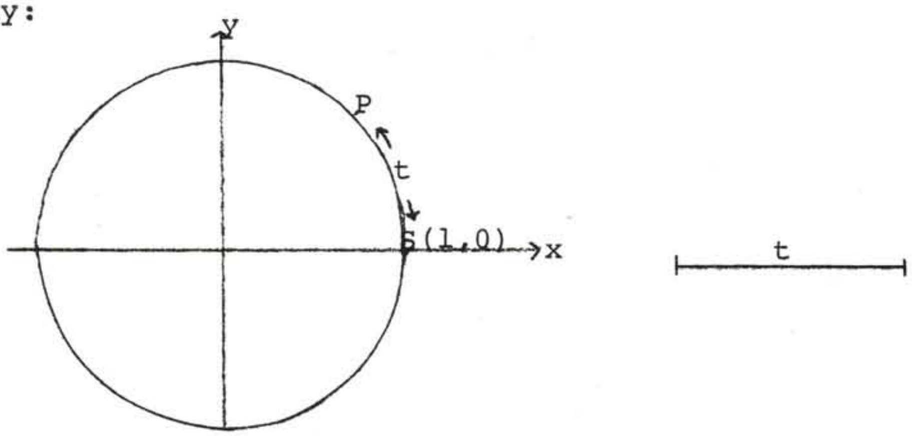
At the same time, we must take care to make sure that our new and more general definitions do not destroy the results that we already know are true for angles. In other words we do not want to wind up with two completely different sets of functions, each of which is called THE trigonometric functions.

B. Trigonometry Without Triangles

As we shall see in a later section, it is very important that we be able to define the trigonometric functions without any reference to angles or triangles.

While it is very nice to be able to motivate any new construction in terms of a practical application, we are now at a stage where it is easier to talk about the construction than it is to motivate why we desire such a construction. At any rate, since we are at liberty to invent any well-defined function whenever we please, let us use this as an initial motivation for our procedure.

The key to our construction will be the unit circle. In terms of coordinate geometry, unless otherwise stated, the unit circle refers to the circle centered at the origin with radius equal to 1 unit. In other words, by the unit circle we mean the circle whose equation is $x^2 + y^2 = 1$. Pictorially:

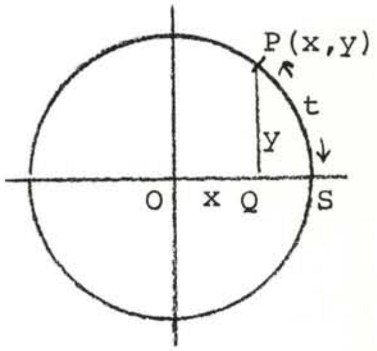


(Figure 10)

Next, let t denote any real number, Then we may view t as a length. We take the length, t, and mark it off along the unit circle starting at S(1,0) (see Figure 10). We move counterclockwise if t is positive and clockwise if t is negative. We label the point at which the length t terminates P. (Note $t \neq \overline{PS}$ since t is marked off <u>along</u> the circle.)

Let us pretend for the moment that the words sine and cosine have never before been invented (of course we know they have been invented and this will certainly affect what we are about to do next). We look at the point

P(x,y) at which the length t terminated and we define sin t to be equal to y and we define cos t to be equal to x. Again in terms of a picture:



(Figure 11)

$\sin t = y = \overrightarrow{QP}$ $\cos t = x = \overrightarrow{OQ}$

(We use the arrows to indicate directed distance. That is, nothing forces us to insure that P lies in the first quadrant. In other quadrants x and/or y need not be positive.)

The most important thing to notice now is that we have managed to define sine and cosine as functions in such a way that both the domain and image of each of these functions consist of real numbers. Moreover, we have accomplished our definition in such a way that sine and cosine obey the same relationships in this new context as they did in the traditional context.

For example, it is easy to show that for any real number, t, $\sin^2 t + \cos^2 t = 1$. Indeed since $\sin t = y$ and $\cos t = x$, the fact that $x^2 + y^2 = 1$ guarantees the desired result.

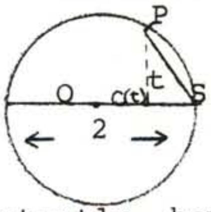
At this point, it might be good to back off for a moment and to consider the danger involved in having the same word mean two different things. For example, we now have two definitions of sin t, one when t is an angle and the other when t is a number. Therefore, whenever we see an expression such as sin t, how can we tell which definition of sine is being used? (This occured once before in our course with the introduction of differentials dy and dx. Namely after we "invented" dy and dx as separate entities, the expression $\frac{dy}{dx}$ took on two meanings - one as the derivative and the other as the quotient of dy and dx. We then had to take care that our new definitions were compatible with the old, at least in the sense that the value of $\frac{dy}{dx}$ did not depend on which of the two possible interpretations were employed.)

In any case, let us now return to our unit circle and study, for example, sin t for some <u>number</u> t.

In Figure 11, we notice that sin t is precisely $\overrightarrow{QP}$. At the same time, since $\overline{OP} = 1$, it is also clear in the traditional sense that $\sin(\sphericalangle POQ) = \overrightarrow{QP}$. In other words, if we were to invent a new unit by which to measure $\sphericalangle POQ$, it is possible that we can write sin t so that no matter which of the two interpretations we use, sin t will have the same value.

With this in mind we invent the notion of a <u>radian</u>. To measure an angle in radians, we imbed it in the unit circle and measure the length of the <u>arc</u>* it subtends. If this length is t units, we define the measure of the angle to be t radians. Therefore, as long as it is understood that the unit for measuring angles is radians, there is no

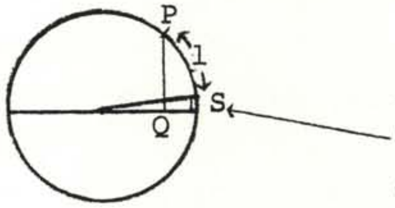
*It is crucial that we recognize that the length t is marked off along the circle. It is not the straight-line distance. Certainly, we have the right to invent such a function if we wish. For example, we could have formed a function, say, C as follows: Given the number t use S as a center and swing an arc of radius t and let P denote the point at which this arc meets the unit circle. We can then define C(t) as being the x-coordinate of P. Of course, it is important to note that if we elect to use such a definition then the domain of C would be the set $\{t: 0 \leqslant t \leqslant 2\}$ since if t exceeds 2 the arc drawn from S will not meet the circle. That is:



With S as center an arc of length in excess of 2 will not meet the circle.

More importantly, however, keep in mind that we can invent functions in any way that we choose, but that if we have a specific aim in mind some well-defined functions will fulfill this aim better than other well-defined functions.

ambiguity in talking about sin t whether we think of t as being a number or as being an angle. Moreover, we are then free to use whichever of the two interpretations we prefer in a given situation depending on which better serves our needs without fear that we can wind up with contradictory results. (This, too, happened with differentials. Namely we could write either $(\frac{dy}{dx})$ or dy ÷ dx depending on which of the two representations was the more convenient in the given problem.) However, if our unit of angular measure were still degrees, then certainly there is a difference between saying sin 1 and saying sin 1°. That is:



sin 1 = $\overline{PQ}$ = sin 1 radian
This "very short" length denotes sin 1°.

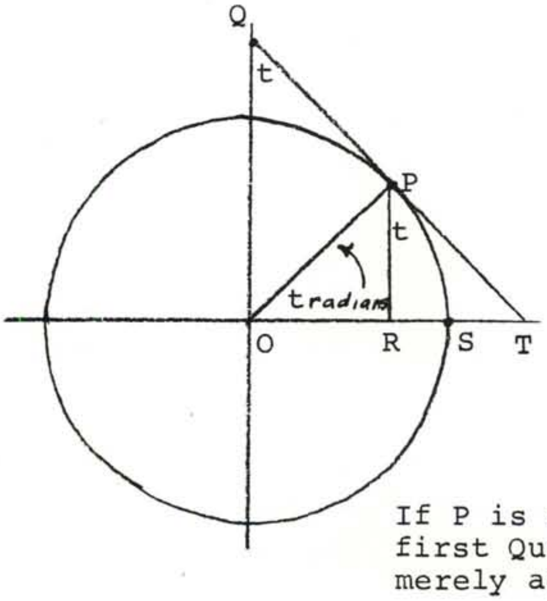
At this point let us take exception to a remark made in most textbooks. It is often said that radians are a dimensionless unit for measuring angles. While, from a practical point of view, this works out to be the case, the true fact is that radians are as much a unit for measuring angles as are degrees. The main idea is that if the unit is radian then the value obtained for the trigonometric function of the angle is the same as it would have been for the pure number.

A final point that we wish to make in this connection is that unless we were trying to identify angles with numbers there would have been no need to invent radian measure.

That is, why should we go to the trouble of having two different ways to measure an amount of rotation (angular measure) unless one way offered us advantages not afforded by the other? Moreover, with regard to angular measure, notice how much more natural degrees are than radians. First of all, it seems fairly natural to talk about dividing a circle into a whole number (360) of equal parts, but to talk about 2π radians (which is the circumference of the unit circle) is far from a natural concept.

In summary, then, we can invent the trigonometric functions so that their domain is the set of real numbers rather than angles, and if we also introduce the notion of radian measure the two definitions of trigonometric functions may be considered as being equivalent.

By way of additional practice, let us observe that viewing sine and cosine as functions of numbers, we can mimic the remaining definitions of the trigonometric functions by letting $\tan t = \frac{\sin t}{\cos t}$, and using the reciprocal relations $\csc t = 1/(\sin t)$, $\sec t = 1/(\cos t)$ and $\cot t = 1/(\tan t)$. Pictorially:



$$\left.\begin{aligned}
\sin t &= \overrightarrow{RP}\\
\cos t &= \overrightarrow{OR}\\
\tan t &= \frac{\overline{PT}}{\overline{OP}} = \overline{PT}\\
\cot t &= \frac{\overline{QP}}{\overline{OP}} = \overline{QP}\\
\csc t &= \frac{\overline{OQ}}{\overline{OP}} = \overline{OQ}\\
\sec t &= \frac{\overline{OT}}{\overline{OP}} = \overline{OT}
\end{aligned}\right\}$$

Thus each of the six trigonometric functions maps the number (length) t into a number (length).

If P is not in the first Quadrant we must merely adjust signs.

We also see that, since the circle has circumference 2π, if f denotes any trigonometric function then $f(x+2\pi)=f(x)$. (Notice here that we are not saying that $x = x + 2\pi$. In terms of modern language we are only saying that these two different elements in the domain have the same image with respect to f.)

We can also show that such results as sin 0 = 0, cos 0 = 1, $\sin \frac{\pi}{2} = 1$, and $\cos \frac{\pi}{2} = 0$ follow merely by reading the coordinates of P. Moreover, these results do not depend on whether we are thinking of numbers or angles, provided that the angles are measured in radians. In this respect, notice that since the circumference of the circle is 2π, an angle of 90° corresponds to an angle of $\frac{2\pi}{4} = \frac{\pi}{2}$ radians. We are not, however, saying that $90 = \frac{\pi}{2}$. Indeed, $\frac{\pi}{2}$ is about 1.57 since π is about 3.14 (in terms of a very elementary analogy, we do not say that 12 = 1, even though it is true that 12 inches = 1 foot).

Since the trigonometric functions are now "legitimate" functions of a real variable, it makes sense to talk about such things as $\lim_{x \to a} f(x)$, where f denotes any trigonometric function. Not only can we apply our limit theorems to the study of the trigonometric functions but we can apply our entire study of derivatives to them as well. This we shall do in the next section. It turns out that one of the key pieces of information we shall need is the fact that:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

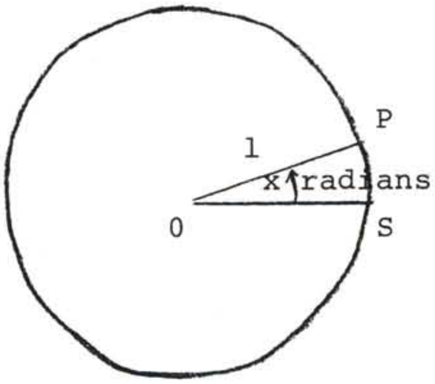
(and let us hope that none among us believes this result is obtained trivially from 0/0, since the numerator and denominator are equal. Recall that we arrive at 0/0 "illegally" by allowing x to equal 0, which is not permitted in the definition of $x \to 0$.)

At any rate, for a review of both the trigonometric functions and limits, let us try to derive the above limit.

While we are thinking of x as being a number, we shall utilize our remarks about radian measure so that we can take advantage of the geometry of the situation. What we shall do is to compute $\lim\limits_{x \to 0^+}$ and $\lim\limits_{x \to 0^-}$. The reason for this is that near 0, sin x is positive for positive values of x and negative for negative values of x. Since we will use inequalities in arriving at the correct answer we will distinguish between these two possibilities, since, as we have seen, multiplying an equality by a negative number reverses the direction of the inequality.

With this in mind, we proceed as follows:

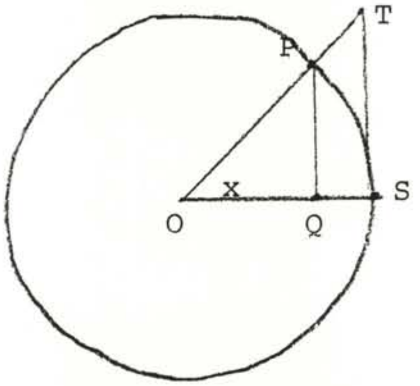
We first let x be a small positive number. Our picture becomes:



Now, since the area of the unit circle is π, the area of sector $\overset{\frown}{POS}$ is $\left(\frac{x}{2\pi}\right)\pi$ because $\measuredangle POS$ is $\frac{x}{2\pi}$ of the measure of the entire circle. (Notice that we do not say $\frac{x}{360}$.

We must use radian measure if sin x is to be unambiguous.)

We now construct PQ perpendicular to OS and TS perpendicular to OS. Thus:



The point is that $\Delta POQ \subset$ sector POS $\subset \Delta OTS$. Hence:

Area of ΔPOQ < Area of sector POS < Area of ΔOTS

$$\therefore \frac{1}{2}(\overline{PQ})(\overline{OQ}) < \left(\frac{x}{2}\right) < \frac{1}{2}(\overline{TS})(\overline{OS})$$

But $\overline{PQ} = \sin x$, $\overline{OQ} = \cos x$, $\overline{TS} = \tan x$ and $\overline{OS} = \overline{OP} = 1$

$$\therefore \frac{1}{2}\sin x \cos x < \frac{x}{2} < \frac{1}{2}\tan x = \frac{1}{2}\ \frac{\sin x}{\cos x}$$

$$\therefore \sin x \cos x < x < \frac{\sin x}{\cos x} \qquad (11)$$

$$\therefore \cos x < \frac{x}{\sin x} < \frac{1}{\cos x} \qquad (12)$$

(Notice that in going from (11) to (12) we required that sin x be positive, otherwise the inequality would be reversed.)

$\therefore$ By the "sandwich theorem":

$$\lim_{x\to 0^+} \cos x \leqslant \lim_{x\to 0^+} \frac{x}{\sin x} \leqslant \lim_{x\to 0^+} \frac{1}{\cos x} \qquad (13)$$

However $\lim\limits_{x\ 0^+} \cos x = 1$, hence (13) yields:

$$1 \leqslant \lim_{x\to 0^+} \frac{x}{\sin x} \leqslant 1 \qquad (14)$$

From (14) we see that $\lim\limits_{x\to 0^+} \dfrac{x}{\sin x} = 1$, (15)

and since $\lim\limits_{x\to a} f(x) = \dfrac{1}{\lim\limits_{x\to a} \dfrac{1}{f(x)}}$ as long as $\lim\limits_{x\to a} f(x) \neq 0$,

(15) yields

$$\lim_{x\to 0^+} \frac{\sin x}{x} = \frac{1}{1} = 1 \qquad (16)$$

If we assume that x is now negative, and sufficiently small, our diagram yields the fourth guadrant. Leaving the details as an exercise, the same sequence of steps as before (except we reverse the sign of the inequality when we divide by sin x) yields:

$$\lim_{x\to 0^-} \frac{\sin x}{x} = 1 \qquad (17)$$

Combining (16) and (17) yields the desired result.

Notice that if we wished to use degrees we could still compute $\lim_{x \to 0} \frac{\sin x}{x}$, but the result would not be 1. More specifically, if the sector is a "slice" of x degrees, we have $\frac{x}{360}$ of the entire circle. Hence the area of the sector is $\left(\frac{x}{360}\right)\pi$.

In this case our inequality is:

$$\frac{1}{2}\,(\overline{PQ})\,(\overline{OQ}) \ < \ \frac{\pi x}{360} \ < \frac{1}{2}(\overline{TS})\,(\overline{OS})$$

Thus (11) could be replaced by:

$$\sin x \cos x < \frac{\pi x}{180} < \frac{\sin x}{\cos x} \qquad (11')$$

or $\frac{180}{\pi} \cos x < \frac{x}{\sin x} < \frac{180}{\pi \cos x}$, for small positive x and this leads to the result that:

$$\lim_{x \to 0} \frac{x}{\sin x^\circ} = \frac{180}{\pi}$$

or

$$\lim_{x \to 0} \frac{\sin x^\circ}{x} = \frac{\pi}{180}$$

At any rate, the fact that if x is a real number $\lim_{x \to 0} \frac{\sin x}{x} = 1$ plays the key role in finding the derivatives of the trigonometric functions. Just how this occurs is the subject of our next section.

C. Derivatives of the Trigonometric Functions

Given that f(x) = sin x, it makes sense to try to find f'(x). The important point is that, regardless of how f is defined, f' is always defined by:

$$f'(x) = \lim_{\Delta x \to 0} \left[\frac{f(x + \Delta x) - f(x)}{\Delta x} \right] \qquad (1)$$

In the special case f(x) = sin x, (1) becomes:

$$f'(x) = \lim_{\Delta x \to 0} \left[\frac{\sin(x + \Delta x) - \sin x}{\Delta x} \right] \qquad (2)$$

To handle (2) we must invoke results concerning the trigonometric functions. For example (and there are other approaches), the recipe for sin(A + B) yields:

$$\sin(x + \Delta x) = \sin x \cos \Delta x + \sin \Delta x \cos x \qquad (3)$$

Putting (3) into (2), we obtain:

$$\left.\begin{aligned} f'(x) &= \lim_{\Delta x \to 0} \left[\frac{\sin x \cos \Delta x + \sin \Delta x \cos x - \sin x}{\Delta x} \right] \\ &= \lim_{\Delta x \to 0} \left[\frac{\sin x (\cos \Delta x - 1) + \sin \Delta x \cos x}{\Delta x} \right] \\ &= \lim_{\Delta x \to 0} \left[\sin x \left(\frac{\cos \Delta x - 1}{\Delta x} \right) + \left(\frac{\sin \Delta x}{\Delta x} \right) \cos x \right] \end{aligned}\right\} \qquad (4)$$

If we now apply our limit theorems to (4) (and again notice that these theorems do not depend on the specific functions in question), we obtain:

$$f'(x) = \left(\lim_{\Delta x\to 0} \sin x\right) \lim_{\Delta x\to 0} \left(\frac{\cos \Delta x - 1}{\Delta x}\right) + \left(\lim_{\Delta x\to 0} \frac{\sin \Delta x}{\Delta x}\right) \lim_{\Delta x\to 0} \cos x$$

$$= \sin x\left[\lim_{\Delta x\to 0} \left(\frac{\cos \Delta x-1}{\Delta x}\right)\right] + \left[\lim_{\Delta x\to 0} \frac{\sin \Delta x}{\Delta x}\right] \cos x \qquad (5)$$

Equation (5) supplies us with the hind-sight motivation as to why we wanted to investigate $\lim_{x\to 0} \frac{\sin x}{x}$ in the previous section. Pedagogically, it was to our advantage to introduce this limit before we needed it, so that we would feel at home with it when we did need it. Pragmatically, it is fair to assume in "real life" that one might have tried to find f'(x) where f(x) = sin x and then arrived at (5). At this point, he would have had ample motivation for investigating $\lim_{x\to 0} \frac{\sin x}{x}$ and its associate $\lim_{x\to 0} \frac{1 - \cos x}{x}$.

Now, in the last section we showed that $\lim_{x\to 0} \frac{\sin x}{x} = 1$ (Again notice that x is not important here. What is important is that $\lim_{[]\to 0} \frac{\sin []}{[]} = 1$. In particular, $\lim_{\Delta x\to 0} \frac{\sin \Delta x}{\Delta x} = 1$.) and that $\lim_{x\to 0} \frac{1 - \cos x}{x} = 0$ (See exercise 3.1 1(2))

Putting these results into (5), we obtain:

$$f'(x) = (\sin x)\ (0) + (1) \cos x$$

In other words, we have now proved:

$$\boxed{\text{If } f(x) = \sin x, \text{ then } f'(x) = \cos x}$$

Our next observation is that we may now apply all of our calculus theorems to this result to obtain even more information.

For example, we may invoke the chain rule to conclude that if y = sin u(x), where u(x) is any differentiable function of x, then $\frac{dy}{dx}$ exists and is given by:

$$\frac{dy}{dx} = \cos u \frac{du}{dx} \tag{6}$$

(That is: $\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \cos u \frac{du}{dx}$)

If we couple (6) with the fact that $\cos x = \sin\left(\frac{\pi}{2} - x\right)$, we also obtain:

$$\frac{d}{dx}(\cos x) = \frac{d}{dx}\sin\left(\frac{\pi}{2} - x\right) = \frac{d\left(\sin\left[\frac{\pi}{2} - x\right]\right)}{d\,\frac{\pi}{2} - x}\;\frac{d\left(\frac{\pi}{2} - x\right)}{dx}$$

$$= \cos\left(\frac{\pi}{2} - x\right)(-1)$$

$$= -\cos\left(\frac{\pi}{2} - x\right)$$

$$= -\sin x$$

and invoking the chain rule again, we obtain:

$$\frac{d}{dx}(\cos u) = -\sin u \frac{du}{dx} \tag{7}$$

From results such as (6) and (7) we may apply the quotient rule to, say, $\tan x = \frac{\sin x}{\cos x}$ to obtain:

$$\frac{d\ (\tan x)}{dx} = \frac{d\left(\frac{\sin x}{\cos x}\right)}{dx} = \frac{\cos x \frac{d(\sin x)}{dx} - \sin x \frac{d(\cos x)}{dx}}{\cos^2 x}$$

$$= \frac{\cos^2 x - \sin x(-\sin x)}{\cos^2 x}$$

$$= \frac{\cos^2 x + \sin^2 x}{\cos^2 x}$$

$$= \frac{1}{\cos^2 x} \quad \text{or } \sec^2 x$$

More results are obtained as standard material in any calculus book and such reading will be assigned to supplement the results derived here.

Our main purpose here was to show how the knowledge that $\lim\limits_{x\to 0} \frac{\sin x}{x} = 1$ leads in a very neat way to the calculus of the trigonometric functions.

There remains, however, one major demonstration which must be shown if we are to keep our promise of motivating the trigonometric functions without regard to angular measure.

To this end, let us look at

$$x = A \sin (\omega t + \alpha) \tag{8}$$

where A, ω, and α are constants.

Taking the derivative in (8) with respect to t, we obtain:

$$\frac{dx}{dt} = A \cos (\omega t + \alpha)\ \omega = A\omega\ \cos (\omega t + \alpha)^{*}$$

and

$$\frac{d^2x}{dt^2} = \frac{d}{dt}\left(\frac{dx}{dt}\right) = - A\omega^2 \sin (\omega t + \alpha) \qquad (9)$$

Recalling from (8) the fact that $x = A \sin (\omega t + \alpha)$, (9) yields:

$$\boxed{\frac{d^2x}{dt^2} = - \omega^2 x} \qquad (10)$$

Now, in the subject called differential equations, one starts with (10) and derives (8). (We shall do this as an exercise later in the course.) This is more difficult than starting with (8), as we did, and deriving (10). (In a manner of speaking, this comes under the adage that it is more difficult to unscramble an egg than to scramble one.) For our purposes, however, let us pretend that we started with (10). Let us think of a particle moving along the x-axis and let t denote time. Then (10) says

*Again we use the chain rule: $\frac{d \sin(\omega t + \alpha)}{dt}$ is <u>not</u> $\cos(\omega t + \alpha)$. Rather $\frac{d \sin u}{du} = \cos u$ means that $\frac{d \sin (\omega t + \alpha)}{d(\omega t + \alpha)} = \cos(\omega t + \alpha)$. Hence by the chain rule, $\frac{d \sin(\omega t + \alpha)}{dt} =$

$$\frac{d \sin(\omega t + \alpha)}{d(\omega t + \alpha)}\ \frac{d(\omega t + \alpha)}{dt} = \cos(\omega t + \alpha)\ \ \omega$$

that at any time t the acceleration of the particle, $\frac{d^2x}{dt^2}$, is proportional to its displacement (x), but with the opposite sign (ω^2 is positive hence - ω^2 is negative. Therefore $\frac{d^2x}{dt^2}$ and x have opposite signs).

Thus (10) is just the statement that the particle is moving in simple harmonic motion and one needs no knowledge of classical trigonometry to understand the physical situation depicted by (10). The solution to (10) (that is, the solution of x as an explicit function of t), however, is (8), and (8) certainly makes use of the trigonometric functions. In other words, $x = A \sin(\omega t + \alpha)$ is a real solution to a real problem which involves no angular measure. In still other words, in the expression $x = A \sin(\omega t + \alpha)$, $\omega t + \alpha$ is not an angle!

To put it still differently, had the trigonometric functions not yet been invented and one had arrived, for some reason or another, at equation (10), then one would have had to invent the trigonometric functions to obtain x explicitly in terms of t.

This completes our attempt to motivate trigonometry without triangles.

D. Circular and Hyperbolic Functions

Later in this course we shall give practical reasons for inventing the hyperbolic functions introduced in this section. For the time being, we would like to point out that once we liberated the so-called trigonometric functions from the study of angular measurement, there was no real reason to keep calling them the trigonometric functions, at least in the traditional sense of trigonometry.

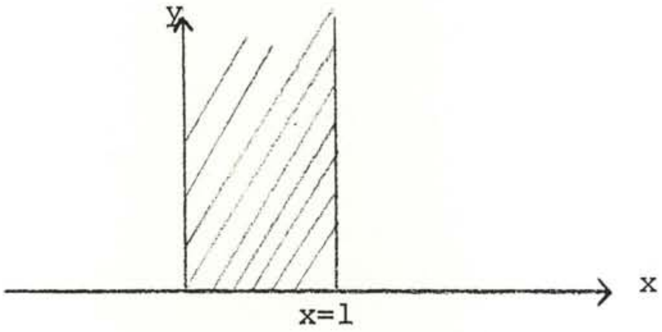
Since the construction of the non-triangular trigonometric functions uses the unit circle, it is customary to refer to these functions as the circular functions rather than as the trigonometric functions. To be sure, in many ways this is merely a question of semantics, but in other ways we can show that the extended definition of trigonometric functions goes beyond their usage as circular functions.

In particular we would like, in this section, to introduce the hyperbolic functions although we will not treat these from a calculus point of view until later in our course. In the same way that the circular functions are based on the curve whose equation is $x^2 + y^2 = 1$, the hyperbolic functions are based on the curve whose equation is $x^2 - y^2 = 1$. This curve is a hyperbola, which, for purposes of identification, we may call the unit hyperbola. In terms of the calculus we have already studied, it is not difficult to sketch the graph of $x^2 - y^2 = 1$. Namely:

We first observe that $x^2 - y^2 = 1$ is symmetric with respect to both the x-and y-axes. Hence it is sufficient to know the graph in the first quadrant. To this end, we study $x^2 - y^2 = 1$ where $x, y \geqslant 0$.

$$\therefore \quad y = +\sqrt{x^2 - 1} \qquad x > 0$$

$\therefore \quad x \geqslant 1$ otherwise y would not be a real number



Next $\frac{dy}{dx} = \frac{2x}{2\sqrt{x^2 - 1}} = \frac{x}{\sqrt{x^2 - 1}}$

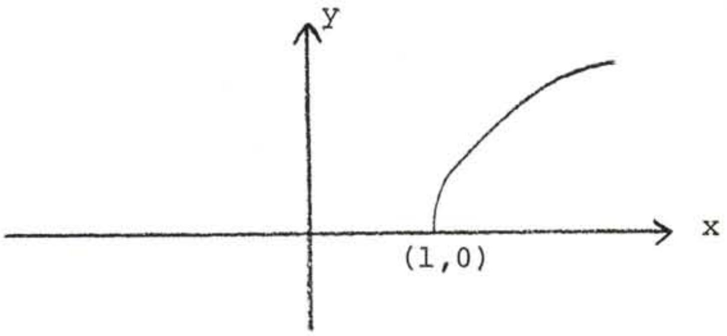
$\therefore$ $x > 0 \rightarrow \frac{dy}{dx} > 0 \rightarrow$ curve rises in first guadrant

Finally $\frac{dy}{dx} = \frac{x}{\sqrt{x^2 - 1}} \rightarrow$

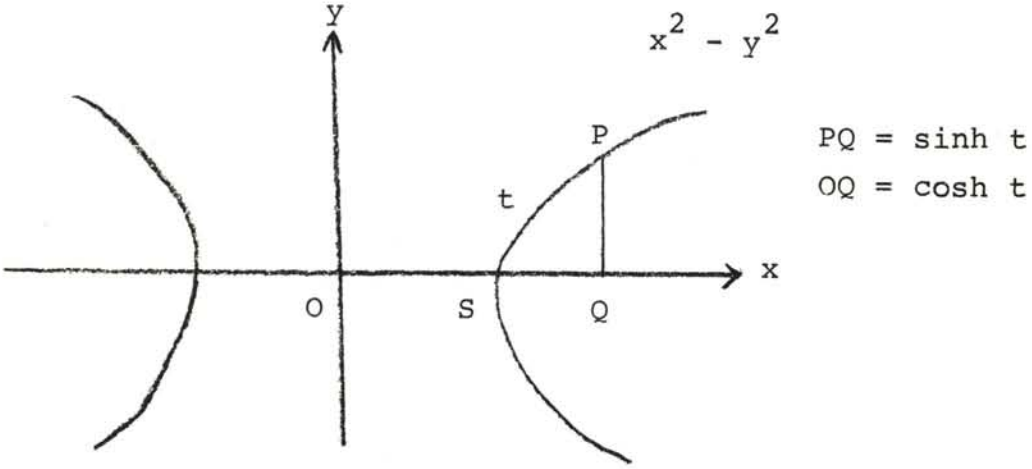
$$\frac{d^2y}{dx^2} = \frac{d\left(\frac{dy}{dx}\right)}{dx} = \frac{\sqrt{x^2 - 1} - x\, \frac{d\sqrt{x^2 - 1}}{dx}}{\left(\sqrt{x^2 - 1}\right)^2} =$$

$$\frac{\sqrt{x^2 - 1} - x\, \frac{x}{\sqrt{x^2-1}}}{x^2 - 1} = \frac{-1}{(x^2-1)^{3/2}} < 0$$

$\therefore$ Curve spills water in first quadrant



Hence, by symmetry, the graph of $x^2 - y^2 = 1$ is given by:

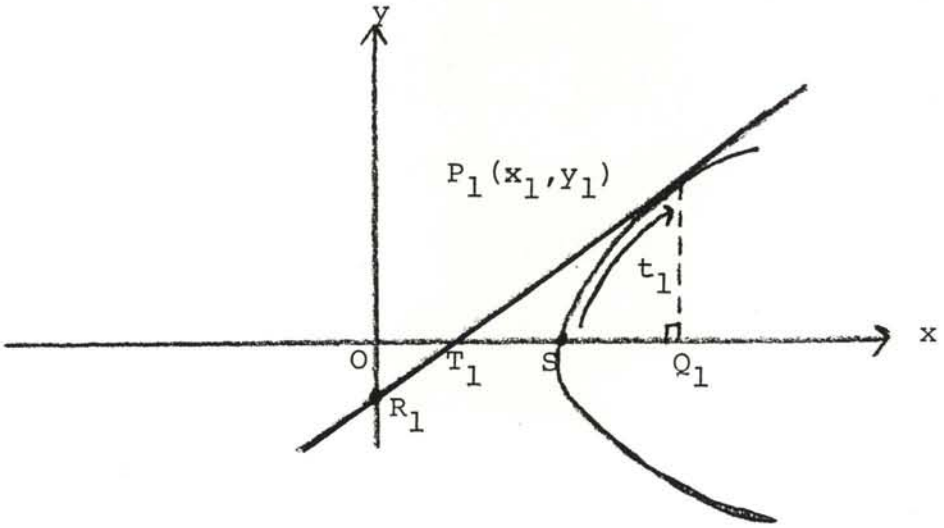


Unlike the circle, the hyperbola has two disconnected sections. We agree to use the section for which x is positive. Letting S denote (1,0), we mimic our procedure for the circular functions. Namely, given the real number t, we mark if off, starting at S, along the branch of

$x^2 - y^2 = 1$ for which x is positive. We mark t off along the upper portion if t is positive and along the lower portion if t is negative. If t terminates at P(x,y) we define x = cosh t, (hyperbolic cosine) and y = sinh t (hyperbolic sine). In a rather obvious way we may now define $\tanh t = \frac{\sinh t}{\cosh t}$, $\coth t = \frac{1}{\tanh t}$, $\operatorname{sech} t = \frac{1}{\cosh t}$, and $\operatorname{csch} t = \frac{1}{\sinh t}$.

Whereas the basic identity for the circular functions was $\sin^2 t + \cos^2 t = 1$, for the hyperbolic functions it is $\cosh^2 t - \sinh^2 t = 1$, since $x^2 - y^2 = 1$.

We can also visualize these functions pictorially. For example, pick a real number t_1, (in the diagram which follows we assume t_1 is positive but this is not crucial). Then, if $P_1(x_1,y_1)$ denotes the point at which the length t_1 terminates, we have $x_1 = \cosh t_1$ and $y_1 = \sinh t_1$.



Let us now find the equation of the line tangent to $x^2 - y^2 = 1$ at P_1. Differentiating $x^2 - y^2 = 1$ implicitly we obtain:

$$2x - 2y \frac{dy}{dx} = 0$$

$$\therefore \frac{dy}{dx} = \frac{x}{y}$$

Therefore at P_1 the tangent has slope $\frac{x_1}{y_1}$; hence its equation is:

$$\frac{y - y_1}{x - x_1} = \frac{x_1}{y_1}$$

or $yy_1 - {y_1}^2 = xx_1 - {x_1}^2$

$$\therefore yy_1 = xx_1 - \left({x_1}^2 - {y_1}^2\right)$$

or $yy_1 = xx_1 - 1$ since (x_1,y_1) satisfies $x^2 - y^2 = 1$.

If $x = 0$, $yy_1 = -1 \therefore$ the y intercept of this line is $-\frac{1}{y_1} = \frac{-1}{\sinh t_1}$ and if $y = 0$, $xx_1 - 1 = 0 \therefore$ the x intercept of the line is $\frac{1}{x_1} = \frac{1}{\cosh t_1}$.

Thus, given t_1 we have constructed both sech t_1 and csch t_1. In terms of our last diagram we have:

$$\overrightarrow{Q_1P_1} = \sinh t$$

$$\overrightarrow{OQ_1} = \cosh t$$

$$\overrightarrow{OT_1} = \operatorname{sech} t$$

$$\overrightarrow{R_1O} = \operatorname{csch} t$$

The hyperbolic functions, while perhaps not quite as familiar, are as real as the circular functions, and they shall play an important role in what comes later.

If we so desired we could introduce hyperbolic radians merely by talking about the arc length on the hyperbola subtended by an angle centered at the origin.

The connection between the circular and the hyperbolic functions becomes even more amazing if we allow ourselves to think in terms of complex rather than real numbers.

For example

$$x^2 - y^2 = 1$$

can be written as

$$x^2 + (iy)^2 = 1$$

Thus, a hyperbola in the x-y-plane would be a circle in the x-iy plane.

It will be shown later that both sinh t and cosh t satisfy the differential equations

$$\frac{d^2x}{dt^2} = \omega^2 x$$

(That is, the acceleration is proportional to the displacement and in the same direction. That is, the acceleration increases without bound!)

However, our immediate aim was to show that the concept of trigonometry is restricted to neither angles nor circles and that we may invent other "trigonometrics" which are quite consistent with the real world.

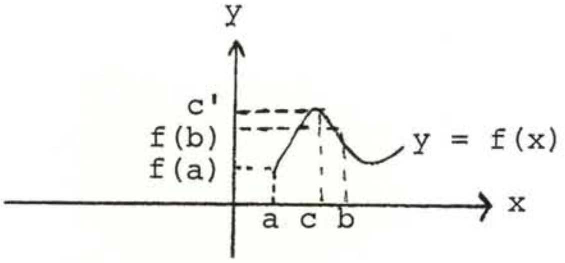
E. The Inverse Circular Functions

In this section we would like to introduce the inverse trigonometric functions and to discuss how one finds the derivatives of these functions. Before doing this, however, it might be wise to take a few minutes and reinforce some of our previous studies of inverse functions.

We have already discussed the fact that if $f:A\to B$ is both one-to-one and onto we can talk about $f^{-1}:B\to A$. We now want to investigate this idea with respect to those functions of mathematical analysis which are differentiable in some interval.

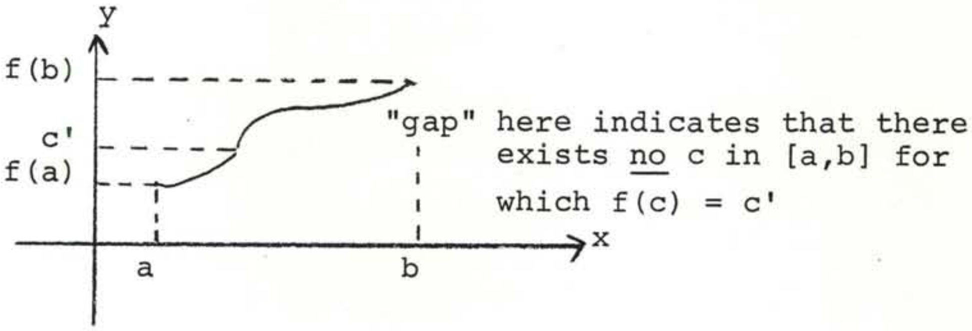
So suppose f is defined on [a,b] and differentiable on (a,b). (Actually it is not crucial that we restrict our attention to finite intervals. It is possible that f is differentiable for all real x. Our restriction is merely for the sake of being able to draw a better picture.)

Recall that since we assume f to be single-valued, we can conclude that for each c in [a,b] there exists a unique (meaning one and only one) real number c' such that f(c) = c'. We cannot conclude, however, that c' lies in [f(a), f(b)]. For example:



(Figure 12)

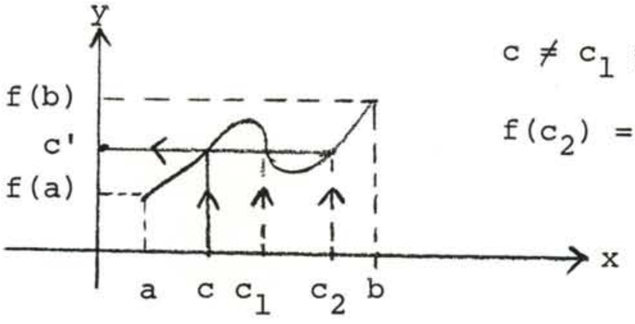
Of course, it is true that if d' lies in [f(a), f(b)] then there exists a number d in [a,b] such that f(d) = d'.* This follows from the INTERMEDIATE VALUE THEOREM about continuous functions. Namely, if f is continuous on [a,b] and if c' lies between f(a) and f(b) then there exists a c in [a,b] for which f(c) = c'. If this were not true, in terms of a graph, there would be a "gap" in the curve in the sense that it "jumps over" the height c'. That is:


y
f(b)
c'
f(a)
"gap" here indicates that there exists no c in [a,b] for which f(c) = c'
x
a
b

(Figure 13)

*We should be a bit more careful with our notation. By convention [a,b] implies that a ⩽ b. Unless more is given about f we do not know whether f(a) < f(b). Thus to conform with our convention we should write [min {f(a), f(b)}, max {f(a), f(b)}] but this is much too cumbersome and the reason we took the above liberty.

The key point now in terms of inverse functions hinges on the fact that, while for each c' in [f(a), f(b)] there exists a number c in [a,b] for which f(c) = c', IT IS BY NOW HOPEFULLY CLEAR THAT c NEED NOT BE UNIQUE. For example:



$c \neq c_1 \neq c_2$ but $f(c)=f(c_1) = f(c_2) = c'$

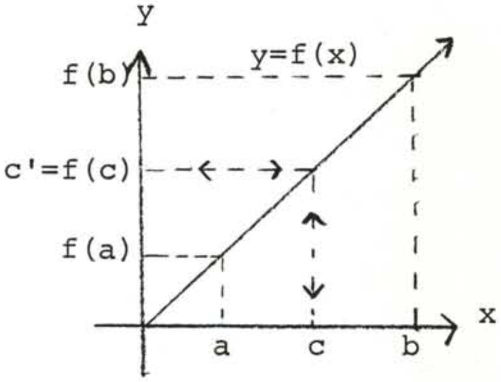
(Figure 14)

In the above diagram we have a situation wherein f is single-valued but not one-to-one. The fact that f is not one-to-one means that if f^{-1} is to exist it must be multi-valued, and we have previously agreed to exclude such functions from our studies.

What happened pictorially that describes why our function was not one-to-one? Evidently the curve doubled back. That is, the curve was allowed to change from rising to falling. In terms of derivatives this is analytically equivalent to saying that f'(x) was allowed to equal 0 for at least one x in (a,b). This motivates, hopefully, our next restriction that not only is f differentiable in (a,b) but also that for each x in (a,b), $f'(x) \neq 0$.

Again to emphasize the picture, this restriction guarantees us that the curve is either always falling or else always rising. The analytical counterpart of this statement is to say that $f(x)$ is either monotonically decreasing or else monotonically increasing. More specifically, $f'(x)$ always positive implies that if $x_1 < x_2$ then $f(x_1) < f(x_2)$, and this is what is meant by monotonically increasing.

The key now is that if f is monotonic (either increasing or decreasing) on $[a,b]$ then f^{-1} exists and has as its domain $[f(a), f(b)]$ if $f(a) < f(b)$ or $[f(b), f(a)]$ if $f(b) < f(a)$. Again, in terms of pictures:

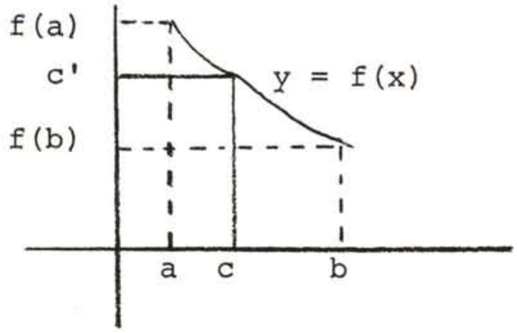


Since $f'(x) > 0$, $c \in (a,b) \rightarrow f(c) = c' \in (f(a), f(b))$. Moreover, $c_1 \neq c \rightarrow f(c_1) \neq c'$. Thus there is a 1-1 correspondence between points in $[a,b]$ and points in $[f(a), f(b)]$.

(Figure 15)

(In words f^{-1}: $[f(a), f(b)] \rightarrow [a,b]$ is defined as follows: pick $c' \in [f(a), f(b)]$ then there exists one and only one number $c \in [a,b]$ for which $f(c) = c'$. Define f^{-1} by $f^{-1}(c') = c$. In this way $f^{-1} \circ f$ is the identity map on $[a,b]$ since for each $c \in [a,b]$, $f^{-1}(f(c) = f^{-1}(c') = c$. In a similar way $f \circ f^{-1}$ is the identity map on $[f(a), f(b)]$.

The same observations apply if $f'(x) < 0$ for all $x \in [a,b]$. The only difference is the role of $f(a)$ and $f(b)$. That is



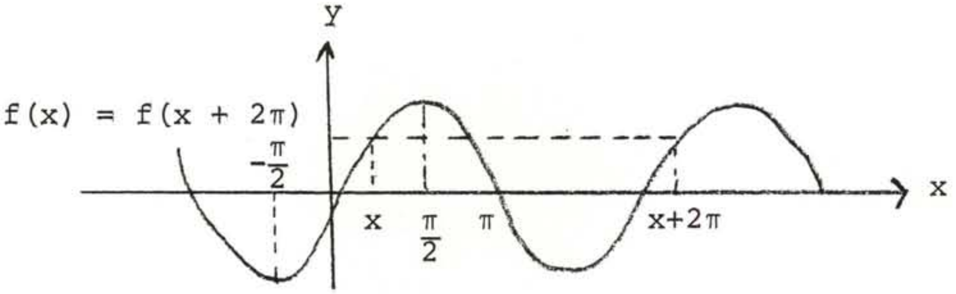
(Figure 16)

(Technical Note:

We do not say that $f \circ f^{-1} = f^{-1} \circ f$ since they may have different domains. (That is, [a,b] need not equal [f(a),f(b)]) More specifically, $f^{-1} \circ f$ is the identity map on [a,b] while $f \circ f^{-1}$ is the identity map on [f(a), f(b)] .)

Let us now apply these results to the trigonometric functions so that we may see how to "invent" the inverse trigonometric functions. For one thing, if $f:A \rightarrow B$ and if f^{-1} exists, then the domain of f^{-1} must be B. In other words, if $f:A \rightarrow B$ then $f^{-1}:B \rightarrow A$. In the case $f(x) = \sin x$, the domain of f would be the set of all real numbers while the image of f would be the interval [-1,1], that is, $B = \{y: -1 \leqslant y \leqslant 1\}$. Thus, if the inverse of the sine function exists, its domain must be B = [-1,1].

However, the sine function far from satisfies the one-to-one property that is so vital if an inverse function is to exist. Indeed, since $f(x) = f(x + 2\pi)$, we see that infinitely many numbers have the same image with respect to f. That is, $f(x) = f(x + 2\pi) = f(x + 4\pi) = \ldots =$ $f(x + 2k\pi)$ where k is any integer. Pictorially, this is reflected by:



(Figure 16)

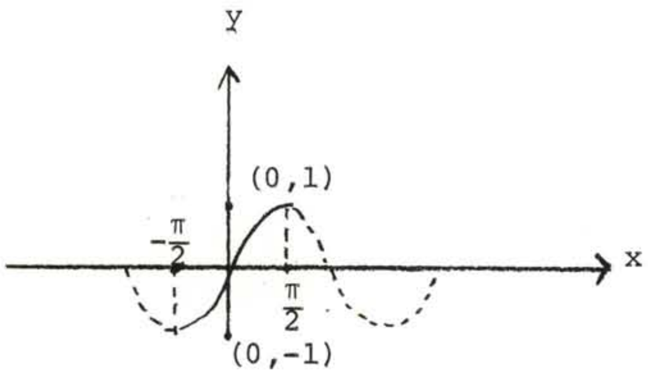
So it appears that unless we invent a new function which "looks like" sin x but which has a more restricted domain, we will be unable to talk about $\sin^{-1}x$, especially if we insist on our earlier convention that all functions must be single valued.

Referring to Figure 16, we observe that if we restrict the domain of sin x to the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, each value of the sine is taken on once and only once. In still other words, let us invent a new function, say S_1, which looks exactly like sine except that it has a different domain. That is:

$$S_1:\left[-\frac{\pi}{2} , \frac{\pi}{2}\right] \to [-1,1]$$

such that if $x \varepsilon \left[-\frac{\pi}{2} , \frac{\pi}{2}\right]$ then $S_1(x) = \sin x$.

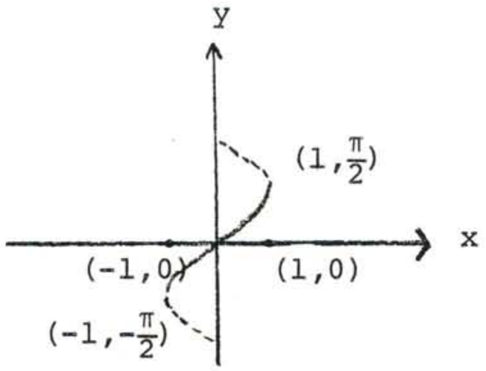
Again, in terms of a picture:



Heavy line denotes the graph of $y=S_1(x)$

(Figure 17)

The point now is that S_1 possesses all the attributes necessary for having an inverse. In fact, from Figure 17, we can readily deduce that the graph of $y = S_1^{-1}(x)$ is given by:



Graph of $y = S_1^{-1}(x)$

That is:

If S_1: $\left[-\frac{\pi}{2},\frac{\pi}{2}\right] \rightarrow [-1,1]$

then S_1^{-1}: $[-1,1] \rightarrow \left[-\frac{\pi}{2},\frac{\pi}{2}\right]$

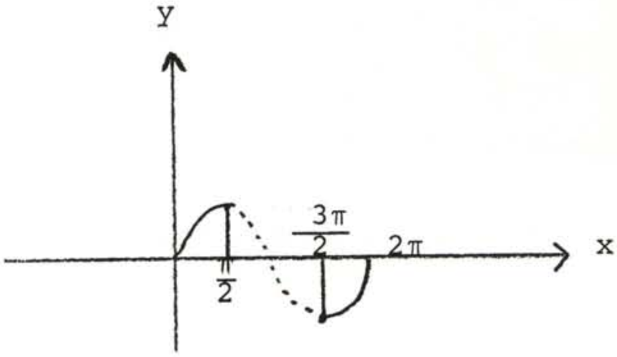
(Figure 18)

There are still a few remarks we should like to make here:

(1) In terms of our earlier remarks that the graphs of $y = f(x)$ and $y = f^{-1}(x)$ are symmetric with respect to the line $y = x$, notice that a comparison of Figures 17 and 18 shows that $y = S_1(x)$ and $y = S_1^{-1}(x)$ satisfy this criterion.

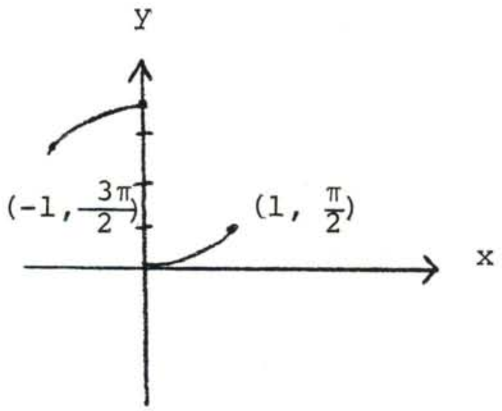
(By the way, there might be some confusion between $y = S_1^{-1}(x)$ and $x = S_1^{-1}(y)$. If S_1 has an inverse, then what is true is that $y = S_1(x)$ and $x = S_1^{-1}(y)$ are two different ways of saying the same thing. However, if we keep in mind the convention that y denotes the dependent variable and x the independent variable, then it is customary to graph S_1^{-1} in the form $y = S_1^{-1}(x)$. In fact, if this idea is not kept in mind, the entire idea of comparing the graph of $y = f(x)$ with the graph of $y = f^{-1}(x)$ would be unnecessary. That is $y = f(x)$ and $x = f^{-1}(y)$ are the <u>same</u> graph since they are but two different ways of stating the <u>same</u> relationship.)

(2) The configuration in Figure 18 is but one of several ways in which we could have invented a function S to be a suitably restricted version of the sine function. For example, $y = S(x)$ as described in Figure 19 has the property that S^{-1} exists. In fact S^{-1} is depicted in Figure 19b. One reason for shying away from such a definition of S is that S^{-1} would then have a very serious discontinuity in a neighborhood of $x = 0$. That is if x is near 0 but positive then $S^{-1}(x)$ is near 0, but if x is negative and near 0, $S^{-1}(x)$ is near 2π.



Heavy line denotes graph of $y = S(x)$ where domain of $S = \left[0, \frac{\pi}{2}\right] \cup \left[\frac{3\pi}{2}, 2\pi\right]$ with either 0 or 2π deleted since $S(0) = S(2\pi) = 0$.

(Figure 19)



Graph of $y = S^{-1}(x)$. $S^{-1}(0)$ equals both 0 and 2π ;
Hence either 0 or 2π must be deleted from the image of S^{-1}.

(Figure 19b)

While nothing prevents us from inventing S as above, it should be clear that S_1 as defined previously is "nicer" than S.

Be this as it may, let us define the "new" function S_1 by $S_1(x) = \sin x$ for all $x \varepsilon \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. A crucial observation is that sin x and $S_1(x)$ are DIFFERENT FUNCTIONS, SINCE THEY HAVE DIFFERENT DOMAINS. For further emphasis, the domain of sine is $(-\infty,\infty)$ while the domain of S_1 is $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

This restriction is not very serious but it is extremely important that you understand that there is such a restriction. It is not serious in the sense that we can define an entire sequence of functions S_n, each covering a different period of the sine function, and such that the union of this family is indeed sine. For example, we could define S_2 by $S_2(x) = \sin x$ for all $x \varepsilon \left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$, etc.

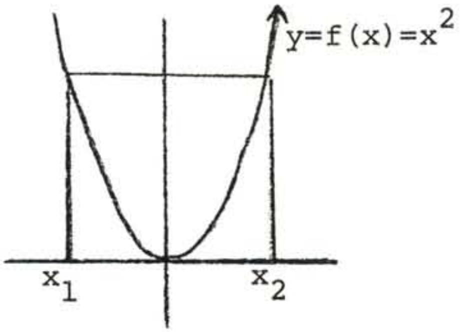
On the other hand, while the assumption is not too constraining, if we don't make it then we <u>must forfeit</u> the right to talk about the inverse functions. Without going into a philosophical discussion of values, suffice it to say that the overwhelming judgement was to make sure that we could define the inverse trig functions, and for this reason we had to introduce the notion of PRINCIPAL VALUES, which of course corresponds to our discussion of $S_1(x)$.

In still other words, then, when we talk about $g(x) = \sin^{-1}x$ we are referring NOT to the inverse of sine but to the inverse of S_1; and if for some physical reason this domain is not acceptable to us we have the right to bring into play such other functions as S_2, S_3, etc, but in a sufficiently orderly way that $\sin^{-1}x$ is a well-defined function at any given time. Once this restriction is made $\sin^{-1}(\sin(x))$ becomes a well-defined unique number - namely x itself.

At this stage of the development, it is now important to see that the properties of f^{-1} (once it exists) do not depend on the physical properties of f. That is, given that $y = f^{-1}(x)$ we may "paraphrase" this as $x = f(y)$. This follows immediately from the definition of f^{-1}. Namely, if $y = f^{-1}(x)$ then $f(y) = f(f^{-1}(x))$ [by substitution]; and $f(f^{-1}(x)) = x$ [by definition of f^{-1}]. Hence $f(y) = x$.

As a particular example this means that we can rewrite $y = \sin^{-1}x$ as $x = S_1(y)$ [or in the language of the usual textbook, $x = \sin y$ where $-\frac{\pi}{2} \leqslant y \leqslant \frac{\pi}{2}$]. AGAIN, THE MAIN POINT IS THAT IF WE DON'T USE S_1 or its equivalent, the inverse, since it is multi-valued, is not defined.

In terms of previous knowledge, the above restrictions are analogous to what happens if we would like to have an inverse to the operation known as <u>squaring a number</u>. A number has one and only one square, but two different numbers can have the same square. This problem is precisely why one talks about a <u>principal square root</u>. Certainly, the negative square root of a positive number is as good a "citizen" as the positive square root. Indeed, -2 is a well-defined number whose square is 4, just as is +2. Why then are we allowed to insist that $\sqrt{x}$ means + $\sqrt{x}$ for $x > 0$? The answer is that $f(x) = x^2$ defines a function that is not one-to-one unless we restrict our attention to an interval in which $f'(x)$ is never 0. This means that we must have an interval which does not contain 0, and this in turn means that we must either restrict f to a positive domain or else to a negative domain; and we arbitrarily chose the positive domain (the other would have done as well, but we could not choose both if we insist on one-to-one-ness). Pictorially we have:



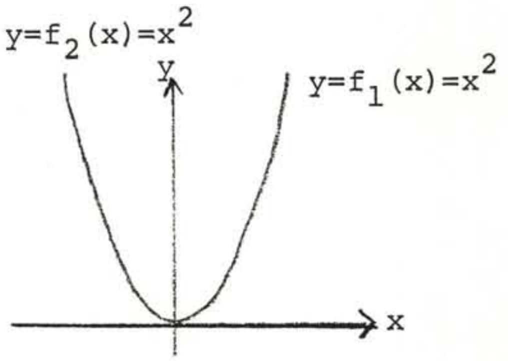
f is 1-1 if its domain is restricted to either $x \geqslant 0$ or $x \leqslant 0$; otherwise it isn't.

(Figure 20)

and when we now write f^{-1} we are not thinking of f being defined by $f(x) = x^2$ but rather as $f(x) = x^2$ FOR ALL POSITIVE x. That is, we have changed the domain of f if we want the inverse of the squaring function to exist.

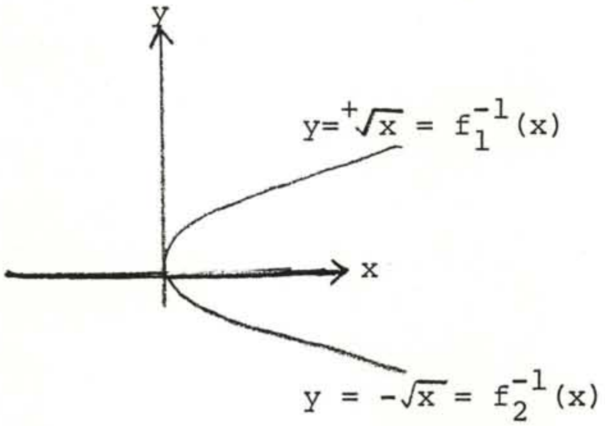
On the other hand, if we let f_1 and f_2 possess inverse functions we can then study the function $f(x) = x^2$ for all real x merely by observing that f is the union of f_1 and f_2.

In still other words, this is what we were doing when we talked about the two separate branches of the curve $y = x^2$. Again, in terms of a picture:



$f(x)=x^2$ is not 1-1; but each of the functions f_1 and f_2 is 1-1.

(Figure 21a)



$y^2 = x$ or $y = \pm \sqrt{x}$ is then the union of $y = f_1^{-1}$ and $y = f_2^{-1}(x)$.

(Figure 21b)

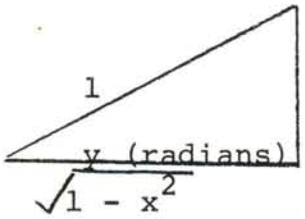
The key computational device in handling inverse functions lies in our previous observation that if $y = f^{-1}(x)$ then $x = f(y)$. Thus a knowledge of $f(x)$ is sufficient for us to determine the structure of $f^{-1}(x)$. By way of an elementary example, when we talked in grade school about subtraction being the inverse of addition, we meant that we could solve any subtraction problem by an appropriate addition problem. Thus, for example, we could view $5 - 3 = ()$ as $() + 3 = 5$.

With respect to the inverse trigonometric functions, suppose that we were given $y = \sin^{-1}x$ and we wished to determine $\frac{dy}{dx}$. Well, the definition of $y = \sin^{-1}x$ implies that $x = \sin y$ where $-\frac{\pi}{2} \leqslant y \leqslant \frac{\pi}{2}$.

From this we can obtain that $\frac{dx}{dy} = \cos y$, whereupon our result that $\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}$ (see solution to Exercise 2.3.9)

tells us that $\frac{dy}{dx} = \sec y$.

This is the correct answer, except that when y is given in terms of x, it is expected that we find $\frac{dy}{dx}$ in terms of x. To this end, the fact that $x = \sin y$ leads us to the "reference" triangle:



$\sin y = x$

From this triangle we see that $\cos y = \sqrt{1 - x^2}$. Actually, we have drawn the angle x in the first quadrant while it could have been a fourth quadrant angle. That is, while the reference triangle suggests $\cos y = \sqrt{1 - x^2}$ the fact is that it can really equal $\pm\sqrt{1 - x^2}$. The saving grace is that for the range $-\frac{\pi}{2} \leqslant y \leqslant \frac{\pi}{2}$, cos y is non-negative so that we need not worry about the minus sign. However, had we chosen a different range for y we might well have been in the predicament of having an ambiguous sign.

Note that we never have to use the reference triangle, except that it is a very convenient visual aid. Had we wished to proceed in a purely analytical way, we know that

$$\sin^2 y + \cos^2 y = 1$$

whence:

$$\cos^2 y = 1 - \sin^2 y = 1 - x^2$$

and consequently:

$$\cos y = \pm \sqrt{1 - x^2}$$

and we discard the minus sign since cos y is non-negative if $-\frac{\pi}{2} \leq y \leq \frac{\pi}{2}$.

In summary, then, we can determine that if $y = \sin^{-1} x$ then $\frac{dy}{dx} = \frac{1}{\sqrt{1-x^2}}$ merely by understanding how to differentiate the sine function and knowing what we mean by an inverse function. Other examples are contained in the reading material from the text as well as in the exercises.

Before leaving this topic, however, it is worth observing that in terms of the antiderivative we have shown that:

$$\int \frac{1}{\sqrt{1-x^2}}\, dx = \sin^{-1} x + C$$

The interesting point is that the integrand has nothing to do with trigonometric functions and we shall also say more about this later in the course.

But for now, this completes our reunion with trigonometry.

CHAPTER VIII
INTEGRAL CALCULUS

A. Introduction

There is a tendency to think of integral calculus as being the "inverse" of differential calculus. Such a tendency is natural in light of the fact that we introduced the terminology, the indefinite integral, as a synonym for the "inverse" of differentiation. Yet, the fact is that integral calculus was being studied by the ancient Greeks as early as 600 B.C., while differential calculus was not developed until 1680 A.D. This is why we tried to shy away from, or at least play down, the phrase "indefinite integral" in favor of "the inverse derivative." The point is that, in the truest sense, the study of the indefinite integral is not part of what is called integral calculus but rather a part of differential calculus. For pedagogical reasons, one usually starts with the study of differential calculus and then proceeds to the study of integral calculus, whereas in "real life" the order was the reverse. This inversion tends to blur the true meaning of integral calculus, which has an existence in its own right, completely independent of differentialcalculus. It is this existence together with its relation to differential calculus that we wish to develop in this chapter.

Our procedure shall be to begin by recounting the ancient Greek contribution to the subject which came in the form of the study of areas of non-rectilinear figures (figures whose boundaries consisted of other than straight line segments) and to generalize the subject so that it becomes independent of area. In much the same way, we earlier found it geometrically convenient to visualize the derivative as the slope of a curve, while recognizing that the general definition of derivative transcends its restriction to slopes of curves.

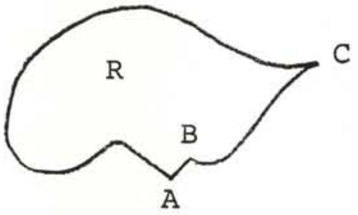
In our development, we shall pay homage to the proper chronology by assuming throughout the initial phases of our presentation that the concept of differential calculus does not exist. This will give us still another way of rediscovering the limit concept and, in fact, will afford us a completely new approach to calculus. Once this mission is completed and we have given the subject of integral calculus proper recognition in its own right, we shall develop the rather remarkable relationship which exists between the two apparently disparate branches of calculus known as differential and integral calculus - a relationship that in terms of hindsight offers a most interesting justification for such terminology as the "indefinite integral" instead of "inverse differentiation."

We shall make no further attempt here to explain this intriguing relationship. Rather, we shall now begin to unfurl our story, one segment at a time, beginning with a "revisit" to the study area.

B. A Discussion of Areas of Non-Rectilinear Figures

As is well-known, the ancient Greeks, using some rather elementary results from plane geometry, were able to determine the area of any rectilinear figure. What is not quite so well known is how they proceeded from this to find the areas of more general regions. It is this procedure that we shall study in this section.

We shall assume that our region is enclosed by a fairly arbitrary curve. For example, our region R might be as depicted in Figure 1. Notice that we do not insist that our curve be smooth but we have assumed that the curve is continuous.*



Boundary is not "smooth" at points A, B, and C.

(Figure 1)

We shall make some more specific remarks about the form of our region R as we go along, but for now let us be content to emphasize the logic of how we proceed from our knowledge of rectilinear regions to obtain some knowledge about the less familiar case of non-rectilinear regions. To begin with, we must, as is in the case of all logical systems, agree on certain properties of area that we are all willing to accept.

*It is interesting to note that for finding areas we do not require that the boundary curve be differentiable (smooth); all we need is that the curve is unbroken (continuous) (and even continuity can be waived in some cases as we shall see later). For example, among other things, we can find the area of rectangles and triangles, and clearly these figures have boundaries which contain "corners" (that is, points at which the curve is not smooth). This, then, is one very important reason for stressing continuous curves. Namely for finding areas, we do not have to insist on curves being anything more than continuous. In this sense, finding areas is more general (less restricted) than finding instantaneous rate of change, where by definition we require differentiability.

To begin our logical quest, let us observe that there is one region for which we could easily define area in an objective way; acceptable to all: namely,

Rule 1:

> The area of a rectangle is the product of the length of its base and the length of its height. In symbols, $A = bh$.

We next observe that it seems non-controversial enough* to assume that as a region gets larger so does its area. (Intuitively, this is precisely the definition of "gets larger".) At any rate, stated more formally,

Rule 2:

> If A_R denotes the area of region R and if A_S denotes the area of region S, then if $R \subset S$, $A_R \leqslant A_S$. In words, if one region is contained within another, the area of the "containing" region is at least as great as that of the "contained" region.

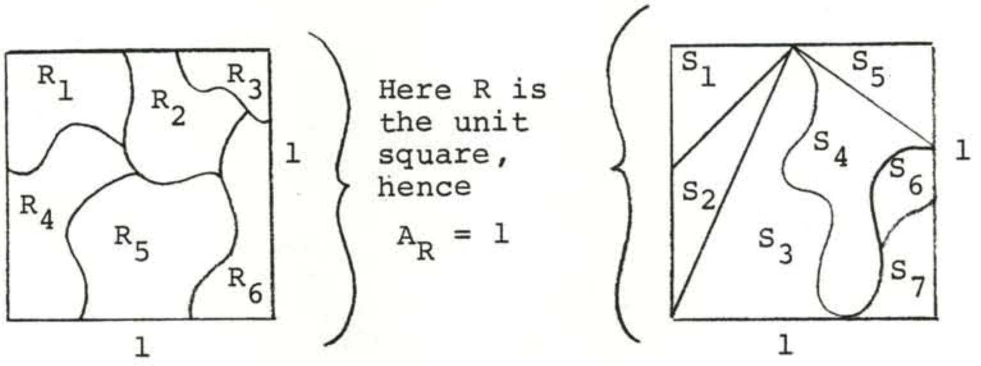
Finally, we invoke the well-known result of elementary geometry that "the whole equals the sum of its parts." Again, more precisely,

Rule 3:

> If a region is the union of regions that share no points in common, other than possibly boundary points, then the area of the region is the sum of the areas of the constituent parts.

*Technically speaking, the notion of "non-controversial" or "self-evident" is highly subjective. What we are really leading up to is that without assumptions there can be no proof. Thus in any logical system we must invoke certain "rules of the game" from which we logically derive the rest of the properties we desire. The idea is that since we want our derived results to be realistic we start with rules of the game which seem realistic, and this is why we mention the idea of "non-controversial rules."

Before proceeding further, it is probably a valuable aside to observe that Rule 3 is really "loaded." For, while it may seem rather obvious, notice that it gives us a very powerful property of area. If we start with a particular region there are infinitely many ways of partitioning it into the type of union mentioned in Rule 3. Rule 3 says that no matter how we perform the partitioning, area must be defined in such a way that the sum of the areas of the pieces is independent of the partitioning! For example:



Here R is the unit square, hence $A_R = 1$

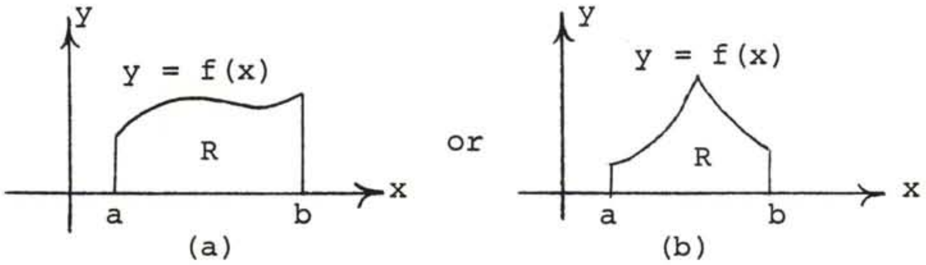
(Figure 2)

Thus, no matter how area is defined, to satisfy Rule 3, we must have: $A_{R_1} + A_{R_2} + A_{R_3} + A_{R_4} + A_{R_5} + A_{R_6} = A_{S_1} + A_{S_2} + A_{S_3} + A_{S_4} + A_{S_5} + A_{S_6} + A_{S_7}$ (=1). Or in more compact notation

$$\sum_{K=1}^{6} A_{R_K} = \sum_{K=1}^{7} A_{S_K}$$

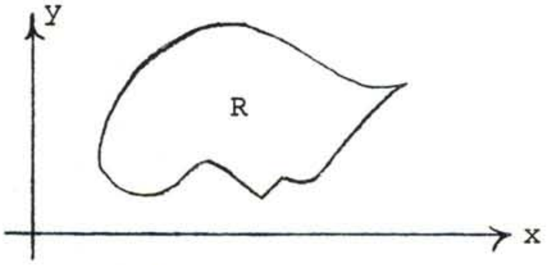
Now, in the event that one is still not convinced of the power of our three rules, let us show that these three rules, together with a few other mathematical concepts which we have already discussed either implicitly or explicitly, are enough to determine the solution to our problem of generalizing the concept of area.

For example, we can now show that we can restrict our study of area to a rather simple special case without any loss of generality. Namely, let our region R be bounded below by the x-axis, above by the curve y = f(x) where f is continuous and non-negative, (that is $f(x) \geqslant 0$ for all x) on the left by the line x = a, and on the right by the line x = b. That is,



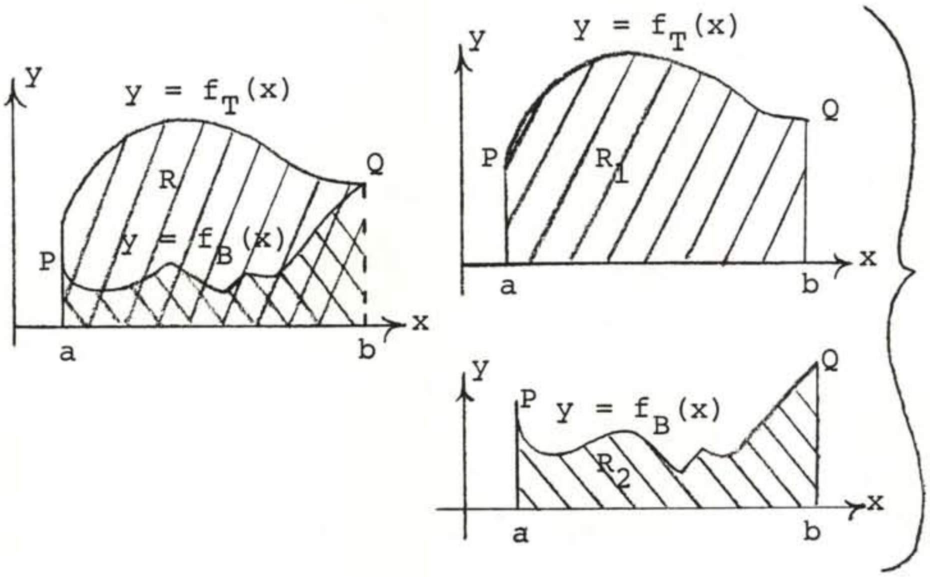
(Figure 3)

You see, the idea is that if we are now given a region such as, say, in Figure 1, we can "superimpose" it onto our coordinate plane so that it never goes below the x-axis. Thus,



(Figure 4)

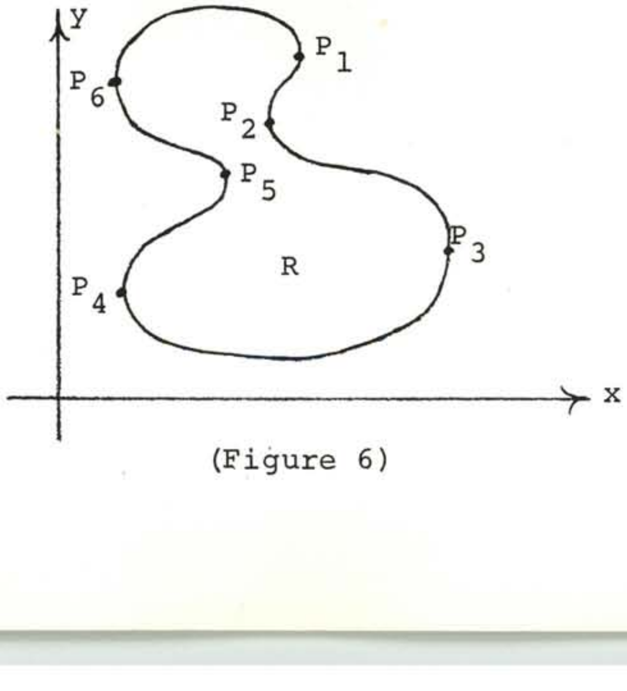
We then locate the points at which the curve "doubles back." Where the curve is smooth, this will occur when we have a vertical tangent, that is, when dx/dy = 0. But, since there is no guarantee that our curve is smooth, we must also check those points where there are "sharp" edges. We can look at the boundary as a union of single-valued curves. In our particular example, we have



By Rule 3, since $R_1 = R \cup R_2$ and R and R_2 share only the boundary $y = f_B(x)$, $A_{R_1} = A_R + A_{R_2}$. $\therefore A_R = A_{R_1} - A_{R_2}$ and R_1 and R_2 are of the type discussed in Figure 3.

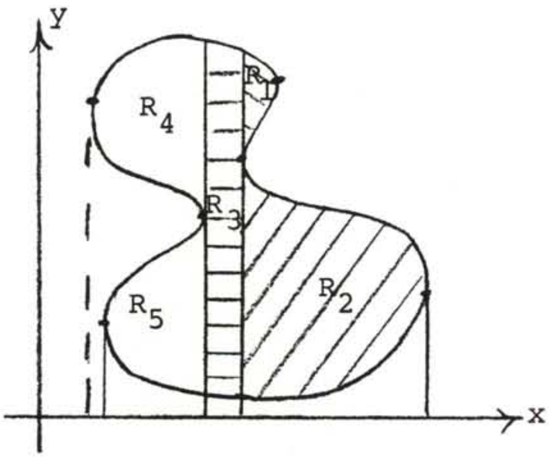
(Figure 5)

Of course as our boundary becomes more "frilly" we must become a bit more careful in keeping track of things, but basically, things remain the same. For example,



(Figure 6)

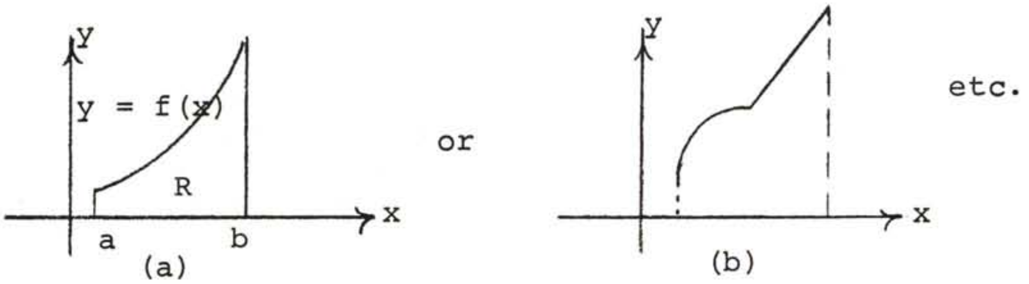
We subdivide R "advantageously" (and there are usually several options at our disposal) so that we get a union of regions of the type we can already handle.



(Figure 7)

It is not our purpose here to explore the various cases which might occur, but rather to indicate how Rule 3 allows us to focus our complete attention on what seems to be an over-simplified special case: namely, the situation depicted in Figure 3.

In fact, our three rules give us even more freedom in undertaking our investigation. For example, we may assume that on the interval [a,b], f(x) is an increasing function. (Why we might want to make such an assumption will, hopefully, become clear later.) That is, we may assume that our region has the form:

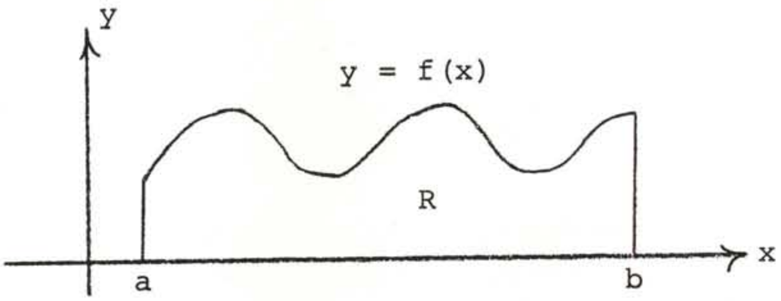


(a)

(b)

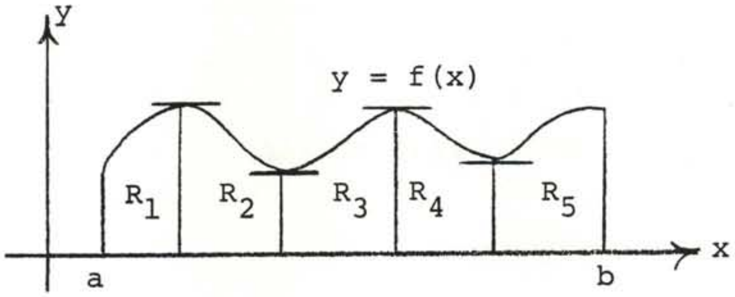
(Figure 8)

Rather than justify this last assumption from an abstract point of view, let us use a specific example. Suppose our region R is as in Figure 9.



(Figure 9)

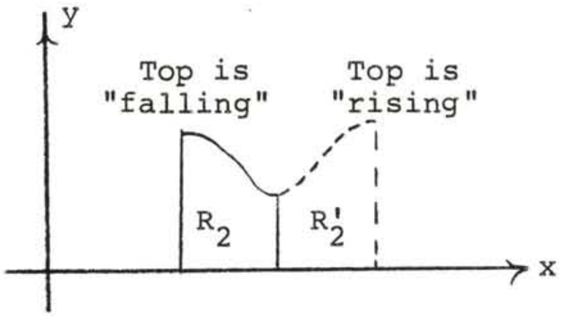
We may partition C into a union of curves each of which is either always rising or always falling. Thus,



(Figure 9a)

In this way, R is partitioned into a union of regions R_1, R_2, R_3, R_4, and R_5 which share only boundary points in common. Hence, we may find the area of R simply by considering regions in which the "top" is either an always rising or an always falling curve. In our example, R_1, R_3, and R_5 have "tops" which are rising, but R_4 and R_2 have "tops" which are falling. However, if we "flip" R_4 or R_2 over we get a new region, say R_4' or R_2' whose top is always rising.

Thus,



(Figure 10)

Since $R_2 \overset{\sim}{=} R_2'$, $A_{R_2} = A_{R_2'}$*

Thus we may replace R_2 by R_2' without loss of generality in finding A_R.

In summary, then, we may study area in general by focusing our attention on the special case in which our region R is bounded above by the continuous, always rising, curve $y = f(x)$, below by the x-axis, on the left by the line $x = a$, and on the right by the line $x = b$.

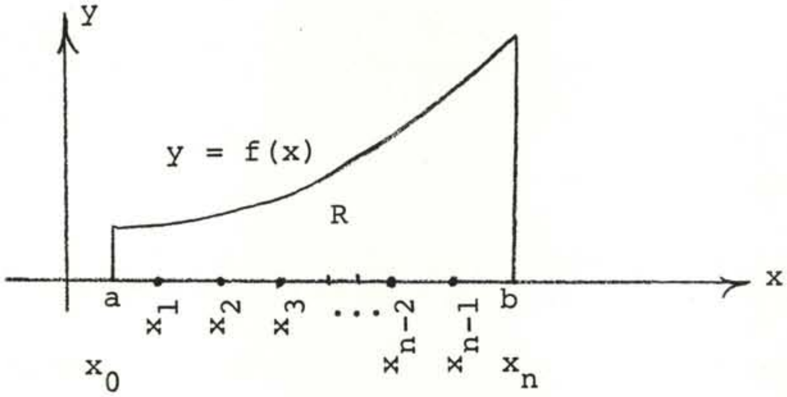
The problem that now remains, however, is that of finding the area of such a region. Our technique will be what the ancient Greeks called the "method of exhaustion" wherein we "squeeze" the desired region between rectilinear regions whose areas we can determine, and, in this way, we can obtain upper and lower bounds for the area of the desired region.

More specifically, we proceed as follows:

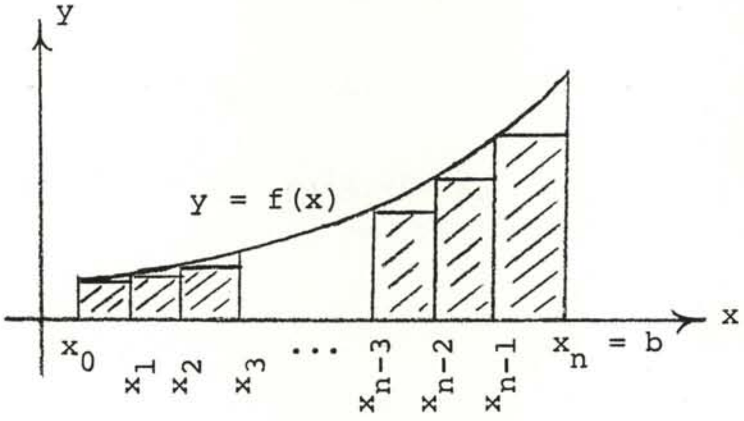
Partition [a,b] into n equal parts, each of size $\Delta x = \frac{b-a}{n}$ and label the points of subdivision $a = x_0, x_1, x_2, \ldots, x_n = b$. Next inscribe rectangles on

*While it is probably "self-evident" that congruent regions have equal areas, the fact is that we may deduce this result from Rule 2. Namely $R \overset{\sim}{=} S$ means that if superimposed properly R and S "coincide." More formally, $R \subset S$ and $S \subset R$. But $R \subset S$ implies that $A_R \leqslant A_S$, while $S \subset R$ implies that $A_S \leqslant A_R$, but, since A_R and A_S are numbers, $A_R \leqslant A_S$ and $A_R \geqslant A_S$ together imply that $A_R = A_S$.

each partition of heights $f(x_0)$, $f(x_1)$, ..,$f(x_{n-1})$.* That is:



(Figure 11a)



(Figure 11b)

*The idea is that in each $[x_{k-1},x_k]$ we pick that number c_k for which $f(c_k) \leqslant f(x)$ for all $x \varepsilon [x_{k-1},x_k]$. Such a c_k always exists by virtue of the fact that f is continuous. In our special case in which $y = f(x)$ is always rising, we have the computational simplification that $c_k = x_{k-1}$, that is the lowest point on each subinterval occurs at the left end-point.

Now the "shaded" region is, by construction, contained in R; hence, its area is no greater than that of R. Let us denote the area of the shaded region by L_n (L to suggest that we have a lower bound on A_R and n to suggest that the area of the shaded region depends on the number of rectangles (n) we inscribe).

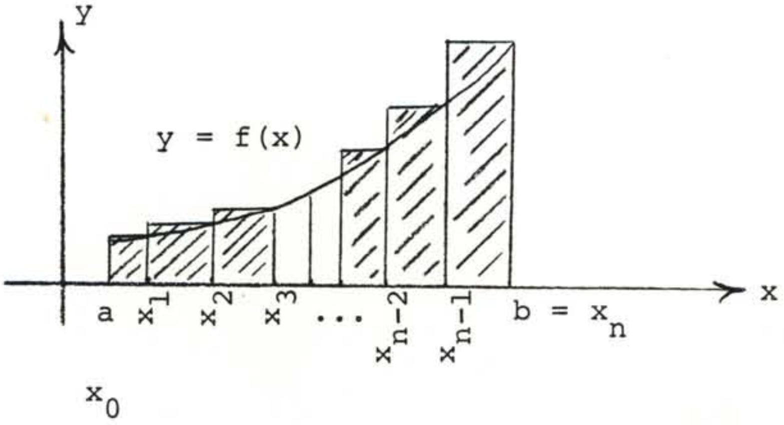
In any event, for all n,

$$L_n \leqslant A_R \tag{1}$$

Moreover, our shaded region is the union of rectangles, each with base Δx and heights of $f(x_0)$, $f(x_1)$, ..., $f(x_{n-1})$. Since we know how to find the area of a rectangle and since the area of a region is the sum of the areas of its constituent parts, we may conclude that

$$L_n = f(x_0)\Delta x + f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_{n-1})\Delta x \tag{2}$$

In a similar way, we may pick the highest point of each partition and circumscribe rectangles which enclose R. That is,



(Figure 11c)

If we let U_n (U representing Upper Bound) denote the area of the region shaded in Figure 11c, we have

$$A_R \leqslant U_n \tag{3}$$

and

$$U_n = f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_{n-1})\Delta x + f(x_n)\Delta x \tag{4}$$

If we combine (1) and (3) we see that for <u>every positive integer</u>, n:

$$L_n \leqslant A_R \leqslant U_n \tag{5}$$

Moreover, from (2) and (4) we see that

$$\begin{aligned} U_n - L_n &= f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_{n-1})\Delta x + f(x_n)\Delta x \\ &\quad - [f(x_0)\Delta x + f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_{n-1})\Delta x] \\ &= f(x_n)\Delta x - f(x_0)\Delta x \\ &= [f(x_n) - f(x_0)]\Delta x \end{aligned}$$

or since $x_n = b$, $x_0 = a$, and $\Delta x = \frac{b-a}{n}$, we may write

$$= [f(b) - f(a)] \left[\frac{b-a}{n}\right] \tag{6}$$

Now, since f(a), f(b), a, and b are constants so also is

$[f(b) - f(a)]\ [b-a]$. Let $c = [f(b) - f(a)]\ (b-a)$.

Then (6) becomes:

$$U_n - L_n = \frac{c}{n} \tag{7}$$

Since $\frac{c}{n} \to 0$ as $n \to \infty$, we sense* that

$$\lim_{n\to\infty}(U_n - L_n) = \lim_{n\to\infty}\frac{c}{n} = 0 \tag{8}$$

Again, assuming** that $\lim_{n\to\infty}(U_n - L_n) = \lim_{n\to\infty}U_n - \lim_{n\to\infty}L_n$, equation (8) yields

$$\lim_{n\to\infty}L_n = \lim_{n\to\infty}U_n \tag{9}$$

Since A_R is a fixed number and is always "caught between" L_n and U_n, it would appear from (9) that

$$\lim_{n\to\infty}L_n \leqslant A_R \leqslant \lim_{n\to\infty}U_n\text{, and, therefore:}$$

*We say "sense" since, technically speaking, we have not studied limits of discrete sequences. That is, when we wrote $\lim_{x\to c} f(x)$ it was assumed that x denoted any real number in a deleted neighborhood of c. When we write $\lim_{n\to\infty}U_n$ n is not any real number but rather any positive integer. Limits of discrete sequences will be discussed in Block VII more rigorously. For now we rely on our intuition and previous experience concerning limits.

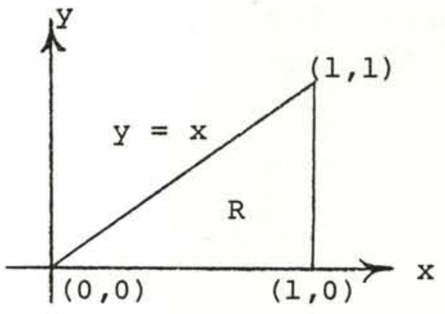
**Just once more for emphasis, when we proved that the limit of a sum (difference) was the sum of the limits, we used the fact that our "inputs" were connected sets of real numbers (e.g. dom f = [a,b]). Here, our inputs are restricted to positive integers and we have not "officially" proven such theorems in this case.

$$\boxed{A_R = \lim_{n\to\infty} L_n = \lim_{n\to\infty} U_n} \tag{10}$$

Equation (10) not only gives us the "recipe" for computing A_R, but it also shows us that, before proceeding further, we must come to grips with the notion of $\lim_{n\to\infty} a_n$. Thus it seems that replacing our study of instantaneous rate of change by a study of area has in no way helped as avoid the concept of limit. At best we have replaced the limit of a "continuous" variable by the limit of a "discrete" variable.

At any rate, lest our present presentation seem too abstract, let us illustrate our results in terms of a concrete, familiar example.

Let us consider the region R where



(Figure 12)

Clearly, since R is a triangle for which b = h = 1, $A_R = \frac{1}{2}$. (We pick such an example so that we may have a "standard" by which we may compare the new method of this section. For example, if we picked a region whose area was not known to us, how could we check the validity of the answer which was obtained by the new method?)

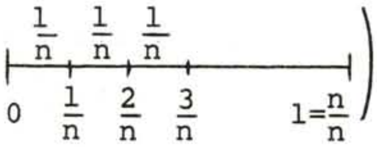
In any event, from (4), we recall that

$$U_n = f(x_1)\Delta x + \ldots + f(x_n)\Delta x.$$

In our particular example $a = 0$, $b = 1$, $f(x) = x$. Hence, $\Delta x = \frac{b-a}{n} = \frac{1}{n}$. Thus, in our example,

$$U_n = \frac{x_1}{n} + \ldots + \frac{x_n}{n} = \frac{x_1 + \ldots + x_n}{n}$$

and $x_1 = \frac{1}{n}$, $x_2 = \frac{2}{n}$, $x_3 = \frac{3}{n}$, etc. (that is



)

$$\therefore\ U_n = \frac{\frac{1}{n} + \frac{2}{n} + \ldots + \frac{n}{n}}{n}$$

$$= \frac{1 + 2 + \ldots + n}{n^2} \qquad (11)$$

Now, in our treatment of Mathematical Induction we have already seen that $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$. Putting this into (11) we obtain

$$U_n = \frac{n(n+1)}{2n^2} = \frac{n+1}{2n} = \frac{1}{2}\left(1 + \frac{1}{n}\right) \qquad (12)$$

whence $\lim_{n\to\infty} U_n = \frac{1}{2}$, whence by (10) $A_R = \frac{1}{2}$ which checks with our previous result.

While we omit the details, it can easily be shown that L_n in this case is given by

$$L_n = \frac{1}{2}\left(1 - \frac{1}{n}\right) \tag{13}$$

whence

$$\lim_{n\to\infty} L_n = \frac{1}{2}$$

as it should be, according to (10).

For those of us who may still "mistrust" limits, notice that equations (12) and (13) are derived without any references to limits. In other words we may conclude from (12) and (13) that for each n

$$\frac{1}{2}\left(1 - \frac{1}{n}\right) < A_R < \frac{1}{2}\left(1 + \frac{1}{n}\right) \tag{14}$$

For example when $n = 10^{12}$ (a large but certainly finite number), equation (14) yields

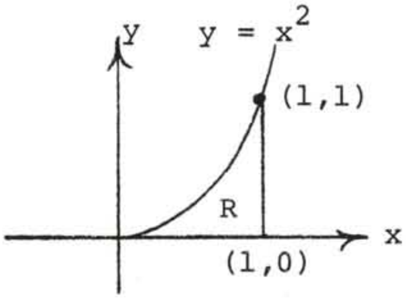
$$\frac{1}{2}\left(1 - 10^{-12}\right) < A_R < \frac{1}{2}\left(1 + 10^{-12}\right)$$

or

$$\frac{1}{2}\,(0.999999999999) < A_R < \frac{1}{2}\,(1.000000000001)$$

and we "get the feeling" that $A_R = \frac{1}{2}$.

As a final example in this section, let us modify our region R by "just a little." Namely, now let R be the region



In this case, $a = 0$, $b = 1$, $\Delta x = \frac{1}{n}$, $f(x) = x^2$, $x_1 = \frac{1}{n}$, $x_2 = \frac{2}{n}$, ..., $x_n = \frac{n}{n}$. Hence:

$$\begin{aligned} U_n &= f(x_1)\Delta x + \ldots + f(x_n)\Delta x \\ &= [f(\tfrac{1}{n})]\ \tfrac{1}{n} + \ldots + f(\tfrac{n}{n})\ \tfrac{1}{n} \\ &= (\tfrac{1}{n})^2\ (\tfrac{1}{n}) + \ldots + (\tfrac{n}{n})^2\ (\tfrac{1}{n}) \\ &= \frac{1^2 + \ldots + n^2}{n^3} \end{aligned} \tag{15}$$

Recalling (see exercises for another way of obtaining this result) that $1^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$, (15) becomes:

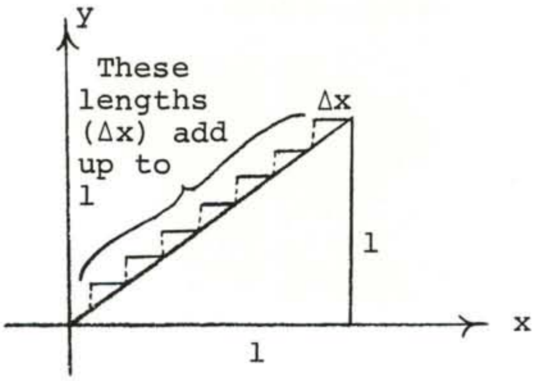
$$\begin{aligned} U_n &= \frac{n(n+1)(2n+1)}{6n^3} \\ &= \frac{1}{6}(\frac{n+1}{n})\ (\frac{2n+1}{n}) \\ &= \frac{1}{6}(1 + \frac{1}{n})\ (2 + \frac{1}{n}) \end{aligned}$$

$$\therefore A_R = \lim_{n\to\infty} U_n = \frac{1}{6}(1 + 0)(2 + 0) = \frac{1}{3} \tag{16}$$

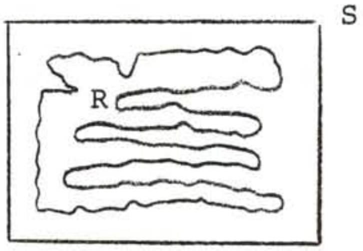
Notice in this example that once we had the proper expression for $\sum_{k=1}^{n} k^2$ it was not any more difficult to find A_R than in the previous example. Yet the "easy" method for the first problem (area of a triangle) cannot be adapted to the new problem.

This completes our introduction to integral calculus except for a few asides:

(1) It was important to find both L_n and U_n in order to "sandwich" A_R. This was the only way we could be sure that all the error was "squeezed out" when we went to the limit. For example, if we return to the region R of our first example and try to use the same technique for finding the length of the hypotenuse of the triangle we could say that $\Delta s \approx \Delta x$ for small Δx. Yet the Δx's add up to 1 while the Δs's add up to $\sqrt{2}$. That is



In essence $R \subset S \rightarrow A_R \leq A_S$ does not apply when we replace area by length (perimeter). For example



$R \subset S$ yet $P_R > P_S$ (where P denotes perimeter)

The point is that we need

$$\begin{tikzcd}[column sep=small] L_n \arrow[dr] & < A_R < & U_n \arrow[dl] \\ & \lim_{n\to\infty} U_n = \lim_{n\to\infty} L_n & \end{tikzcd}$$

to guarantee that we have located A_R exactly.

(2) We could have picked any sum between U_n and L_n, say S_n and defined $A_R = \lim_{n\to\infty} S_n$, where $S_n = f(c_1)\Delta x + \ldots + f(c_n)\Delta x$ where c_k is any number (point) in $[x_{k-1}, x_k]$.

Namely, if

$$\begin{gathered} x_0 \leqslant c_1 \leqslant x_1 \\ \vdots \\ x_{n-1} \leqslant c_n \leqslant x_n \end{gathered}$$

Then

$$f(x_0)+\ldots+f(x_{n-1}) \leqslant f(c_1)+\ldots+f(c_n) \leqslant f(x_1)$$
$$+\ldots+f(x_n)^*$$

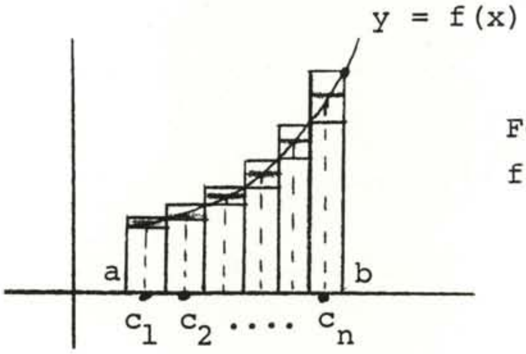
$$\therefore\ [f(x_0)+\ldots+f(x_{n-1})]\Delta x \leqslant [f(c_1)+\ldots+f(c_n)]\Delta x$$
$$\leqslant [f(x_1)+\ldots+f(x_n)]\Delta x$$

or

$$L_n \leqslant S_n \leqslant U_n$$

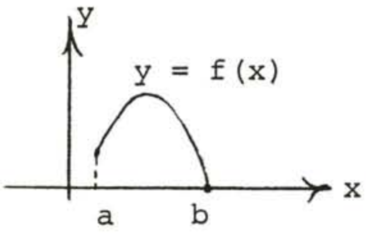
$\therefore \lim\limits_{n\to\infty} L_n = \lim\limits_{n\to\infty} U_n = A_R$ guarantees by the "sandwiching principle" that $\lim\limits_{n\to\infty} S_n = \lim\limits_{n\to\infty} L_n = \lim\limits_{n\to\infty} U_n = A_R$.

Pictorially,



For each k
$f(x_{k-1})\Delta x \leqslant f(c_k)\Delta x \leqslant f(x_k)\Delta x$

*Here we are invoking the fact that f is an increasing function. Thus, for example, $x_0 \leqslant c_1 \leqslant x_1 \rightarrow f(x_0) \leqslant f(c_1) \leqslant f(x_1)$. Obviously if f is not an increasing function this need not be true. Example:



$$\begin{cases} a < b \\ f(a) > f(b) \end{cases}$$

(3) It is not important that [a,b] be divided into n equal parts. What is important is that the size of the biggest interval approaches 0 as $n \to \infty$. That is, suppose we make any partition of [a,b] into n parts not necessarily equal. Say

$$a = x_0 < x_1 < \ldots < x_n = b$$

and let $\Delta x_k = [x_{k-1}, x_k]$.

Let

$$\overline{U}_n = f(x_1)\Delta x_1 + \ldots + f(x_n)\Delta x_n$$

$$\overline{L}_n = f(x_0)\Delta x_1 + \ldots + f(x_{n-1})\Delta x_{n-1}$$

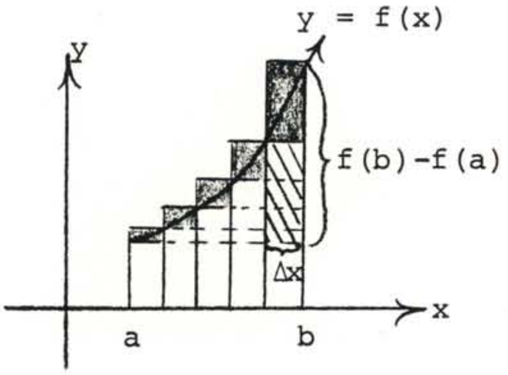
Then

$$\overline{U}_n - \overline{L}_n = f(x_n)\Delta x_n - f(x_0)\Delta x_1$$

$$= f(b)\Delta x_n - f(a)\Delta x_1$$

$$\therefore \lim_{n \to \infty}(\overline{U}_n - \overline{L}_n) = 0, \text{ provided each } \Delta x \to 0.$$

Pictorially,

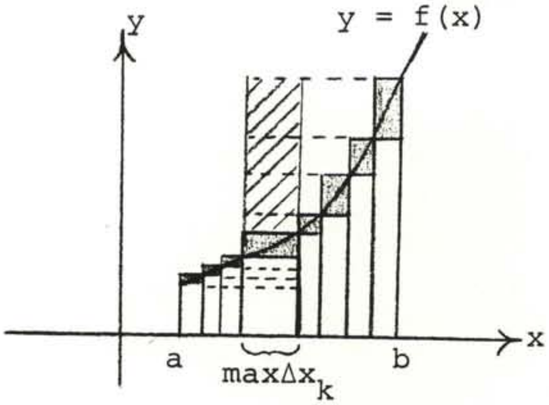


Solid region represents $U_n - L_n$. These pieces can be stacked to form a rectangle of dimensions

$[f(b)-f(a)]$ by $\frac{b-a}{n} = \Delta x$

$\therefore U_n - L_n = [f(b)-f(a)] \left(\frac{b-a}{n}\right)$

(Case 1: All Δx's are equal)



Solid region is $\overline{U}_n - \overline{L}_n$. We may slide these onto the largest interval to form a rectangle $[f(b)-f(a)]$ by $\max\Delta x_k$ whose area exceeds $\overline{U}_n - \overline{L}_n$.

$\therefore \max\Delta x_k \rightarrow 0 \rightarrow \lim_{n\to\infty}(\overline{U}_n - \overline{L}_n) = 0$

(Case 2: Δx's are not all equal)

In this case $\overline{L}_n \leqslant A_R \leqslant \overline{U}_n$.

$$\therefore \lim_{n\to\infty}\overline{L}_n = \lim_{n\to\infty}\overline{U}_n \rightarrow A_R = \lim_{n\to\infty}\overline{U}_n = \lim_{n\to\infty}\overline{L}_n$$

<u>Note</u>: From an abstract point of view this result is not obvious. Namely L_n and $\overline{L}_n$ are different since they denote areas of different regions. How then can we be sure that $\lim_{n\to\infty} L_n = \lim_{n\to\infty}\overline{L}_n$? The abstract proof is fairly sophisticated and requires the concept of uniform continuity (this is developed in the text). From an intuitive point of view, however, A_R is well-defined. Hence if $A_R = \lim_{n\to\infty} L_n$ and at the same time $A_R = \lim_{n\to\infty}\overline{L}_n$ then $\lim_{n\to\infty}\overline{L}_n = \lim_{n\to\infty} L_n$.

(4) While area gives us a nice geometric model, notice that we can define $\lim_{n\to\infty} U_n$ etc. without reference to it. For example let f be continuous, dom f = [a,b], $x_1 < x_2 \rightarrow f(x_1) < f(x_2)$. Then we may form the partition

$$a = x_0 < x_1 < \ldots < x_n = b,$$

choose $c_k \varepsilon [x_{k-1}, x_k]$, and form:

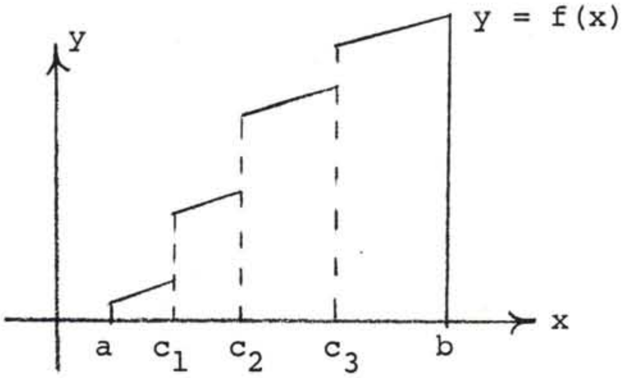
$$f(c_k)\Delta x_1 + \ldots + f(c_n)\Delta x_n$$

and look at

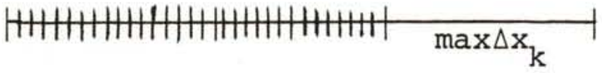
$$\lim_{\max\Delta x_k \to 0^*} [f(c_1)\Delta x_1 + \ldots + f(c_n)\Delta x_n]$$

Of course, such limits may be extraordinarily difficult to compute, and it may also be difficult, devoid of geometric interpretation, to prove that such limits are independent of the method of partitioning, but despite these difficulties the concept of such a limit exists in its own right apart from the interpretation as an area.

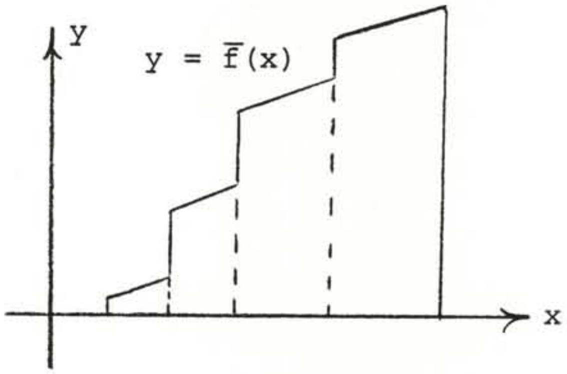
(5) Even the condition that f is continuous on [a,b] can be "weakened." For example, suppose f has a finite number of "jump" discontinuities on [a,b] (in which case f is called <u>piecewise-continuous</u> on [a,b]). Pictorially,



*Notice that $\max\Delta x_k \to 0$ says more than $n \to \infty$. For example, we can leave one interval intact and subdivide the others so that $n \to \infty$ yet $\max\Delta x_k$ doesn't approach 0.

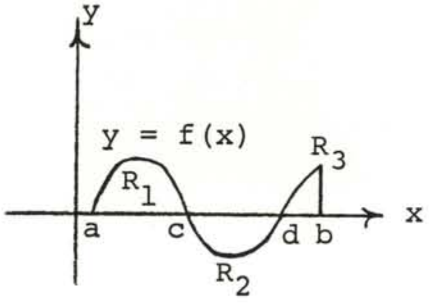


Technically speaking, such a region, not being "enclosed," has no area. However, since a line having no thickness has no area, we could "close up" the region and still talk about area. That is



To be sure $y = \overline{f}(x)$ is not single valued at the "jumps" of $y = f(x)$ but while $y = \overline{f}(x)$ and $y = f(x)$ are different curves they enclose the same area.

(6) If we think about the concept of net area rather than total area we may even remove the restriction that $f(x)$ be non-negative. For example, if



Then $f(c_k) < 0$ for any $c_k \in [c,d]$. Thus for this part of the partition $\sum f(c_k)\Delta x_k$ is negative. More specifically, the idea is that in this case if we form $\lim_{\max\Delta x_k \to 0} \sum_{k=1}^{n}$ $f(c_k)\Delta x_k$ we find $A_{R_1} + A_{R_3} - A_{R_2}$.

In summary, then, the ancient Greeks may be viewed as the fathers of integral calculus in that they were the first to come to grips with an expression of the form:

$$\lim_{n\to\infty} \sum_{k=1}^{n} f(c_k)\Delta x$$

To be sure, their usage of such limits were restricted to geometry in the quest for areas, volumes, and arc lengths. In the more general sense, such sums can represent more "practical" physical concepts such as distance, velocity, work, etc. These ideas are developed in the text in great detail, and we shall treat them in due course. For now, it is our purpose only to emphasize that the concept of integral calculus as the limit of an infinite sum requires no a priori knowledge of differential calculus.

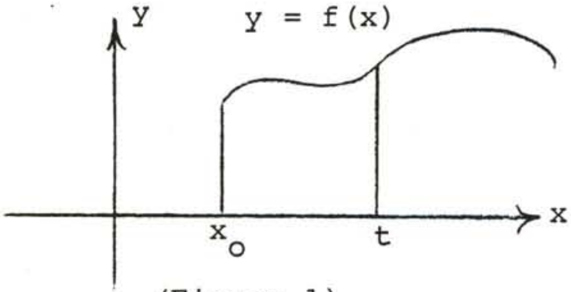
With the hope that this point is very clear, we devote the remainder of this chapter toward showing the wonderful relation between integral calculus and differential calculus.

C. Area as a Differential Equation

While it may not be apparent, at least at first glance, why one might have been motivated to "invent" integral calculus as a sequel to differential calculus (had differential calculus been invented first), it is not difficult to imagine how one might have been motivated to "invent" differential calculus as a natural consequence of integral calculus (although this did not happen in the true course of events.)

For example, let us again imagine that the function f is continuous and non-negative, and merely for the sake of convenience we shall assume that the domain of f is the set of all real numbers. We consider the region R which is bounded above by the curve y = f(x), below by the x-axis, on the left by the line $x = x_o$, where x_o is a specifically chosen constant for our investigation (had the domain of f been [a,b] then x_o would have been restricted to this interval but otherwise nothing different would occur), and on the right by the line x = t. It is clear that the size of R depends on t and hence that the area of R, A_R, is, therefore, a function of t. Thus, it is not at all far-fetched to ask how the area changes as t changes. Obviously, the change is negative if t decreases and positive if t increases. For the sake of simplicity, however, we shall investigate how the area increases as t increases.

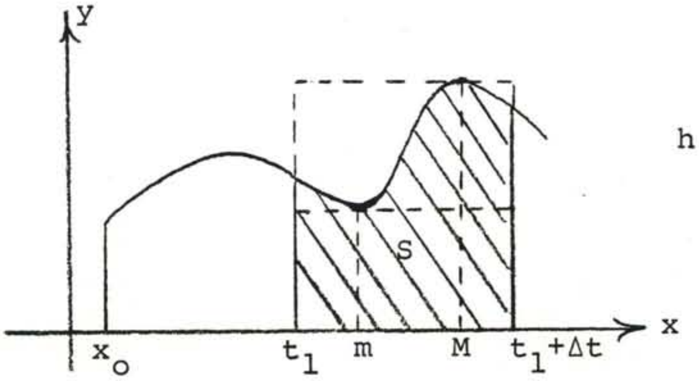
Before continuing, let us pause for a moment and observe that if the non-negative restriction on f is removed, our investigation remains the same, provided only that we replace "area" by "<u>net</u> area" in accordance with our remarks of the previous section. Notice also, as we discussed in the previous section, that the continuity requirement on f can be weakened to the extent that we need only assume that f is piecewise-continuous. At any rate, we have:



(Figure 1)

Let us denote the area of R by A(t) rather than by A_R in order to emphasize that the area is indeed a function of t. We could then have discussed (as we did in our treatment of differential calculus) the change in A as t changed from t_1 to $t_1 + h$, denoted by $A(t_1 + h) - A(t_1)$, from which we could have proceeded to a discussion of the average rate of change $\frac{A(t_1 + h) - A(t_1)}{h}$, and finally to the limit of this quotient as h approached zero. In this way, we would have "invented" the concept of A'(t); and it is in this sense that differential calculus could have been a sequel to what was called integral calculus.

Now,



$h = \Delta t$

(Figure 2)

$A(t_1+\Delta t) - A(t_1)$ denotes the area of the region S which we have shaded in Figure 2. Let m denote the value of $x \in [t_1, t_1+\Delta t]$ for which f(x) is minimum. (Locating m is trivial pictorially; from the more abstract point of view, the existence of m is guaranteed by the fact that a continuous function on a closed interval assumes its minimum and maximum values on that interval.) Then the rectangle of area $f(m)\Delta t$ is inscribed in S. Hence:

$$f(m)\Delta t \leqslant A_S \qquad (1)$$

Similarly, if M denotes the value of $x \in [t_1, t_1+\Delta t]$ at which f(x) is maximum then

$$f(M)\Delta t \geqslant A_S \tag{2}$$

Recalling that $A_S = \Delta A = A(t_1+\Delta t) - A(t_1)$ we may combine (1) and (2) to conclude

$$f(m)\Delta t \leqslant A(t_1+\Delta t) - A(t_1) \leqslant f(M)\Delta t \qquad (m,\ M \in [t_1,\ t_1+\Delta t]) \tag{3}$$

Dividing through by Δt in (3) we obtain

$$f(m) \leqslant \frac{A(t_1+\Delta t) - A(t_1)}{\Delta t} \leqslant f(M)^* \tag{4}$$

Letting $\Delta t \to 0$ in (4), the fact that m and M are in $[t_1, t_1+\Delta t]$ guarantees that both m and $M \to t_1$. Now, however, since f is continuous, we have that $f(m) \to f(t_1)$ as $\Delta t \to 0$ and $f(M) \to f(t_1)$ as $\Delta t \to 0$**. Thus if we let $\Delta t \to 0$, (4) becomes

$$f(t_1) \leqslant \lim_{\Delta t \to 0} \frac{A(t_1+\Delta t) - A(t_1)}{\Delta t} \leqslant f(t_1) \tag{5}$$

*Notice that we are assuming that $\Delta t > 0$, for if Δt were negative the sense of the inequality would be changed. That is if $a < b < c$ then $ta > tb > tc$ if t is negative. In effect, when we are finished we will <u>not</u> have found $\lim_{\Delta t \to 0} \frac{\Delta A}{\Delta t}$ but rather $\lim_{\Delta t \to 0^+} \frac{\Delta A}{\Delta t}$. A result similar to (4), however, can be easily obtained when $\Delta t < 0$ so that our final stated result will be true.

**Notice that continuity is indeed required here. We are saying, essentially, that $\lim_{m \to t_1} f(m) = f(t_1)$, and this is precisely the <u>definition</u> that f is continuous at t_1.

Hence:

$$\lim_{\Delta t \to 0} \left[\frac{A(t_1+\Delta t) - A(t_1)}{\Delta t}\right] = f(t_1);$$

and therefore:

$$A'(t_1) = f(t_1) \qquad (6)$$

where $A'(t_1)$ is, of course, $\lim_{\Delta t \to 0} \frac{A(t_1+\Delta t) - A(t_1)}{\Delta t}$.

Since t_1 could have denoted any point in the domain of f, we may remove the subscript and rewrite (6) as

$$A'(t) = f(t) \qquad (7)$$

Equation (7) gives us a rather remarkable expression for A(t) as a differential equation. It also says that the rate of change of area at a given instant is numerically equal to the height of the curve above the x-axis at that instant. At least part of this result should not come as too big a surprise since it seems intuitively clear that A'(t) should at least be a direct function of f(t); what may be surprising is the simplicity of the relationship.

If we now recall our earlier treatment of the inverse of differentiation, we may see in (7) a hint of things to come. Namely, suppose we assume that G is any function for which G' = f. Then, since G' = A', it follows that:

$$A(t) = G(t) + c \qquad (8)$$

and since $A(x_0) = 0$ (since a region of zero width has zero area), we obtain from (8) that:

$$0 = A(x_o) = G(x_o) + c,$$

whence,

$$c = -G(x_o)$$

and putting this result into (8), we finally obtain:

$$A(t) = G(t) - G(x_o), \text{where } G' = f \qquad (9)$$

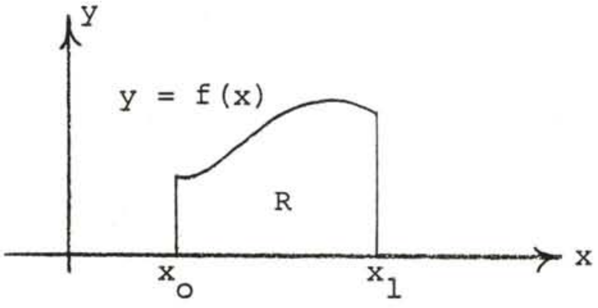
It turns out that equations (7) and (9) are the keys to the remarkable interrelation between integral and differential calculus.

We shall explore this relationship further in the next section, but for now we would like to connect these last results with the results of our previous section.

In the previous section we considered a region R bounded above by $y = f(x)$, below by the x-axis, on the left by $x = a$ and on the right by $x = b$ and we studied A_R. Notice that R in this case corresponds to our present treatment in which we let $x_o = a$ and $t = b$. That is, to relate R of the previous section to the R of this section, all we need do is observe that we may consider our area under the curve as a function of t where t takes on all values from a to b. Another way of saying this without reference to the previous section is to observe that if we now pick a value of t greater than x_o, say $t = x_1$, then

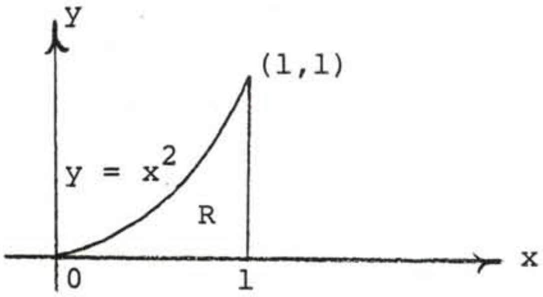
$$A_R = G(x_1) - G(x_o),\ G' = f,$$

where



(Figure 3)

As a quick check we can re-investigate A_R where



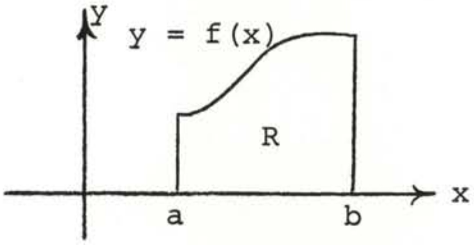
In the last section we showed that $A_R = \frac{1}{3}$. At that time we used the fact that $A_R = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{k^2}{n^3}$.

According to our newly derived result, we have that $A_R = G(1) - G(0)$ where $G'(x) = x^2$

$$= \frac{1}{3} x^3 \Big|_0^1$$

$$= \frac{1}{3}$$

It thus appears that inverse differentiation gives us a rather neat way for evaluating sums of the form $\lim\limits_{\max\Delta x_k \to 0} \sum\limits_{k=1}^{n} f(c_k)\Delta x_k$. Actually, the result of (7) is far more important than this, and we shall continue this discussion in the next section. For now, relish the luxury of being able to compute A_R by $G(b) - G(a)$, $G' = f$ where



D. The Fundamental Theorems of Integral Calculus

The fact that a fairly involved infinite sum can be computed rather simply by means of an inverse derivative is a remarkable result, at least in the sense that the concepts of infinite sum and inverse derivative are apparently quite independent. For this reason, this result is given a very special name: THE FIRST FUNDAMENTAL THEOREM OF INTEGRAL CALCULUS.

Actually, this theorem refers to a more general situation than area under a continuous non-negative curve (even though this is the only case for which we supplied the rigorous details, it is not too difficult, especially with access to the text, to supply the "missing links" for the more general case) and it states:

First Fundamental Theorem of Integral Calculus

Let f be piecewise-continuous on [a,b] and let $a = x_o < x_1 < \ldots . < x_n = b$ be any partitioning of [a,b]. Let $c_k \varepsilon [x_{k-1}, x_k]$ and let $\Delta x_k = x_k - x_{k-1}$. Then: $\lim\limits_{\max \Delta x_k \to 0} \sum\limits_{k=1}^{n} f(c_k)\Delta x_k$ exists and is unique (that is, it is invariant with respect to the partitioning). Moreover, this limit is precisely G(b) - G(a) where G' = f. Notice that in the special case in which f is non-negative on [a,b], the limit denotes the area under the graph of f, and this is the case we investigated in detail in the last section.

It might also be interesting to observe at this point that historically the symbol $\int_a^b f(x)dx$ was invented to denote $\lim\limits_{\max \Delta x_k \to 0} \sum\limits_{k=1}^{n} f(c_k)\Delta x_k$. In other words, $\int_a^b f(x)dx$, which is called the definite integral of f from a to b, was invented to denote an infinite sum not an inverse derivative. The fact that we didn't get into "trouble" in our treatment of inverse derivatives when we wrote

$$\int_a^b f(x)dx = G(x) \Big|_a^b , G' = f$$

is equivalent to the result stated in the first fundamental theorem.

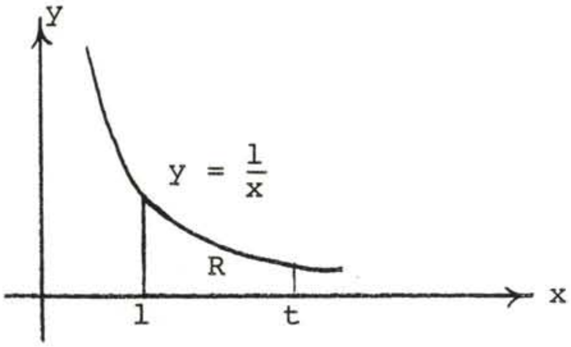
Also, in this same context, notice how $\int$ as an "elongated S" is a rather logical symbol (if symbols need be logical) for denoting an infinite sum.

Aside from the fact that the "integral sign" is better motivated in terms of an infinite sum rather than as an inverse derivative, there is yet another flaw in trying to view $\int_a^b f(x)dx$ as meaning $G(b) - G(a)$ where $G'(x) = f(x)$. The point is that as an area the <u>number</u> $\int_a^b f(x)dx$ may exist independently of whether we can exhibit a function G for which $G'(x) = f(x)$.

For example, let us consider $\int_1^t \frac{dx}{x}$ $(t \geqslant 1)$. We could certainly say that $\int_1^t \frac{dx}{x} = G(t) - G(1)$ where G is any function for which $G'(t) = \frac{1}{t}$. The problem is that, at least at the present stage of our course, we cannot explicitly produce a function G such that $G'(t) = \frac{1}{t}$*.

On the other hand, $\int_1^t \frac{dx}{x}$ is <u>exactly</u> the area of a region R where:

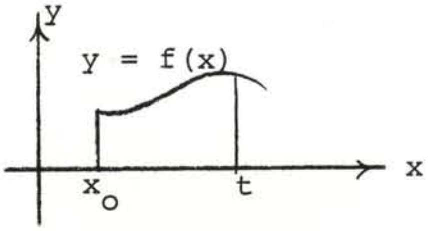
*In a "revisited" course such as ours there is the danger that we anticipate the function G for which $G'(t) = \frac{1}{t}$ since we have learned it when we first took the course. Should this be the case, replace $\int_1^t \frac{dx}{x}$ by, say, $\int_1^t e^{x^2}dx$ or $\int_1^t \frac{\sin x}{x}dx$. These latter integrals lend themselves to the same discussion as $\int_1^t \frac{dx}{x}$ but they are not integrals that were "solved" in the usual calculus course.



Certainly we are not about to argue that R has no area merely because we do not know a function G for which $G'(t) = \frac{1}{t}$!

In summary $\int_a^b f(x)dx$, where f is continuous on [a,b], is a well-defined number which can be found as a limit of a sum (or if worst comes to worst, can at least be approximated to as a great a degree of accuracy as we may require) <u>regardless</u> of whether we can exhibit a function G such that $G'(x) = f(x)$. To be sure, <u>if</u> we can exhibit G then $\int_a^b f(x)dx$ can be computed very conveniently by $G(b) - G(a)$. (Even if we can't exhibit G explicitly, it is still correct to say that $\int_a^b f(x)dx = G(b) - G(a)$, but now we are not helped at all because we dont' know more about G.)

Of even more importance, notice that in computing the area we actually explicitly produce a function whose derivative is f. Namely, (7) yields the desired result. That is, to find a function whose derivative is f(t) where f is piecewise continuous, we need only consider the region R, where



Then $A(t)$ $(= A_R) = \int_{x_o}^{t} f(x)dx$, and as we have seen from (7) $A'(t) = f(t)$.

For example, to construct a function G for which $G'(t) = \frac{1}{t}$ $(t > 1)$, we compute $\int_{x_o}^{t} \frac{dt}{t} = A(t)$.

By (7)

$$\frac{d}{dt}\left[\int_{x_o}^{t} \frac{dx}{x}\right] = \frac{1}{t}$$

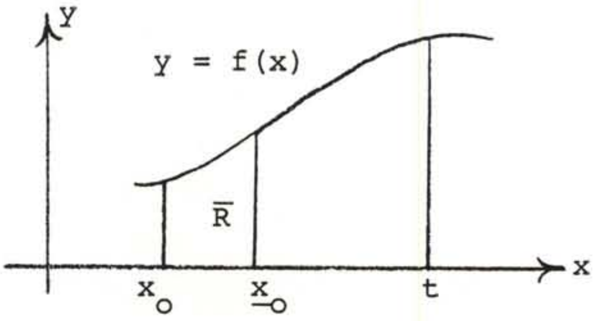
It is extremely important to notice that $A(t)$ gives us more than just the name of a function whose derivative is t^{-1}. Namely, $A(t)$ can be computed either exactly or to as close an approximation as we wish in terms of area. In other words, $A(t)$ provides us with an explicit expression for a function whose derivative is the desired function.

Let us notice also that the function $A(t)$ as we constructed it is not unique! The reason is that we chose the initial point x_o at our convenience. If we change x_o we can still talk about $A(t)$ but all we are saying is that A is actually a function of both t and x_o. How does changing x_o change the property of $A(t)$ that $A' = f$? The answer is that it doesn't. To see this more intuitively, let us use the result of the first fundamental theorem as well.

We observe that $\int_{x_0}^{t} f(x)dx = G(t) - G(x_0)$ while $\int_{\underline{x}_0}^{t} f(x)dx = G(t) - G(\underline{x}_0)$ where $G' = f$.

Now in either case the derivative of the right hand side is $G'(t) = f(t)$; hence, our assertion follows.

By the way, the fact that $\int_{x_0}^{t} f(x)dx$ depends on x_0 allows us to think of $\int_{x_0}^{t} f(x)dx$ as being an indefinite integral in the sense that the derivative will be $f(t)$ regardless of the choice of x_0, but $A(t)$, for a given t, changes by a constant as we change x_0. Pictorially:



The difference between $\int_{x_0}^{t} f(x)dx$ and $\int_{\underline{x}_0}^{t} f(x)dx$ is precisely $A_{\overline{R}}$.

In any event, our discussion can now be summarized as follows.

Second Fundamental Theorem of Integral Calculus

Let f be piecewise continuous on [a,b] and let $G(t) = \int_a^t f(x)dx$ for $t \varepsilon [a,b]$ (where now $\int_a^t f(x)dx$ denotes a particular infinite sum, not an inverse derivative).

Then G is differentiable on [a,b] and in particular $G'(t) = f(t)$ for each $t \varepsilon (a,b)$.

In summary, the two fundamental theorems show us how integral calculus and differential calculus are related. By the first theorem we see how certain infinite sums can be evaluated if we know enough about inverse differentiation, and by the second theorem we see how we can find inverse derivatives by computing suitable infinite sums.

The remainder of this block is devoted to applying the principles of this chapter to various "real" situations. We shall leave the further development to the text, exercises, and lectures. Our hope is that these three sections have provided the underlying threads that unite integral and differential calculus, and that we now fully understand that, conceptually, these two branches of calculus, while marvelously related, are still independent.

Chapter IX
LOGARITHMS REVISITED

A. Logarithms Without Exponents

In our earlier treatment of trigonometry, we stressed the fact that one could have invented the trigonometric functions even if there had been no such things as angles. Later we showed that there was a natural relationship between the "new" trigonometry and the "traditional" trigonometry. In this chapter, we shall do a similar treatment of logarithms. More specifically we shall show how logarithms could have been invented entirely within the framework of calculus, and in the next section we shall show how the "new" approach is actually completely equivalent to the old. Our main hope is that the "new" approach will give us a new insight into the nature of logarithms and make us feel a bit more at home with them. In particular, we hope to make the base of the natural logarithm system seem really natural.

Since it is often helpful to introduce a new concept in the form of a meaningful physical situation, let us consider the following plausible situation. In a certain experiment, it is observed that the rate of change of a certain quantity is proportional to the amount present. It is our wish to express the amount present at any time by an explicit mathematical formula.

The point is that we have a rather simple differential equation that tells the story. Namely,

$$\frac{dm}{dt} = km \qquad (1)$$

Separating the variables in (1), we obtain

$$\frac{dm}{m} = kdt \qquad (2)$$

If we integrate (2), we have that:

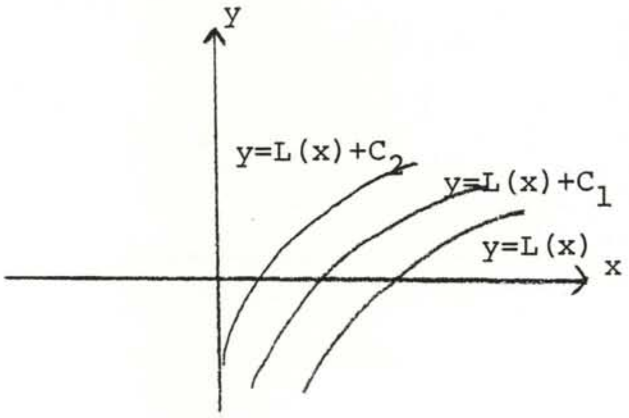
$$\int \frac{dm}{m} = kt + C \qquad (3)$$

The interesting point is that, while there is nothing at all strange about the situation depicted by (1), it turns out that the resulting solution (3) requires us to come to grips with an integral which, at least at the moment, is one that we have not learned to handle.

In fact, from another point of view, notice that $\int \frac{dx}{x}$ is the form $\int x^n dx$ with $n = -1$. Thus, whether it is to solve (3) or to find a formula for $\int x^n dx$ when $n = -1$, we are required to <u>construct</u> a function, which we shall for now denote by L, such that $L'(x) = \frac{1}{x}$.

From an intuitive point of view, the most obvious "evidence" that such a function exists lies in the realm of differential calculus. Specifically, if we are told to sketch the curve $y = L(x)$ where $L'(x) = 1/x$, we have at once that $L''(x) = -1/x^2$. This tells us that we know the slope and the concavity at any point on the curve, and this is certainly enough information to sketch the curve up to the constant of integration; that is, as usual, once one curve fills the bill, an infinite family also does. In this case, once we find one curve of the form $y = L(x)$, any other curve $y = L(x) + C$ will also fill the bill.

In fact, if we restrict the domain of L, for the moment, to positive values of x (the main thing is that we do not want $x = 0$ in the domain of L since $1/x$ and $-1/x^2$ are not defined for this value of x), we see almost at a glance that our function L has the form



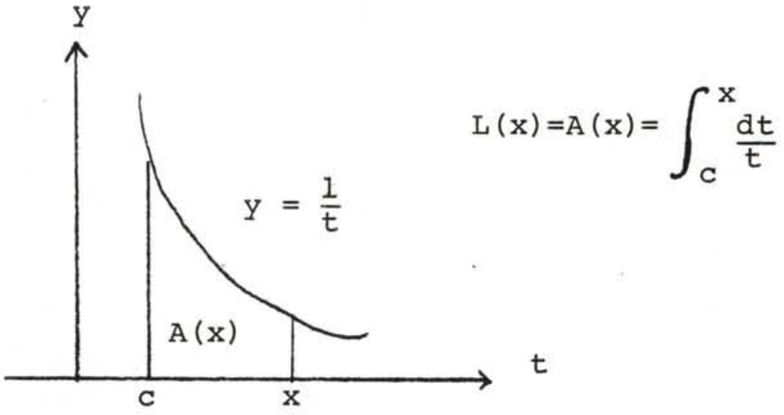
(Figure 1)

Of course, while the differential calculus approach makes the existence of L plausible, it is not too conducive towards making a more "exact" quantity out of L. One way of constructing L is through the second fundamental theorem of integral calculus. What we do in this approach is pick a number c which is greater than zero. We then can define L(x) by:

$$L(x) = \int_c^x \frac{dt}{t}$$

Clearly $L(c) = \int_c^c \frac{dt}{t} = 0$ and by the second fundamental theorem $L'(x) = \frac{1}{x}$.

Pictorially,



(Figure 2)

Why will L ultimately be called a logarithmic function? To begin with, a function f is said to be logarithmic if for all x and y in its domain, f(xy) = f(x) + f(y). Notice that this describes "ordinary" logarithms since for any base b, $\log_b(xy) = \log_b x + \log_b y$. As we shall see, every computation that makes use of ordinary logarithms takes advantage of this single property. The main question for us, however, is:

what does all this have to do with our function L? To answer this, we shall show that L is, at least, "almost-logarithmic." What this means is the following. Let us see how L(ax) and L(a) + L(x) are related. If it were to happen that for each constant, a, that L(ax) = L(a) + L(x), then, by definition, L would be a logarithmic function.

Now, we know that L'(x) = 1/x. This is precisely the definition of L. By use of the chain rule, it follows that

$$\frac{dL(u)}{dx} = \frac{dL(u)}{du} \quad \frac{du}{dx}$$

$$= \frac{1}{u}\frac{du}{dx} \qquad (4)$$

From (4) it follows that, for the case u = ax,

$$\frac{dL(ax)}{dx} = \left(\frac{1}{ax}\right) a = \frac{1}{x} \qquad (5)$$

On the other hand,

$$\frac{d(L(a)+L(x))}{dx} = \frac{1}{x} \text{ (since } L(a) \text{ is constant)} \qquad (6)$$

Keeping in mind the corollary to the mean value theorem, a comparison of (5) and (6) shows that L(ax) and L(a)+L(x) can differ by at most a constant. Thus:

$$L(ax) = L(a) + L(x) + C \qquad (7)$$

To evaluate the constant in (7), it is quite judicious to let $x = 1$ since we can then cancel the $L(a)$ terms in equation (7). With this in mind, we find that (7) becomes:

$$L(a) = L(a) + L(1) + C \qquad (8)$$

From (8) it follows that C is zero if and only if $L(1) = 0$. In other words, L will be logarithmic (from (7)) if and only if $L(1) = 0$.

Armed with this information, we now add the following restriction to our sought-for function L. Namely, we now require, as before, that $L'(x) = 1/x$, but now we also require that $L(1) = 0$. We shall give this particular member of our family a special name. We shall call it ln x (where ln is read as the natural logarithm).

As far as we are concerned, at least in this section, it is not necessary that logarithm refer to exponent. That is we use the word logarithm to indicate that ln has the general logarithmic property, and we use the word natural to indicate that the birth of ln came from a rather common occurence of nature - that the rate of change was proportional to the amount present.

In summary, then, ln is defined as follows:

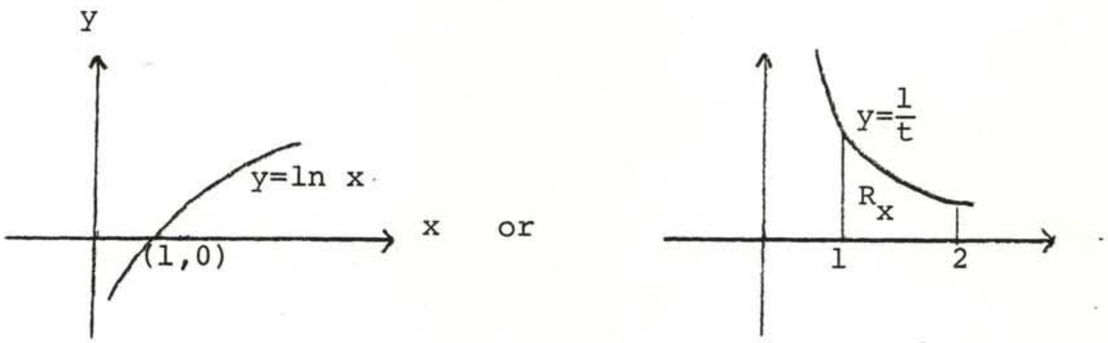
(1) the domain of ln is the set of positive real numbers

(2) for any positive x,

$$\frac{d(\ln x)}{dx} = \frac{1}{x}$$

(3) $\ln 1 = 0$

Pictorially:



$$\ln x = A_{R_x}$$

$$= \int_1^x \frac{dt}{t}$$

(Figure 3)

Let us now turn our attention to common properties shared by all logarithmic functions (including, of course, ln x).

Recall that the only property required of f to be logarithmic is:

$$f(xy) = f(x) + f(y) \qquad (9)$$

Property #1

$$f(x^2) = 2f(x)$$

Proof:

Let $x = y$ in (9) - That's it!

Corollary

For any positive integer n, $f(x^n) = n\,f(x)$

Proof:

Use Induction on Property #1. More informally,

$$f(x^n) = f[\underbrace{x\cdot x\ldots\cdot x}_{\text{n times}}] = \underbrace{f(x) + \ldots + f(x)}_{\text{n times}} = nf(x)$$

Property #2

If $x = 1$ is in the domain of f, then $f(1) = 0$.

Proof:

Let, for example, $y = 1$ in (9)

We obtain:

$$f(x) = f(x) + f(1)$$

whence, $f(1) = 0$

Property #3

If $\frac{1}{x}$ is in the domain of f, then $f\left(\frac{1}{x}\right) = -f(x)$

Proof:

Use (9) with $y = \frac{1}{x}$. Then:

$$f\left(x\cdot\frac{1}{x}\right) = f(x) + f\left(\frac{1}{x}\right)$$

at the same time, since $x\cdot\frac{1}{x} = 1$, $f\left(x\cdot\frac{1}{x}\right) = f(1) = 0$ (Property #2)

Hence: $0 = f\left(x\cdot\frac{1}{x}\right) = f(x) + f\left(\frac{1}{x}\right)$

$$\therefore\ f\left(\frac{1}{x}\right) = -f(x)$$

Property #4

$$f\left(\frac{x}{y}\right) = f(x) - f(y) \text{ (where } x, y, \text{ and } \frac{x}{y} \text{ are in domain of } f)$$

Proof:

$$f\left(\frac{x}{y}\right) = f\left(x\cdot\frac{1}{y}\right) = f(x) + f\left(\frac{1}{y}\right) = f(x) - f(y)$$

The point is that, by and large, these properties of logarithmic functions are what we use in computations. In this same vein, since ln has these same properties, we can use the natural logarithm to aid us in calculus computations. Specifically, by the use of ln we can convert products to sums, quotients to differences, etc.

By way of illustration, let us rederive the quotient rule using the natural logarithm. Suppose we are given that $y = \frac{u}{v}$ where u and v are differentiable functions of x. Then

$$y = \frac{u}{v} \rightarrow \ln y = \ln\left(\frac{u}{v}\right) = \ln u - \ln v^{*} \text{ (Property \#4)}$$

$$\therefore\ \frac{d(\ln y)}{dx} = \frac{d}{dx}(\ln u - \ln v)$$

$$\therefore\ \frac{1}{y}\frac{dy}{dx} = \frac{1}{u}\frac{du}{dv} - \frac{1}{v}\frac{dv}{dx}$$

*Technically speaking, ln u requires that u>0. If u<0 we can ignore the minus sign until the end of the problem, just as in our treatment of traditional logarithms. For example to compute (2)(-3) by base-ten logarithm, we might let N=(2)(-3)= -1[(2)(3)]. We would use logarithms to compute 2x3 and then affix the minus sign.

$$\frac{dy}{dx} = \frac{y}{u}\frac{du}{dv} - \frac{y}{v}\frac{dv}{dx}\text{, and since } y = \frac{u}{v}\text{;}$$

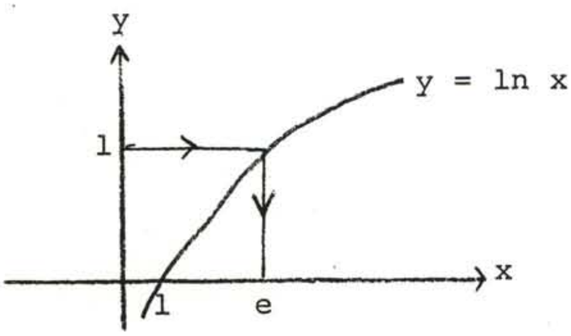
$$\frac{dy}{dx} = \frac{1}{v}\frac{du}{dx} - \frac{u}{v^2}\frac{dv}{dx}$$

$$= \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}$$

We shall save other computations for the exercises. For now, we would like to make mention of the rather special number, e, which either directly or indirectly, suggests itself here.

The idea is that if ln has genuine logarithmic properties it must also possess the analog of a base. Notice that in base b, b is characterized by $\log_b b = 1$. In this context, let us define e by ln e = 1. In this way, we make an artificial connection between ln x and $\log_e x$ (which we shall explore in more detail in the next section).

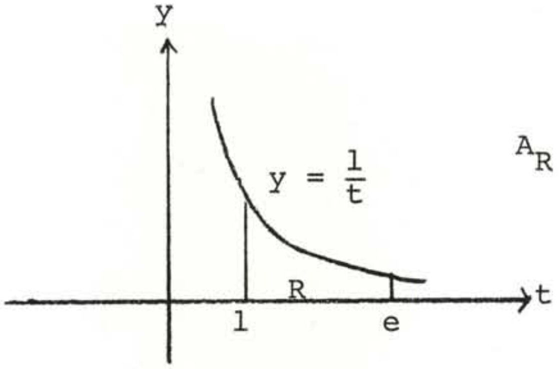
Geometrically, the number e is defined either by



(Figure 4)

or

or by



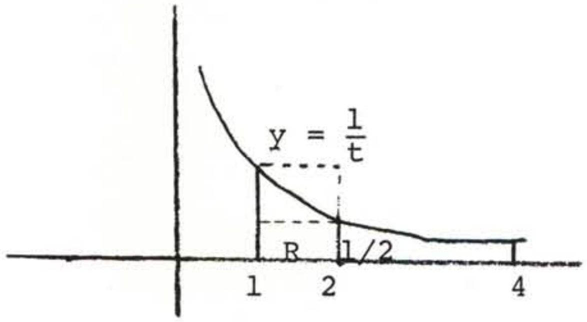
$$A_R = 1 = \int_1^e \frac{dt}{t}$$

(Figure 5)

From Figure 5 it is easy to show that the number e defined by ln e = 1 indeed exists. In fact, for the same price we can show that

$$\underline{2 < e < 4}$$

Namely,



$$\ln 2 = A(R)$$

$$1\left(\frac{1}{2}\right) < A(R) < 1(1)$$

$$\frac{1}{2} < \ln 2 < 1 \qquad (10)$$

In particular, $\ln 2 < 1$ (11)

On the other hand $\ln 2 > \frac{1}{2} \rightarrow 2 \ln 2 > 1$, but

$$2 \ln 2 = \ln 2^2 = \ln 4 \text{ (Property \#1)}$$

$$\ln 4 > 1 \tag{12}$$

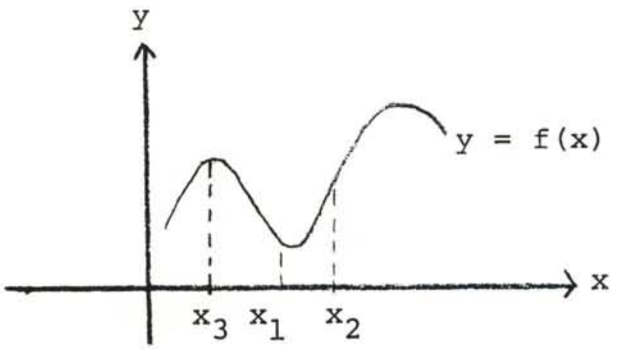
Combining (11) and (12) with the fact that $\ln e = 1$, we obtain:

$$\ln 2 < \ln e < \ln 4 \tag{13}$$

Finally the fact that ln is 1 - 1 and increasing, allows us to conclude from (13):

$$\underline{2 < e < 4}$$

(Notice how we need the 1 - 1 property here. In general $f(x_1)<f(x_2)<f(x_3)$ tells us nothing about the ordering of x_1, x_2, and x_3. For example,



in the above diagram $f(x_1)<f(x_2)<f(x_3)$ but it is neither true that $x_1<x_2<x_3$ nor $x_1>x_2>x_3$.)

In Summary,

The function $y = \ln x$ is a natural outgrowth of the problem $\frac{dm}{dt} = km$. In fact, the solution is precisely

$$\frac{dm}{m} = kdt$$

$$\therefore \int \frac{dm}{m} = kt + c$$

$$\therefore \quad \ln|m|^{*} = kt + c$$

Without the physical application, we have

$$\int x^{n}dx = \begin{cases} \frac{1}{n+1}\, x^{n+1} + c & n \neq -1 \\ \ln|x| + c & n = -1 \end{cases}$$

At no time during our presentation are we <u>required</u> to acknowledge the traditional concept of logarithms. In the next section, we shall attempt to unite the "new" and "old" logarithms.

[*] Again $\ln x$ requires that x be positive. If $x<0$, $\int \frac{dx}{x}$ makes sense but $\ln x$ doesn't. In this case we argue as follows: If $x<0$ then $x = -u$ where $u>0$. Then $dx = -du$

$$\therefore \int \frac{dx}{x} = \int \frac{-du}{-u} = \int \frac{du}{u} = \ln u + c \qquad \text{(since } u>0\text{)}$$

$$= \ln(-x) + c \quad \text{, but } x<0 \rightarrow |x| = -x$$

$$\therefore \int \frac{dx}{x} = \ln|x| + c$$

B. A Note on the Connection Between ln x and $\log_e x$

For those of us who would like to see a more concrete demonstration that ln x = $\log_e x$, we may use the following approach. Let $f(x) = \log_b x$ where b is some arbitrary base and $\log_b x$ is the "traditional" logarithm.

Then, as usual:

$$f'(x_1) = \lim_{\Delta x \to 0} \left[\frac{\log_b (x_1 + \Delta x) - \log_b x_1}{\Delta x} \right] \qquad (1)$$

Now:

$$\log_b (x_1 + \Delta x) - \log_b x_1 = \log_b \frac{x_1 + \Delta x}{x_1}$$

$$= \log_b \left(1 + \frac{\Delta x}{x_1} \right) \qquad (2)$$

Putting (2) into (1) yields:

$$f'(x_1) = \lim_{\Delta x \to 0} \left[\frac{\log_b \left(1 + \frac{\Delta x}{x_1} \right)}{\Delta x_1} \right]$$

$$= \lim_{\Delta x \to 0} \left[\frac{1}{\Delta x} \log_b \left(1 + \frac{\Delta x}{x_1} \right) \right] \qquad (3)$$

Then recalling that $c \log_b u = \log_b u^c$, we may rewrite (3) as:

$$f'(x_1) = \lim_{\Delta x \to 0} \left[\log_b \left(1 + \frac{\Delta x}{x_1}\right)^{\frac{1}{\Delta x}} \right] \qquad (4)$$

(Using hindsight in the sense that we suspect that $f'(x)$ should involve $\frac{1}{x}$, we might be tempted to try a substitution such as $u = \frac{\Delta x}{x_1}$ in the hope of obtaining $\frac{1}{x}$, as a factor.)

Letting $u = \frac{\Delta x}{x_1}$ we have $\Delta x = ux_1$ or $\frac{1}{\Delta x} = \frac{1}{ux_1}$, and also that $u \to 0$ as $\Delta x \to 0$. Putting these facts into (4), we find:

$$f'(x_1) = \lim_{u \to 0} \left[\log_b (1+u)^{\frac{1}{ux_1}} \right]$$

$$= \lim_{u \to 0} \left[\log_b \left\{ (1+u)^{\frac{1}{u}} \right\}^{\frac{1}{x_1}} \right] \qquad (5)$$

Equation (5) gives us our chance to obtain $\frac{1}{x_1}$ as a factor. Namely

$$\log_b(\ )^{\frac{1}{x_1}} = \frac{1}{x_1} \log_b(\ ).$$

Hence, (5) becomes

$$f'(x_1) = \lim_{u \to 0} \left[\frac{1}{x_1} \log_b (1+u)^{\frac{1}{u}} \right] \qquad (6)$$

and since x_1 is a constant, (6) becomes:

$$f'(x_1) = \frac{1}{x_1} \lim_{u\to 0}\left[\log_b (1+u)^{\frac{1}{u}}\right]$$

$$= \frac{1}{x_1} \log_b\left[\lim_{u\to 0} (1+u)^{\frac{1}{u}}\right]^{*} \qquad (7)$$

Among other things, equation (7) shows us that the basic limit problem in developing the calculus of traditional logarithms lies in evaluating

$$\boxed{\lim_{u\to 0} (1+u)^{\frac{1}{u}}} \qquad (8)$$

In essence, this limit is to the development of logarithms what $\lim\limits_{u\to 0} \frac{\sin u}{u}$ was to the development of the (circular) trigonometric functions.

[*] When f and g are continuous, $\lim\limits_{x\to a} f(g(x)) = f(\lim\limits_{x\to a} g(x))$. Namely, $\lim\limits_{x\to a} g(x) = g(a)$ implies that $\lim\limits_{x\to a} f(g(x)) = \lim\limits_{g(x)\to g(a)} f(g(x))$ $= f(g(a)) = f(\lim\limits_{x\to a} g(x))$. While we may be tempted to interchange the order of operations as it suits us, we must learn to be cautious (as we shall see especially in Block VII) and to make sure that the interchange is valid.

To maintain our flow of thought here, let us assume, without taking time out to prove our assertion, that the limit in (8) exists. As an interesting aside, let us observe in (8) that if we "illegally" replace u by 0 we wind up with 1^{∞} which is another indeterminate form.*

At any rate, again using hindsight, let us define the number e by:

$$e = \lim_{u\to 0} (1+u)^{\frac{1}{u}} \tag{9}$$

Using (9) in conjunction with (7) we now have the following fundamental result:

$$\boxed{\begin{array}{l}\text{Let } f(x) = \log_b x \ (x>0). \text{ Then } f \text{ is} \\ \text{differentiable, and, in particular} \\ \qquad f'(x) = \frac{1}{x}\log_b e\end{array}} \tag{10}$$

Equation (10) now tells us how b must be chosen if we desire that $f'(x) = \frac{1}{x}$. Namely, from (10) $f'(x) = \frac{1}{x}$ if and only if $\log_b e = 1$, and this in turn requires that $b = e$.

*It is sometimes difficult to think of 1^{∞} as being indeterminate since we might view $1^{\infty} = 1\text{x}1\text{x}1\ldots$. If this were the case 1^{∞} would equal 1. The fact is, however, that 1^{∞} was arrived at by looking at an endless product whose factors approached 1 as a limit. For example, in regard to (8) if we let $u = 0.01$, $(1+u)^{1/u} = (1.01)^{100}$. Granted that 1.01 is near 1 it is still greater than 1; hence, raised to the 100th power, it might be appreciably greater than 1. In fact, if $N = (1.01)^{100}$, then $\log_{10}N = 100 \ \log_{10}1.01=100(.0043)=0.43$

$\therefore N = 2.69$

We are now ready to identify ln x with $\log_e x$. More specifically, both ln x (as seen in the previous section) and $\log_e x$ (as seen from (10)) have $\frac{1}{x}$ as their derivative. Hence, they differ by a constant. That is:

$$\ln x = \log_e x + c \tag{11}$$

Letting x = 1 in (11), we find

$$\ln 1 = \log_e 1 + c \quad \text{or } 0 = 0 + c$$

$\therefore$ c = 0 and (11) becomes:

$$\boxed{\ln x \equiv \log_e x} \tag{12}$$

The connection between ln x and $\log_e x$ may now be stated as follows: Suppose we seek a function L(x) subject to the two conditions (1) L(1)=0 and (2) $L'(x)=\frac{1}{x}$, x>0. Then:

$$L(x) = \log_e x$$

In summary, let us observe that in the previous section we named the above L(x) by ln x and in this context there was no need to understand traditional logarithms. In this section, we showed that we might just as well talk about traditional logarithms since ln x was a synonym for $\log_e x$.

As a final note on this topic, let us observe that the choice of e was not crucial other than in the sense it helps us avoid keeping track of a cumbersome constant. For example, if $y = \log_{10}x$ then $\frac{dy}{dx} = \frac{1}{x}\log_{10}e = \frac{k}{x}$ where k in this case equals $\log_{10}e$ or approximately 0.43. No matter what the base b, if $y = \log_b x$ then $\frac{dy}{dx} = \frac{c_1}{x}$, where $c_1 = \log_b e$ This, in turn, yeilds the equation

$$\underline{\frac{dx}{x} = \frac{dy}{c_1} = c_2 dy}$$

which is basically the same form as before.

Chapter X
Sequences and Series

A. Introduction.

The concept of infinity defies our intuition, if only because our intuition is unable to tell us about infinity. That is, every man-made operation is finite, and we tend to associate infinity with an extremely large but *finite* number. Yet, unless we are extremely careful, we get into tremendous difficulty if we do not separate the concepts of *extremely large* and *infinite*.

For example, let us denote by M an extremely large number. Perhaps M would take several years to write if we were to use place value notation. But, no matter how large we choose it, M would be followed by M + 1, M + 2, etc., and we would be no closer to the "end" of our number system than when we started. In fact, it seems that we are still at the beginning of the system with M as our new reference point. In other words, perhaps the first point we should come to grips with is that no matter how large a finite number we have, it is no "nearer" to infinity than one at the beginning of the number system. In still other words, compared with infinity, any finite number is "small."

We may illustrate this point in other ways. For example, it is intuitively clear to us that there are as many even whole numbers as there are odd whole numbers. Suppose we now agree to list the whole numbers by starting with the first two odd numbers, then the first even number, and continue in this way. We obtain

1,3,2,5,7,4,9,11,6,13,15,8,17,19,10,21,23,12,...

Notice that, no matter what even number we stop at, there will always be twice as many odd numbers in our list as even numbers, AND YET WE WILL NOT RUN OUT OF ODD NUMBERS

BEFORE WE RUN OUT OF EVEN NUMBERS. The paradox that there now seem to be twice as many odd numbers as even numbers is immediately resolved if we observe that we said "no matter what even number we STOP at." In other words, hopefully, this example shows the difference between going on FOREVER and going as far as we want provided that we eventually stop. Notice that the variations on the theme here are endless. We could have "proved" that there were twice as many even numbers as odd ones by writing

2,4,1,6,8,3,10,12,5,....

or that there were five times as many odds as evens by writing

1,3,5,7,9,2,11,13,15,17,19,4,....

etc.

In terms of ordinary arithmetic, we can also show the difference between large and infinite. For example, in adding any finite set of numbers we may change the order of the terms (in the "new" language, addition is commutative), and once the terms are in a given order we can group them in any way we wish (addition is associative). The amazing thing is that these properties, as "self-evident" as they may seem, are not always obeyed by infinite sets of numbers. As a simple illustration consider the sum

$$1 + (-1) + 1 + (-1) + 1 + (-1) + \ldots.$$

If we assume the "voice inflection" (grouping)

$$[1 + (-1)] + [1 + (-1)] + [1 + (-1)] + \ldots.$$

we obtain $\underline{0}$ as our "sum."

On the other hand, if we use the grouping

$$1 + [(-1) + 1] + [(-1) + 1] + \ldots.$$

we obtain $\underline{1}$ as our "sum."

There is no contradiction here. Rather, what we have shown is that the sum of an infinite amount of numbers may well depend on how the numbers are grouped.

As a final example, let us consider the notion of dividing one polynomial by another. As a specific illustration, let us divide 1 by 1 - x. If we carry out the division through n steps we find that

$$\frac{1}{1 - x} = 1 + x + x^2 + \ldots + x^n + \frac{x^{n+1}}{1 - x} \quad . \qquad (1)$$

The point is that equation (1) is true for any value of x, no matter how great n is, provided (as we have done) we eventually stop and "tack on" the remainder term.

For example:

$$\frac{1}{1 - x} \equiv 1 + x + x^2 + \ldots + x^{99} + \frac{x^{100}}{1 - x} \quad .$$

On the other hand, if we were now to carry out the division forever (that is, without ever stopping and "tacking on" the remainder term), we would obtain:

$$\frac{1}{1 - x} = 1 + x + x^2 + x^3 + \ldots + x^n + \ldots . \qquad (2)$$

But equation (2) is no longer an identity. For example, if we let x = 2, the left hand side of (2) becomes -1 while the right hand side becomes 1 + 2 + 4 + 8 + 16 + ... or infinity.

At any rate, it should now be clear that if we are to come to grips with the arithmetic of infinite sets of numbers we are going to have to "revisit" the elements of finite arithmetic in order to decide what remains valid and what doesn't, and this will be the concern of the remainder of this chapter. As to why we would want to investigate "infinite arithmetic," it is hoped that our previous experience in the course, especially with regard to the infinite sums defined by definite integrals, is sufficient to make the answer to this query obvious.

B. (Infinite) Sequences.

In most of the important mathematical operations, we utilize the fact that we may accomplish our mission in a finite number of steps, even though this finite number might be very large. For example, suppose we wanted to arrange a billion numbers according to size. To find the smallest we could compare the first two members and discard the greater. We could then compare the "survivor" with the third member and again discard the greater. We proceed in this way making a billion (less one) comparisons, and the last "survivor" is the smallest number in the collection. We could then, of course, repeat this procedure with the remaining numbers to find the next smallest number, etc., but the key point is that while it might take a long time to arrange these billion numbers, it can be done and in a finite number of steps.

The crucial point is that the above procedure does not apply to infinite sets, since we never run out of elements to be compared. Notice, especially in terms of some earlier remarks, that this is a much stronger statement than saying it might take several years to complete making the necessary comparisons.

In other words, the problem with infinite sets is that, by definition, no matter how we elect to order the members, there is no last member. Yet, somehow or other, we would like to have the analog of a "last term" even when we deal with the problem of ordering an infinite set. (By the way, when a set is arranged so that its members appear in a given order, we refer to the elements in the given order as a sequence. In particular, an infinite sequence is a listing on an infinite set.)

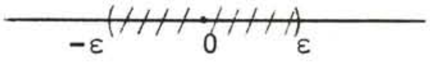
While we shall, in a moment, talk about this more generally, notice that we come to grips with this problem in our study of area when we form areas with inscribed and circumscribed rectangles. We called these sums L_n and U_n respectively, and we then investigated what happened as n was allowed to increase without bound. It was in this context that the notations $\lim\limits_{n\to\infty} U_n$ and $\lim\limits_{n\to\infty} L_n$ were introduced, and our definition of limit was such that, from a practical point of , it played the role of a "last term." We can generalize this notion without reference to area.

For example, suppose a_n denotes the nth element of set A where $a_n = \frac{1}{n}$; n any positive whole number. According to the implied listing $a_1, a_2, a_3, \ldots, a_n, \ldots$ we have

$$A = \{1, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \ldots, \tfrac{1}{n}, \ldots\} \quad .$$

Notice that according to our above listing, A has no last member. To be sure $\lim\limits_{n\to\infty} \frac{1}{n} = 0$, but 0 is not the last term in A if only because $0 \notin A$! That is, there exists no whole number n such that $\frac{1}{n} = 0$.

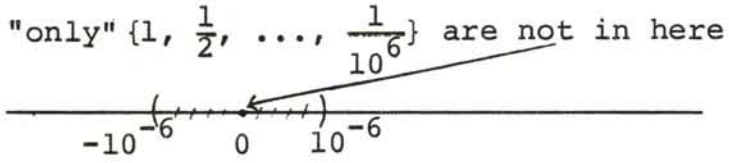
Here is where the notion of limit helps us resolve the dilemma of a "last term" for an infinite sequence. Namely, with respect to the given example, $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} \frac{1}{n} = 0$. This, in turn means that if we choose any positive number ε and look at the interval $(-\varepsilon, \varepsilon)$,



.

That "after a while" every a_n belongs to $(-\varepsilon, \varepsilon)$. By "after a while," of course, we mean that, for the given ε , we can find a number N such that $n > N$ implies $a_n \varepsilon (-\varepsilon, \varepsilon)$. How large N must be depends on the size of ε. The important point is that, no matter how small we choose ε, the desired N is still finite even though it might be very large. For example if we desire that $(0 <) \frac{1}{n} < 10^{-6}$ then we must choose $n > 10^6$. In other words, all but the first million terms of $\{\frac{1}{n}\}$ belong to $(-10^{-6}, 10^{-6})$ and a million terms is "negligible" when we are concerned with infinitely many terms.

To explain this idea further, suppose we wanted to "keep our eye" on every term of the sequence $\{\frac{1}{n}\}$. Picking, say, $\varepsilon = 10^{-6}$ we have:


"only" $\{1, \frac{1}{2}, \ldots, \frac{1}{10^6}\}$ are not in here
-10^{-6}
0
10^{-6}

In other words we now have $\{\frac{1}{n}\}$ "under surveillance" in the sense that we have a finite number of specific terms (namely $1, \frac{1}{2}, \ldots, \frac{1}{10^6}$) and the "rest" of the terms squeezed into the "narrow" interval $(-10^{-6}, 10^{-6})$.*

In fact, if we wish to get more geometrical in our approach, we may again talk about dots versus points. Notice that $(-10^{-6}, 10^{-6})$ may be viewed as a dot (a "thick point") while $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{10^6}$ may be viewed as points. Thus, the concept of limit allows us to replace infinitely many points by a finite number of points plus one "dot," with the actual finite number of points being dependent upon the thickness of the "dot." (The "dot," of course, is the geometric analog of the ε-neighborhood of 0 .)

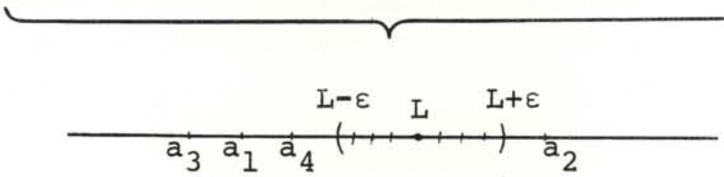
More generally, by way of review, a sequence $\{a_n\}$, $\lim_{n\to\infty} a_n = L$ (or that the sequence $\{a_n\}$ converges to L).

*Notice in this example that $n > 0$ implies that $\frac{1}{n} > 0$. Thus no a_n is in $(-10^{-6}, 0)$. However, since $(-10^{-6}, 10^{-6})$ contains $(-10^{-6}, 0)$ as a subset no harm is done by our notation.

means that for each $\varepsilon > 0$ we can find a number N (which, in general, depends on ε) such that $n > N \longrightarrow |a_n - L| < \varepsilon$.

Pictorially,

After a certain term (the "Nth") all the rest are in $(L - \varepsilon, L + \varepsilon)$.



In other words, to find upper and lower bounds for the convergent sequence $\{a_n\}$ (recall that if $\lim\limits_{n\to\infty} a_n$ exists we call the sequence convergent), we need only check the finite* sequence of numbers:

$$a_1, a_2, \ldots, a_n, L - \varepsilon, L + \varepsilon \quad .$$

We shall explore some of these ideas more computationally in the exercises, but for now, as long as we have come this far, we would like to establish a few precise definitions and notions about bounds - especially those which will be of help to us later in this chapter.

In what follows, let S denote any set whose numbers are real numbers.

Definition 1

M is called an upper bound for $S \longleftrightarrow s \leq M$, for all $s \in S$.

*Notice that it is crucial that only a finite number of terms be outside the interval $(L - \varepsilon, L + \varepsilon)$. This is much stronger than saying infinitely many terms are in $(L - \varepsilon, L + \varepsilon)$. For example, in the sequence 1,-1,1,-1,1,-1,... infinitely many terms lie in $(\frac{1}{2}, \frac{3}{2})$: namely, each odd-numbered term is 1 and $\frac{1}{2} < 1 < \frac{3}{2}$. Yet, infinitely many terms also lie outside this interval, since $-1 \not\in (\frac{1}{2}, \frac{3}{2})$.

Definition 1'

m is called a lower bound for S $\longleftrightarrow$ m $\leqslant$ s for all s ε S .

Definition 2

S is said to be bounded above if it has an upper bound. (Notice that not all sets are bounded above; for example, since there is no largest whole number, the whole numbers are not bounded above.)

Definition 2'

S is said to be bounded below if it has a lower bound. (The set of negative integers is not bounded below.)

Definition 2"

S is said to be bounded if it is both bounded above and bounded below. More symbolically, S is said to be bounded if there exists mumbers m and M such that $m \leqslant s \leqslant M$ for all s ε S .

Note:

The notion of bounds is trivial for finite sets. For example, if S has N elements, we may arrange them according to size, say,

$$a_1 \leqslant a_2 \leqslant \ldots \leqslant a_{N-1} \leqslant a_N$$

whence a_1 is a lower bound for S and a_N is an upper bound for S. Indeed, S is bounded since

$$a_1 \leqslant s \leqslant a_N \quad \text{for all} \quad s \ \varepsilon \ S \quad .$$

In the case that S is finite, as above, we can even go one step further. Namely, any number less than a_N cannot be an upper bound for S (since a_N is in S and exceeds it) while any number greater than a_1 cannot be a lower bound for S. This leads to:

Definition 3

M is called a least upper bound (lub) for S if

(a) M is an upper bound for S, and

(b) if $K < M$ then K is not an upper bound for S.

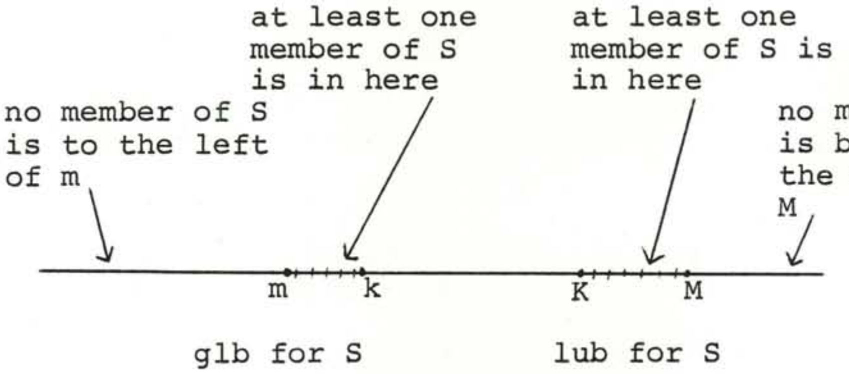
Definition 3'

m is called a greatest lower bound (glb) for S if

(a) m is a lower bound for S, and

(b) if $k > m$ then k is not a lower bound for S.

Pictorially:


at least one member of S is in here
at least one member of S is in here
no member of S is to the left of m
no member of S is beyond (to the right of) M
m
k
K
M
glb for S
lub for S

These results are not so obvious for infinite sets. In fact we have already seen that infinite sets need not be bounded, either above or below. In terms of additional examples, consider:

$$S = (0,1) = \{x: < 0 < x < 1\} \quad .$$

Clearly, if $M \geqslant 1$ then M is an upper bound for S, while if $m \leqslant 0$, m is a lower bound for S. It is also easy to see that any number less than 1 is not an upper bound for S. Therefore $\begin{cases} 1 \text{ is the lub for } S \\ 0 \text{ is the glb for } S \end{cases}$.

Yet neither 0 nor 1 belong to S! In other words for an infinite set the least upper bound and/or the greatest lower bound need not belong to the set itself.

The basic idea that we shall invoke here as an axiom stated without proof, but which is self-evident for finite sets, is:

> If S is bounded above (below) then it possesses a least upper bound (greatest lower bound). However, neither the lub nor glb need be members of S.

We shall return to these concepts later.

C. Cauchy Sequences.

In a manner of speaking, this section should have been included as part of the previous section. Our textbook, however, does not treat this particular topic, and, as a result, we felt that because the material is new, it should be accentuated in the form of a separate section.

The key point that motivates this section is the fact that there are many times when we want to know that a sequance converges, even if we do not need to know what the limit is explicitly. For example, in terms of a trivial but real-life illustration, we often have to use $\sqrt{2}$ in some type of computation, and we must express it in decimal form. Now $\sqrt{2}$ cannot be expressed explicitly as a decimal, but it can be defined as the limit of a sequence of rational numbers. Namely, 1, 1.4, 1.41, 1.414, 1.4142, ... each approximate $\sqrt{2}$ to one more decimal place, and by one technique or another, we can compute each succeeding member of the sequence. The point is that since we know that the sequence converges to $\sqrt{2}$, we can "chop off" the sequence anywhere we want, once we are certain that we have the desired degree of accuracy. Thus, if we determine that we do not need more than two decimal places, it is of little consequence that we cannot compute $\sqrt{2}$ exactly as a decimal, since in our application it would be indistinguishable from 1.41 .

If we desire a less familiar and, hence, more challenging application of this idea, we may consider the following sequence:

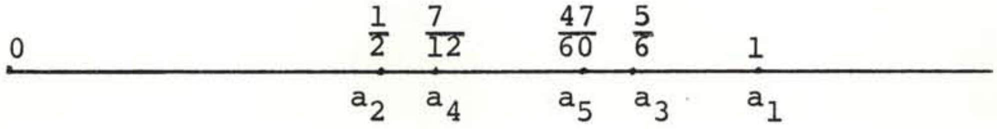
$$a_1 = 1,\ a_2 = 1 - \frac{1}{2},\ a_3 = 1 - \frac{1}{2} + \frac{1}{3},\ a_4 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \ldots,$$

$$a_n = \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} \ .^{*}$$

In this case, our sequence becomes

$$1,\ \frac{1}{2},\ \frac{5}{6},\ \frac{7}{12},\ \frac{47}{60},\ \ldots .$$

Let us convince ourselves that the given sequence converges. Perhaps the easiest way is geometrically, in terms of the number line. We have:

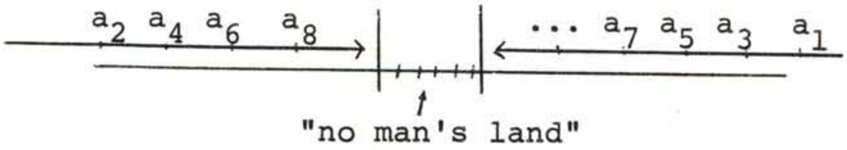


While we haven't proved the validity yet (but the proof is intuitively obvious as we shall soon see), a pattern seems to be emerging. The odd numbered term =

*While it may look complicated the factor $(-1)^k$ is a "sign-alternator." That is, $(-1)^k$ is -1 when k is odd and +1 when k is even. In our example, we want the positive terms to occur in the odd-numbered positions; that is, when k = 1, 3, 5, ... We merely observe that when k is odd, k+1 is even. Thus, $(-1)^{k+1}$ is positive when k is odd and negative when k is even.

$a_1, a_3, a_5, \ldots$ form a decreasing sequence (i.e., the points $a_1, a_3, a_5, \ldots$ move "steadily" from right to left) while the even numbered terms form an increasing sequence. Moreover, these two sequences are "segregated" in the sense that no odd numbered term ever comes between two even numbered terms and vice versa.

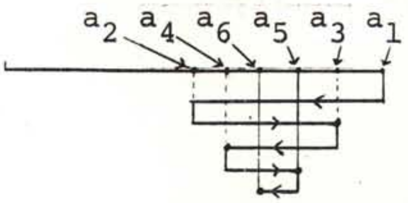
That is,



Finally, notice that, since we go from a_n to a_{n+1} by adding or subtracting $\frac{1}{n+1}$, and, since $\lim_{n\to\infty} \frac{1}{n+1} = 0$, the size of our "no man's land" approaches 0 as n approaches infinity. That is, there is some number (point) L to which the sequences $a_1, a_3, a_5, \ldots$ and $a_2, a_4, a_6, \ldots$ both converge, and this number L (even if we can't express it explicitly) is the limit of our sequence.*

To see this result from another point of view, notice that, because of the alternating signs, each even numbered term must appear to the left of the immediately preceeding odd numbered term (since we subtract the even numbered terms), and, similarly, each odd numbered term must appear to the right of the immediately preceeding even numbered term. Moreover, since $\frac{1}{n}$ decreases as n increases we see that each "shift" (in magnitude) is less than the previous one, and this is why the odd-numbered and even-numbered sequences are "segregated." That is:

*We will prove it later in this chapter but the limit of our sequence $1, \frac{1}{2}, \frac{5}{6}, \ldots, \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k}, \ldots$ is actually $\ell n\ 2$! For those of us who did not already know this, how likely is it that we would have discovered this no matter how many terms in the sequence we investigated?



Finally, since $a_{n+1} - a_n = \pm\frac{1}{n+1}$ (the sign depends on whether n is even or odd, but unambiguously $|a_{n+1} - a_n| = \frac{1}{n+1}$), we have $\lim_{n\to\infty} (a_{n+1} - a_n) = 0$.

Thus, since either n or n+1 is even and the other is odd, we see that the size of our "shifts" approaches 0 as n increases without bound. That is, our sequence "zeroes-in" on some limit L. In terms of a final schematic representation we have:

Now, the fairly lengthy preceeding example is merely meant to focus our attention on the crucial point that we might want to test a sequence for convergence even when we do not have an explicit guess as to what the limit, if it exists, is. It is in this context that one talks about a Cauchy sequence or the Cauchy Criterion for Convergence.

Definition:

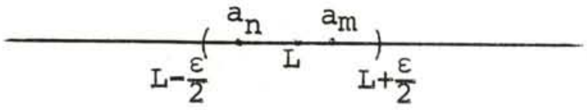
The sequence $\{a_n\}$ is called a Cauchy sequence if for each $\varepsilon > 0$ we can find a number N (which in general will depend on the choice of ε) such that for every n and m for which $n > N$ and $m > N$, $|a_n - a_m| < \varepsilon$. The key point is that

> A sequence of real numbers converges if and only if it is a Cauchy sequence.

This key result is probably the single most important result on the study of sequential convergence. Among other things, it allows us to study convergence whether or not we have any idea as to what the limit is. That is, $|a_n - a_m| < \varepsilon$ never requires reference to the limit L.

From the most intuitive point of view, assuming that we now have mastered the geometric interpretation of absolute values, it is easy to see why a Cauchy sequence is a convergent sequence. Namely, what the definition of a Cauchy sequence says is that, beyond a certain term, the difference, (i.e., the distance) between any two of the the remaining terms (points) can be made as small as we please. If we call this arbitrarily-small number ε, then there is a "band-width" of size ε into which all the remaining terms must fit. Since ε was chosen arbitrarily, L must be somewhere in this band-width.

Conversely, if we know that the sequence $\{a_n\}$ converges to L, we can reverse the preceeding process. Namely, given $\varepsilon > 0$, we let $\varepsilon_1 = \varepsilon/2$. Then we can find N such that $n > N$ implies that $|a_n - L| < \varepsilon_1$. Pictorially,



$$\left.\begin{array}{l} n > N \longrightarrow a_n \; \varepsilon (L - \frac{\varepsilon}{2}, L + \frac{\varepsilon}{2}) \\ \\ \\ m > N \longrightarrow a_m \; \varepsilon (L - \frac{\varepsilon}{2}, L + \frac{\varepsilon}{2}) \end{array}\right\} \begin{array}{c} a_n, a_m \; \varepsilon (L - \frac{\varepsilon}{2}, L + \frac{\varepsilon}{2}) \longrightarrow \\ |a_n - a_m| < \varepsilon \end{array}$$

(Geometrically, the distance between a_n and a_m cannot exceed the distance between $L - \frac{\varepsilon}{2}$ and $L + \frac{\varepsilon}{2}$.)

While we do not think it is important for our purposes to "beat this idea to death" rigorously, it is important that we recognize that all of our intuitive (geometric) ideas can be developed analytically. For illustrative purposes, let us translate our last result into more analytic terms.

We are given that $\lim_{n\to\infty} a_n = L$, and we want to show that, given $\varepsilon > 0$, we can find N such that $n, m > N \longrightarrow$

$|a_n - a_m| < \varepsilon$.

Now, by the analytic properties of absolute values, we have

$$|a_n - a_m| = |(a_n - L) + (L - a_m)| \leqslant |a_n - L| + |L - a_m|$$

$$= |a_n - L| + |a_m - L| \quad .$$

Thus $|a_n - a_m|$ will be less than ε as soon as, for example, $|a_n - L|$ and $|a_m - L|$ are each less than $\frac{\varepsilon}{2}$. But by definition of $\lim_{n\to\infty} a_n = L$, given $\varepsilon > 0$, let $\varepsilon_1 = \frac{\varepsilon}{2}$; then, there exists N such that $k > N \longrightarrow |a_k - L| < \varepsilon_1 = \frac{\varepsilon}{2}$. Hence, $n > N$, $m > N \longrightarrow |a_n - L|$ and $|a_m - L|$ are each less than $\frac{\varepsilon}{2}$.

Restated, with all motivation ommitted, we have

<u>Given</u> $\lim_{k\to\infty} a_k$* $= L$

<u>To prove</u> Given $\varepsilon > 0$, there exists N such that $n, m > N \longrightarrow |a_n - a_m| < \varepsilon$.

<u>Proof</u> Let $\varepsilon_1 = \frac{\varepsilon}{2}$. Then $\lim_{k\to\infty} a_k = L$ means we can find N such that $k > N \longrightarrow |a_k - L| < \varepsilon_1$. Thus $n > N$ and $m > N \longrightarrow |a_n - L| < \varepsilon_1$ and $|a_m - L| < \varepsilon_1$.

*Notice that our subscript is a dummy variable. We switched from the dummy variable n to dummy variable k so that n would not be used to stand for two different things in the same problem. Namely we use n as an index and we also talk about n and m > N .

$\therefore$ If $n > N$, $m > N$, we have:

$$|a_n - a_m| = |(a_n - L) + (L - a_m)|$$

$$\leq |a_n - L| + |L - a_m| = |a_n - L| + |a_m - L|$$

$$< \varepsilon_1 + \varepsilon_1 = 2\varepsilon_1 = 2(\tfrac{\varepsilon}{2}) = \varepsilon$$

q.e.d.

We shall, at least for the time being, not pursue the analytic aspects of Cauchy sequences any further. Our main point was that we wanted to emphasize that one did not have to suspect what the limit of a sequence was in order to test the sequence for convergence. Later, we shall investigate still other tests for convergence that do not require knowledge of the limit.

More importantly, not only is the idea of Cauchy sequences used extensively in more advanced analysis courses (and therefore this would be reason enough to be familiar with the concept) but there will be occasions in the remainder of our discussion when it will be to our advantage to be able to use the Cauchy criterion rather than the more "conventional" definition of convergence.

At any rate, this completes our introduction to the concept of sequences.

D. Series.

The process of addition may be viewed in terms of a sequence - a sequence of partial sums. For example, to compute 1 + 2 + 3 + 4, we, consciously or otherwise, compute the partial sums: 1, 1 + 2, (1 + 2) + 3, [(1 + 2) + 3] + 4, or 1, 3, 6, 10 .

The actual sum is our last partial sum which, in this case, is 10.

We must, of course, be careful not to confuse the sequence of partial sums with the sequence of numbers being added. That is, the sequence of numbers being added is 1, 2, 3, 4, . The sequence of partial sums is 1, 3, 6, 10 . We may picture the difference in terms of a simple desk calculator. For example 1, 2, 3, and 4 are successively "punched in" and in the "answer slot" we see the successive sums 1, 3, 6, 10 .

Also, notice that 1 + 2 + 3 + 4 is one number (10) not four numbers; it is the sum of four numbers.

At any rate, with these ideas in mind, let us now turn our attention to infinite sums. The point is that, with an endless number of summands (terms), our sequence of partial sums becomes infinite and consequently there is no last term in this sequence. We, therefore, invoke the analog of a last term in a finite sequence. (namely the limit), and we define the sum of an infinite number of terms to be the limit of the sequence of partial sums.

Before pursuing this idea further, let us make sure that we can distinguish between sequences and series.

Simply identify (infinite) sequences with listings while series correspond to adding an infinite sequence of terms. The connection between the two is that the sum of a series is the limit of the sequence of partial sums. In still other words, the study of series is a special case of the study of sequences-sequences of partial sums.

While our discussion may seem rather formal, let us recall that we were dealing with series (even though we didn't call them that) when we were studying areas (the definite integral).

For example, suppose we are given the series

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{2^n} + \ldots$$

We proceed as in the finite case by listing the sequence of partial sums. Thus:

$$s_1 = \frac{1}{2}$$

$$s_2 = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

$$s_3 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{3}{4} + \frac{1}{8} = \frac{7}{8} \quad .$$

With luck, we would now begin to suspect that the denominator of s_n is 2^n and the numerator is one less than the denominator. That is:

$$s_n = \frac{2^n - 1}{2^n} = 1 - \frac{1}{2^n} \quad . \qquad (1)$$

The validity of this observation may be verified by mathematical induction.*

We may spot check (1) for specific values of n. For example, if n = 10, we find:

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{1024} = \frac{1023}{1024} = 1 - \frac{1}{2^{10}} \quad .$$

The point is that if we <u>stop</u> after the nth term, equation (1) gives us the <u>exact</u> sum of the n numbers $\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^n}$ and this follows from the fact that (1) yields s_n which is the nth partial sum.

If we want the sum of the <u>series</u>, then the sequence of partial sums becomes infinite. We then <u>define</u> the sum to be $\lim_{n\to\infty} s_n$.

*Later in this section, we shall give a more general way of evaluating the given type of series which requires no guess work.

In other words:

$\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots$ is defined* to mean

$\lim\limits_{n\to\infty} (\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n})$ or $\lim\limits_{n\to\infty} s_n$ where $s_n = \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n}$.

Since $\lim\limits_{n\to\infty} \frac{1}{2^n} = 0$, we see in this case that

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} (1 - \frac{1}{2^n}) = 1 - \lim_{n\to\infty} \frac{1}{2^n} = 1^{**} \quad .$$

Hence:

$$\begin{aligned} \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots &= \lim_{n\to\infty} (\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n}) \\ &= \lim_{n\to\infty} (1 - \frac{1}{2^n}) \\ &= 1 \quad . \end{aligned}$$

Thus, by definition, the sum of the series $\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots$ is 1 - not approximately 1, but exactly 1. What is "approximately" 1 is s_n for large [but finite] values of n.

*As usual, one is not required to "prove" a definition. It should be noted, however, that there are often several ways of "inventing" a definition and we chose, wherever possible, a definition that agrees with what we believe to be reality. In this respect, notice that our definition of the sum of an infinite series is in agreement with our "method of exhaustion" for finding areas whereby we "squeezed" the given region between rectangular networks.

**We have been using limit theorems concerning sequences without proving them according to the definition of $\lim\limits_{n\to\infty} a_n$. (All our proofs were for continuous variables, that is, for $\lim\limits_{x\to a} f(x)$.) We shall supply the more rigorous proofs, for discrete sequences as part of our exercises.

If we now introduce the notation $\sum_{n=1}^{\infty} \frac{1}{2^n}$ to denote $\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots$ (that is, $\sum_{n=1}^{\infty} \frac{1}{2^n} = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{2^k}$) we would write:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = 1 \quad .$$

Perhaps now, with the present illustration in mind, it might be wise to generalize our discussion:

Given the sequence of real numbers $\{a_n\}$, we may elect to form their "sum." That is, $a_1 + \ldots + a_n + \ldots = \sum_{n=1}^{\infty} a_n$. We define the "sum" to mean

$$\lim_{n\to\infty} (a_1 + \ldots + a_n) \quad \text{or} \lim_{n\to\infty} \sum_{k=1}^{n} a_k \quad .$$

$a_1 + \ldots + a_n$ is the nth partial sum of the terms and if we denote this by s_n we have:

Definition

By $\sum_{n=1}^{\infty} a_n$ we mean $\lim_{n\to\infty} s_n$ where $s_n = a_1 + \ldots + a_n$ provided the limit exists. If the limit exists, the limit is called the sum of the series $\sum_{n=1}^{\infty} a_n$. In still other words, $\sum_{n=1}^{\infty} a_n = L$ means $\lim_{n\to\infty} (a_1 + \ldots + a_n) = L$.

Note:

We bring the notion of a Cauchy sequence into our discussion of series. Namely the convergence of $\sum_{k=1}^{\infty} a_k$ involves the convergence of $\{s_k\}$ where $s_k = a_1 + \ldots + a_k$.

In turn, by the Cauchy criterion, the convergence of $\{s_k\}$ requires that for each $\varepsilon > 0$ we can find N such that $n > N$ and $m > N$ implies $|s_n - s_m| < \varepsilon$. However, $s_n - s_m = (a_1 + \ldots + a_n) - (a_1 + \ldots a_m)$. If we now assume, without loss of generality, that $n > m$, we have $s_n - s_m = a_{m+1} + \ldots + a_n = \sum_{k=m+1}^{n} a_k$. This leads to

The series $\sum_{k=1}^{\infty} a_k$ converges if and only if for each $\varepsilon > 0$ there exists a number N, depending on ε, such that $\left|\sum_{k=m+1}^{n} a_k\right| < \varepsilon$ whenever $n > m > N$* .

We shall have occasion to use the Cauchy criteron later, but for now our note is complete.

Before closing this section, we would like to say a few words about the generalization of the series $\sum_{n=1}^{\infty} \frac{1}{2^n}$. Back in high school algebra, we usually studied the topic called "Geometric Progressions." Recall that a geometric progression was one in which the ratio between any term and its immediate predecessor was the same. Thus, if this ratio is denoted by r and the first term by a, the geometric progression(sequence) is given by:

$$a, ar, ar^2, \ldots, ar^n, \ldots$$

*We should not take the symbol m+1 too literally. Since m denotes a "sufficiently large" number so also does m+1. For this reason, one often finds the definition of "Cauchy-convergent" given in the form ".....

$$\left|\sum_{k=m}^{n} a_k\right| < \varepsilon \quad \text{whenever } n \geqslant m > N."$$

With this notation in mind, a geometric series is one of the form

$$\sum_{n=0}^{\infty} ar^n = a + ar + ar^2 + \dots + ar^n + \dots \text{ *}$$

In this case, we have

$s_0 = a$

$s_1 = a + ar$

$s_2 = a + ar + ar^2$

$s_n = a + ar + ar^2 + \dots + ar^n$,

and we wish to compute $\lim_{n\to\infty} s_n$.

*Do not be alarmed that we have suddenly switched from $\sum_{n=1}^{\infty}$ to $\sum_{n=0}^{\infty}$. For example we have ten fingers and this does not depend on whether we enumerate them from 1 through 10 or 0 through 9. In fact the ten digits are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Without belaboring this point, the idea is that we choose $\sum_{n=1}^{\infty}$ or $\sum_{n=0}^{\infty}$ depending on which is the more convenient. For example, $\sum_{n=1}^{\infty} ar^n$ $= ar + ar^2 + \dots + ar^n + \dots$ and this is not the same as $a + ar + \dots + ar^n + \dots$ while $\sum_{n=0}^{\infty} ar^n$ is. Had we insisted on the form $\sum_{n=1}^{\infty}$, we would have to write $\sum_{n=1}^{\infty} ar^{n-1}$ ($= a + ar + \dots$) and this would be okay but slightly more cumbersome than $\sum_{n=0}^{\infty} ar^n$.

It turns out there is a particularly "cute" way of computing $\lim_{n\to\infty} s_n$ (other less cute ways are also available) if we are dealing with a geometric series. The "cute way" is a generalization of "tacking on a zero" in our decimal arithmetic when we want to multiply by 10. The key idea is that when we multiply s_n by r we "almost get" s_{n+1}. That is:

$$s_n = a + ar + \dots + ar^{n} \quad .$$

Therefore

$$rs_n = ar + ar^2 + \dots + ar^{n+1} \quad .$$

Hence:

$$\begin{aligned} s_n - rs_n &= (a + ar + \dots + ar^n)-(ar + \dots + ar^n + ar^{n+1}) \\ &= a - ar^{n+1} \\ &= a(1 - r^{n+1}) \quad . \end{aligned}$$

But, $s_n - rs_n = (1 - r)s_n$.

Therefore:

$$(1 - r)s_n = a(1 - r^{n+1}) \quad .$$

Thus, if $r \neq 1$* we obtain

$$s_n = \frac{a(1 - r^{n+1})}{1 - r} \quad .$$

*If r = 1 the series is a + a + ... + a + ... which clearly diverges unless a = 0 . That is $sn = \underbrace{a + \dots + a}_{n \text{ times}} = na$

$\therefore a \neq 0 \longrightarrow \lim_{n\to\infty} s_n = \lim_{n\to\infty} na = \infty$.

The next key point is that $\lim_{n\to\infty} r^{n+1} = 0$ if $|r| < 1$. (If $r > 1$ $\lim_{n\to\infty} r^{n+1} = \infty$.) Hence if $|r| < 1$ then

$$\lim_{n\to\infty} s_n = \frac{a}{1-r} \quad .$$

In summary

If $|r| < 1$, then the sum of the series $\sum_{n=0}^{\infty} ar^n$ is $\frac{a}{1-r}$.

Again, there is no need to memorize the abstract derivation. For example, suppose we were given

$$2 + \frac{2}{3} + \frac{2}{9} + \frac{2}{27} + \ldots + \frac{2}{3^n} + \ldots = \sum_{n=0}^{\infty} 2(\tfrac{1}{3})^n \quad .$$

This is a geometric series with $a = 2$ and $r = \frac{1}{3}$. The key is that:

$$s_n = 2 + \frac{2}{3} + \ldots + \frac{2}{3^n}$$

$$\therefore \frac{1}{3} s_n = \qquad \frac{2}{3} + \ldots + \frac{2}{3^n} + \frac{2}{3^{n+1}}$$

$$\therefore s_n - \frac{1}{3} s_n = 2 - \frac{2}{3^{n+1}}$$

$$\text{or } \frac{2}{3} s_n = 2 - \frac{2}{3^{n+1}}$$

$$\therefore \frac{2}{3} \lim_{n\to\infty} s_n = 2 - 0 = 2 \quad \text{or} \quad \lim_{n\to\infty} s_n = \frac{3}{2} \times 2 = 3$$

$$2 + \frac{2}{3} + \ldots + \frac{2}{3^n} + \ldots = 3$$

and this checks with $\frac{a}{1-r}$ with $a = 2$, $r = \frac{1}{3}$, namely:

$$\frac{a}{1-r} = \frac{2}{1-1/3} = \frac{2}{2/3} = 3 \quad .$$

We shall explore geometric series further in the exercises. We wish to conclude this section with the remark that while geometric series are highly specialized they do, nevertheless, play an important role in the general theory of series.

E. Absolute Convergence and Rearrangements.

Note:

Much of the material in this and subsequent sections is not in the text. Thus, to supplement the lectures, the new ideas are discussed rather informally and in more generality than in the lectures. Since the interested student might desire a more formal treatment (for example, in terms of having theorems and proofs made available) while other students might only be distracted by such a procedure, we have elected to leave the theorems and their proofs for the end of each section. In this way, the student who wishes to omit the proofs can move on to the next section without loss of continuity, while the student who wants to see the various proofs then has the opportunity to do so. In any event we want only to emphasize that proofs are included only to make the course self-contained.

So far we have stressed positive series. Obviously, however, a series need not be positive. The aim of this section is to discuss series which may have negative terms and to show how such series are related to appropriate positive series.

To introduce our topic, let us observe that if our series has both positive and negative terms then there are two basically different ways in which the series can converge. On the one hand, it might be that the magnitudes of the terms themselves get small fast enough to cause the series to converge. On the other hand, it might be that the magnitudes do not get small fast enough but that positive terms and negative terms cancel one another to bring about the convergence.

Let us discuss the first case. If the magnitudes of the terms get small fast enough, then to all intents and purposes we are saying that the series would still converge even if we replaced every term by its magnitude. This, in turn, means that we leave the positive terms alone and "delete" the minus signs on the negative terms. More mathematically, we are saying to replace by $\sum_{n=1}^{\infty} a_n$, by $\sum_{n=1}^{\infty} |a_n|$.

It can be shown (for the proof see Theorem 1 at the end of this section) that the convergence of $\sum_{n=1}^{\infty} |a_n|$ implies the convergence of $\sum_{n=1}^{\infty} a_n$. With this in mind, we have:

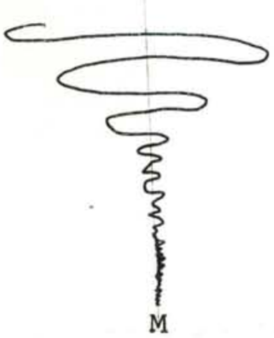
<u>Definition</u>

The series $\sum_{n=1}^{\infty} a_n$ is said to be <u>absolutely</u> <u>convergent</u> if $\sum_{n=1}^{\infty} |a_n|$ is convergent. (Notice it is $\sum_{n=1}^{\infty} a_n$, which we are calling absolutely convergent.)

The significance of absolute convergence is that if $\sum a_n$ converges absolutely then the sum of the series does not depend on how the a_n's are rearranged (see Theorem 2). At first glance, this result might seem trivial, since, in fact, the result <u>is</u> trivial in the case of the sum of a finite number of terms. That is, when we add a finite number of terms, the sum does not depend on how the terms are grouped or rearranged. Consequently, we might "expect" the same to be true when we deal with tne sum of infinitely many terms. It turns out, however, that if $\sum_{n=1}^{\infty} a_n$ converges but not absolutely, we can change the sum merely by changing the order of the terms! More surprisingly, we can find rearrangements in which the resulting series actually diverges.

More specifically, what we shall show is that if a series converges but not absolutely, then the subseries which consists of the positive terms and the subseries which consists of the negative terms each diverge to infinity (see Theorem 3)! Given any number M, we add the positive terms until the sum exceeds M (which must happen since the positive series diverges to infinity). We then annex the sum of the negative terms until the sum falls below M. We then take up the positive terms where we left off until the sum exceeds M again (and this must happen since the positive series diverges to infinity and we have only used up a finite number of (finite) terms so that what remains must still be infinite) and then we annex negative terms until the sum falls below M again. We continue in this way, ad infinitum, and we converge to M as the limit. The proof that M is the limit lies in the fact that both the positive and negative terms must approach zero in magnitude otherwise $\sum_{n=1}^{\infty} a_n$ wouldn't converge (recall that for a series to converge its nth term must approach zero as a limit). Thus every time we again exceed or fall below M we do it by smaller and smaller amounts since the size of each "jump" approaches zero as n approaches infinity.

Pictorially,



Using the above discussion as motivation, we have:

Definition

$\sum_{n=1}^{\infty} a_n$ is said to be conditionally convergent if $\sum_{n=1}^{\infty} a_n$ is convergent but $\sum_{n=1}^{\infty} |a_n|$ diverges (to infinity).

The key point is that while a conditionally convergent series is a bona fide convergent series, we must be sure to form the sequence of partial sums in the precise order in which the terms are given. If we rearrange the terms (and observe that such a rearrangement certainly changes the sequence of partial sums, and from that point of view it is not quite so surprising to observe that different sequences can have different limits) we get a different series which may not only have a different sum than that of the original series but may not even converge. Again, this does not negate the convergence of the given series, but it forces us to take no liberties with it.

On the other hand, if the series converges absolutely, we can rearrange the terms to suit our convenience without changing the sum.

In other words, the chief benefit of an absolutely convergent series is that we may treat it in just the same way, with respect to rearranging the terms, as we would any finite number of terms.

In still other words:

If $\sum_{n=1}^{\infty} a_n$ converges absolutely, the sum is independent of the order in which we arrange the terms. If, however, $\sum_{n=1}^{\infty} a_n$ converges conditionally and if c denotes any real number then there is a rearrangement of $\sum_{n=1}^{\infty} a_n$, say $\sum_{n=1}^{\infty} r_n$, such that $\sum_{n=1}^{\infty} r_n = c$.

Theorem 1

If $\sum_{k=1}^{\infty} |a_k|$ converges, then so also does $\sum_{k=1}^{\infty} a_k$.

Paraphrase

If a series converges absolutely then it converges. At first glance, this may sound like a truism but the point is that the definition of absolute convergence of $\sum_{k=1}^{\infty} a_k$ utilizes only properties of $\sum_{k=1}^{\infty} |a_k|$. To be sure, it is very likely that if Theorem 1 were not true, we would never have invented the notion of calling $\sum_{k=1}^{\infty} a_k$ absolutely convergent if $\sum_{k=1}^{\infty} |a_k|$ converged.

Strategy behind the proof

We know that $\left|\sum_{k=m}^{n} a_k\right| \leqslant \sum_{k=m}^{n} |a_k|$. We also know that since $\sum_{k=1}^{\infty} |a_k|$ converges we can find N such that for a given $\varepsilon > 0, n \geqslant m > N \longrightarrow \sum_{k=m}^{n} |a_k| < \varepsilon$. Therefore

$n, m > N \longrightarrow |\sum_{k=m}^{n} a_k| \leq \sum_{k=m}^{n} |a_k| < \varepsilon$, and this is precisely the Cauchy condition that $\sum_{k=1}^{\infty} a_k$ converges .

Formal Proof

We shall show that, given $\varepsilon > 0$, we can find N such that $n \geqslant m > N \longrightarrow |\sum_{k=m}^{n} a_k| < \varepsilon$.

Since $\sum_{k=1}^{\infty} |a_k|$ converges, we can find N such that $n \geqslant m > N$ implies

$$\sum_{k=m}^{n} |a_k| < \varepsilon \quad .$$

But $|\sum_{k=m}^{n} a_k| \leqslant \sum_{k=m}^{n} |a_k|$

$\therefore n \geqslant m > N \longrightarrow |\sum_{k=m}^{n} a_k| < \varepsilon$ q.e.d.

Theorem 2

Suppose $\sum_{k=1}^{\infty} a_k = L$ and that $\sum_{k=1}^{\infty} a_k$ is absolutely convergent. Let $\sum_{k=1}^{\infty} r_k$ denote any series obtained by rearranging the a_k's . Then $\sum_{k=1}^{\infty} r_k = L$.

Paraphrase:

If a series converges absolutely then its sum is independent of the order in which the terms are added.

Strategy

Given $\varepsilon > 0$, we know that we can find N such that $\sum_{k=N+1}^{\infty} |a_k| < \varepsilon$ (since $\sum_{k=1}^{\infty} |a_k|$ converges). We then look at the rearrangement $r_1, r_2, \ldots$ and proceed until all the terms $a_1, a_2, \ldots, a_N$ appear in the list. (This must happen since the r_k's are merely a rearrangement of the a_k's; hence, the finite set of numbers $\{a_1, \ldots, a_N\}$ must eventually appear in the rearrangement.) Let's suppose $a_1, \ldots, a_N$ are contained among $r_1, \ldots, r_M$. We then look at $\sum_{k=M+1}^{\infty} |r_k|$. Since $\{a_1, \ldots, a_N\} \subseteq \{r_1, \ldots, r_M\}$ it follows that r_{M+1}, r_{M+2}, etc. must be included in $\{a_{N+1}, a_{N+2}, \ldots\}$. (That is if $A \subset B$ then $A' \supset B'$ where A' denotes the complement of A. In our example, $\{a_{N+1}, a_{N+2}, \ldots\} = \{a_1, \ldots, a_N\}'$ while $\{r_{M+1}, r_{M+2}, \ldots\} = \{r_1, \ldots, r_M\}'$ }.) Therefore,

$$\sum_{k=M+1}^{\infty} |r_k| \leqslant \sum_{k=N+1}^{\infty} |a_k| \text{ and } \sum_{k=M+L}^{\infty} |a_k| < \varepsilon \quad .$$

This proves that $\sum r_k$ is at least absolutely convergent and gives us some experience with new essential notation.

The remainder of our strategy is as follows:

Using the same notation as before, we have:

$$\sum_{k=1}^{\infty} r_k = \sum_{k=1}^{M} r_k + \sum_{k=M+1}^{\infty} r_k \quad .$$

The point is that $\{a_1, \ldots, a_N\} \subseteq \{r_1, \ldots, r_M\}$; hence, $\sum_{k=1}^{M} r_k = \sum_{k=1}^{N} a_k$ plus $(M - N)$ terms chosen from $\{a_{N+1}, a_{N+2}, \ldots\}$. Let us label these by $a_{i_1}, a_{i_2}, a_{i_{M-n}}$ (the double subscript is used to indicate that we have $(M - N)$ a_k's, but we do not know how they are ordered among the r_k's) and that $\sum_{j=1}^{M-N} |a_{i_j}| \leqslant \sum_{k=N+1}^{\infty} |a_k|$

$$\sum_{k=1}^{\infty} r_k = \sum_{k=1}^{N} a_k + \sum_{j=1}^{M-N} a_{i_j} + \sum_{k=M+1}^{\infty} r_k$$

$$\therefore \left|\sum_{k=1}^{\infty} r_k - \sum_{k=1}^{N} a_k\right| \leqslant \left|\sum_{j=1}^{M-N} a_{i_j} + \sum_{k=M+1}^{\infty} r_k\right|$$

$$\leqslant \sum_{j=1}^{M-N} |a_{i_j}| + \sum_{k=M+1}^{\infty} |r_k|$$

$$\leqslant \sum_{k=N+1}^{\infty} |a_k| + \sum_{k=M+1}^{\infty} |r_k| \quad .$$

But since $\sum_{k=1}^{\infty} |a_k|$ and $\sum_{k=1}^{\infty} |r_k|$ are convergent both $\sum_{k=N+1}^{\infty} |a_k|$ and $\sum_{r=M+1}^{\infty} |r_k|$ can be made as small as we wish just by taking N (and hence M since $M \geqslant N$) sufficiently large. That is:

$$\lim_{N\to\infty} \left|\sum_{k=1}^{\infty} r_k - \sum_{k=1}^{N} a_k\right| = 0$$

$$\therefore \sum_{k=1}^{\infty} r_k = \lim_{N\to\infty} \sum_{k=1}^{N} a_k$$

$$= \sum_{k=1}^{\infty} a_k$$

$$= L \quad .$$

Formal Proof

We must show that

$$\sum_{k=1}^{\infty} r_k - \sum_{k=1}^{\infty} a_k = 0 \quad .$$

Let $\varepsilon > 0$ be given. Set $\varepsilon_1 = \frac{\varepsilon}{2}$. Then we can find N such that

$$\sum_{k=N+1}^{\infty} |a_k| < \varepsilon_1 \quad . \qquad (1)$$

We can then find M such that $\{a_1, \ldots, a_N\} \subseteq \{r_1, \ldots, r_M\}$

$$\therefore \sum_{k=M+1}^{\infty} |r_k| < \varepsilon_1 \quad (\text{since } \sum_{k=M+1}^{\infty} |r_k| \leqslant \sum_{k=N+1}^{\infty} |a_k|)_{(2)} \cdot$$

Therefore:

$$\sum_{k=1}^{\infty} r_k = \sum_{k=1}^{M} r_k + \sum_{k=M+1}^{\infty} r_k$$

$$= \sum_{k=1}^{N} a_k + \sum_{j=1}^{M-N} a_{i_j} + \sum_{k=M+1}^{\infty} r_k \quad \text{where } a_{i_j} \epsilon \{a_{N+1}, a_{N+2}, \ldots\}$$

$$\left|\sum_{k=1}^{\infty} r_k - \sum_{k=1}^{N} a_k\right| = \left|\sum_{j=1}^{M-N} a_{i_j} + \sum_{k=M+1}^{\infty} r_k\right|$$

$$\leq \sum_{j=1}^{M-N} |a_{i_j}| + \sum_{k=M+1}^{\infty} |r_k|$$

$$\leq \sum_{k=N+1}^{\infty} |a_k| + \sum_{k=M+1}^{\infty} |r_k|$$

$$< \varepsilon_1 + \varepsilon_1 \quad \text{(by (1) and (2))}$$

$$< \varepsilon \quad .$$

∴ For each $\varepsilon > 0$ there exists N such that

$$\left|\sum_{k=1}^{\infty} r_k - \sum_{k=1}^{N} a_k\right| < \varepsilon$$

$$\therefore \lim_{N\to\infty} \sum_{k=1}^{N} a_k = \sum_{k=1}^{\infty} r_k \qquad \text{q.e.d.}$$

<u>Theorem 3</u>

Suppose $\sum_{n=1}^{\infty} a_n$ converges conditionally. Then $\sum_{a_n>0} a_n = \infty$

and $\sum_{a_n<0} a_n = -\infty$ (where $\sum_{a_n>0} a_n$ means the sum of all positive terms and $\sum_{a_n<0} a_n$ means the sum of all negative terms. More generally if S denotes any set of numbers $\sum_{s\varepsilon S} s$ is used to denote the sum of the elements in S.)

Paraphrase

The sum of the positive terms in a conditionally convergent series is infinite. Likewise, the sum of the negative terms is negative but infinite in magnitude.

Proof

Let $p_n = \dfrac{a_n + |a_n|}{2}$ and $q_n = \dfrac{a_n - |a_n|}{2}$.

Since $|a_n| = a_n$ if $a_n > 0$ while $|a_n| = -a_n$ if $a_n < 0$ we see that $p_n = \begin{cases} a_n & \text{if } a_n > 0 \\ 0 & \text{if } a_n \leqslant 0 \end{cases}$ while $q_n = \begin{cases} 0 & \text{if } a_n \geqslant 0 \\ a_n & \text{if } a_n < 0 \end{cases}$.

Therefore:

$$\sum_{n=1}^{\infty} p_n = \sum_{a_n>0} a_n \quad \text{(since } p_n = a_n \text{ if } a_n > 0 \text{ while } p_n = 0 \text{ if } a_n \leqslant 0 \text{ .)}$$

and $\displaystyle\sum_{n=1}^{\infty} q_n = \sum_{a_n<0} a_n$.

We now show that neither $\sum_{n=1}^{\infty} p_n$ nor $\sum_{n=1}^{\infty} q_n$ can converge. For example, suppose $\sum_{n=1}^{\infty} q_n$ converges. Well, we know that $\sum_{n=1}^{\infty} a_n$ converges and therefore $\sum_{n=1}^{\infty} (\frac{a_n}{2} - q_n)$

converges*. Since $q_n = \frac{a_n - |a_n|}{2}$ we see that

$$\frac{a_n}{2} - q_n = \frac{|a_n|}{2}$$

$$\therefore \sum_{n=1}^{\infty} \frac{|a_n|}{2} \text{ converges}$$

$$\therefore \sum_{n=1}^{\infty} |a_n| \text{ converges}$$

$\therefore \sum_{n=1}^{\infty} a_n$ converges absolutely - but this contradicts the hypothesis that $\sum_{n=1}^{\infty} a_n$ is conditionally convergent.

Similarly, if $\sum_{n=1}^{\infty} p_n$ converges, we conclude that $\sum_{n=1}^{\infty} (p_n - \frac{a_n}{2}) (= \sum_{n=1}^{\infty} \frac{|a_n|}{2})$ converges and the same contradiction follows.

Hence $\sum_{n=1}^{\infty} p_n \ (= \sum_{a_n > 0} a_n)$ and $\sum_{n=1}^{\infty} q_n (= \sum_{a_n < 0} a_n)$ both diverge.

*Recall that $\lim_{n\to\infty} (a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n$. Thus, in terms of sequences of partial sums

$$\sum_{n=1}^{\infty} a_n, \sum_{n=1}^{\infty} b_n \text{ both convergent} \longrightarrow$$

$\sum_{n=1}^{\infty} (a_n + b_n)$ is convergent and is equal to $\sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n$.

F. Sequences of Functions.

Up to now, we have considered sequences of numbers. We can generalize our consideration rather nicely by replacing numbers by functions. For example, let f_1, f_2, ... denote a sequence of functions and suppose that the domain of each is $[a,b]$*. We now pick any number $c \varepsilon [a,b]$ and look at the sequence $f_1(c)$, $f_2(c)$, $f_3(c)$, ..., $f_n(c)$, ... The key point is that this sequence is a sequence of numbers, and as such it makes sense for us to investigate $\lim\limits_{n\to\infty} f_n(c)$. Moreover, we can make this study for each $c \varepsilon [a,b]$. The limit may exist for each c, or for some but not for others, or for none of the c's. By way of illustration, consider the following examples:

Examples

(a) Let $f_n(x) = x^n$, dom $f_n = [0,1]$. Then for $c \varepsilon [0,1]$ we have $f_1(c) = c$, $f_2(c) = c^2$, $f_3(c) = c^3$, ... $f_n(c) = c^n$, ... More concretely, let $c = \frac{1}{2}$. Then our sequence $f_1(c)$, $f_2(c)$, ... becomes $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, ..., $\frac{1}{2^n}$.

Now, if $c = 1$, $\lim\limits_{n\to\infty} c^n = 1$ while if $0 \leqslant c < 1$ $\lim\limits_{n\to\infty} c^n = 0$. Thus, in this example, $\lim\limits_{n\to\infty} f_n(c)$ exists for each $c \varepsilon [0,1]$.

*It is really not important that each function in the sequence has the same domain. What is important is that if we pick a number c then $f_1(c)$, $f_2(c)$, $f_3(c)$, ... must all be defined. In more precise language, c must belong to the intersection of the domain of f_1, the domain of f_2, etc. In more symbolic language when we study a sequence of functions we consider as our "inputs" only members of

$$\bigcap_{n=1}^{\infty} \text{dom } f_n = \text{dom } f_1 \cap \text{dom } f_2 \cap \ldots \cap \text{dom } f_n \cap \ldots$$

(b) Let $f_n(x) = x^n$, dom $f_n = [0,2]$. As we have just seen $\lim_{n\to\infty} f_n(x)$ exists if $0 \leqslant x \leqslant 1$. Now let $c > 1$. Then $f_n(c) = c^n$ implies $\lim_{n\to\infty} f_n(c) = \infty$. (In other words, we are making use of the fact that $\lim_{n\to\infty} c^n = \begin{cases} 0, \text{ if } 0 \leqslant c < 1 \\ 1, \text{ if } c = 1 \\ \infty, \text{ if } c > 1 \end{cases}$.)

Here, the point is that for $c > 1$, c^n is well-defined no matter how large n is as long as it is finite, but $\lim_{n\to\infty} c^n = \infty$.

In this example, then, $\lim_{n\to\infty} f_n(x)$ exists if $x \in [0,1]$, but it doesn't exist if $x \in (1,2]$.

(c) Let $f_n(x) = x^n$, dom $f_n = [2,3]$. Then $\lim_{n\to\infty} f_n(x) = \infty$ for each $x \in [2,3]$. For example, if we let $x = 2$, our sequence is $2,4,8,16,32,64,128,\ldots 2^n, \ldots$

While our three illustrations are simple and very similar to each other, they serve to illustrate the three possibilities we have described.

A natural consequence of our discussion of sequences of functions is the notion of a limit function. That is, if we return to our general form where the sequence is denoted by $f_1, f_2, \ldots, f_n, \ldots$ and dom $f_n = [a,b]$ for each n, we may "induce" a new function f as follows. We choose any $x \in [a,b]$ and let $f(x) = \lim_{n\to\infty} f_n(x)$. Of course, as we have just discussed, it is possible that for some of our choices of x, $\lim_{n\to\infty} f_n(x)$ need not exist. Since we require that the "output" of f be a real number, such an x is not allowed as a member of dom f. In other words, we may define f as follows:

Let $A = \{x\colon \lim_{n\to\infty} f_n(x) \text{ exists}, x \in [a,b]\}$.

(In other words, A is a subset of [a,b]). Then for each $x \in A$ let $f(x) = \lim_{n\to\infty} f_n(x)$.

Again by way of examples, let us return to our earlier example. We have already seen that $\lim\limits_{n\to\infty} c^n = 0$ if $0 \leqslant c < 1$ while $\lim\limits_{n\to\infty} 1^n = 1$. Thus, if we let $f(c) = \lim\limits_{n\to\infty} f_n(c) = \lim\limits_{n\to\infty} c^n$ for each $c \varepsilon [0,1]$, we have that $\lim\limits_{n\to\infty} f_n(c) = 0 = f(x)$ if $c \neq 1$ while $f_n(1) = 1 = f(1)$. Written in the more usual notation:

$$f(x) = \begin{cases} 0, \text{ if } 0 < x < 1 \\ 1, \text{ if } x = 1 \quad . \end{cases}$$

As a second example, let f_n be defined by

$$f_n(x) = \frac{nx^2}{2n+1} \quad ,$$

and, just for the sake of argument, suppose dom $f_n = [0,1]$. (Actually, in this example, the domain of each f_n could just as well be any set of real numbers since for each n, $f_n(x)$ is defined for every real x.)

If we pick $x = 0$, we see that $f_n(x) = f_n(0) = \frac{n0^2}{2n+1} = 0$. Hence $f(0) = \lim\limits_{n\to\infty} f_n\ (0) = \lim\limits_{n\to\infty} 0 = 0$. If we pick $x = 1$, we see that $f_n(x) = f_n(1) = \frac{n1^2}{2n+1} = \frac{n}{2n+1}$; hence, $f(1) = \lim\limits_{n\to\infty} f_n(c) = \lim\limits_{n\to\infty} \frac{n}{2n+1} = \frac{1}{2}$.

More generally, in this example, if we let $x = c$, we see that

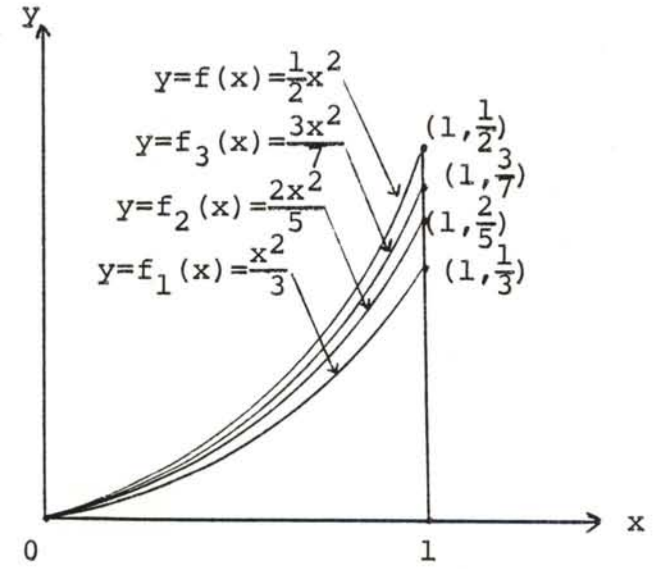
$$f(c) = \lim_{n\to\infty} f_n(c) = \lim_{n\to\infty} \frac{nc^2}{2n+1} = \left(\lim_{n\to\infty} \frac{n}{2n+1}\right)\left(\lim_{n\to\infty} c^2\right) = \frac{1}{2}c^2 \quad .$$

In other words, in this example, f is defined by

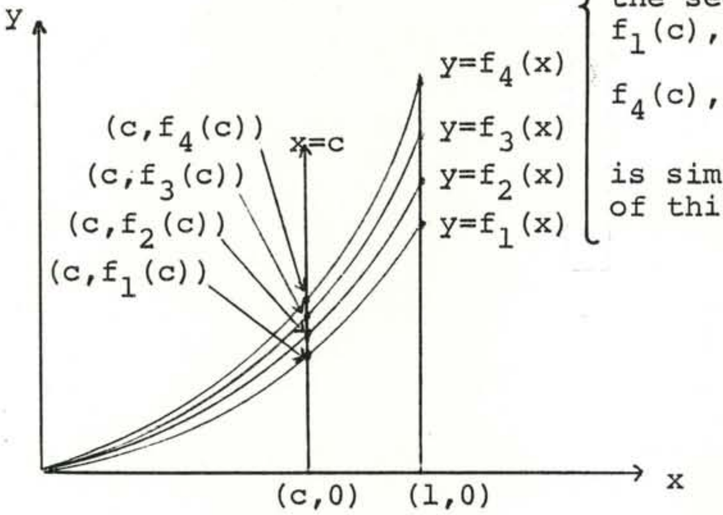
$$f(x) = \frac{x^2}{2} \text{ for each } x \varepsilon [0,1] \quad .$$

For those of us who would like to view the discussion more pictorially, we may look at the last example (and the method will apply equally well to any other example) as follows.

For each n, f_n is defined by $f_n(x) = nx^2/(2n + 1)$. Hence, $f_1(x) = x^2/3$, $f_2(x) = 2x^2/5$, $f_3(x) = 3x^2/7$, etc. We may then sketch each of the curves $y = f_n(x)$. That is, we draw $y = x^2/3$, $y = 2x^2/5$, $y = 3x^2/7$, etc. Finally, we sketch the graph of the limit function $y = x^2/2$. Thus, we have



We then "freeze" on a point (c,0) in [0,1] and draw the line x = c. We then get a sequence of numbers which consist of the y-coordinates of the points on $y = f_1(x)$, $y = f_2(x)$, etc. where the line x = c meets the curve. That is:



What $\lim_{n\to\infty} f(c)$ means pictorially is that we may view $x=c$ as the number line which contains the sequence of points $f_1(c)$, $f_2(c)$, $f_3(c)$, $f_4(c)$, ... and $\lim_{n\to\infty} f_n(c)$ is simply the limit of this sequence.

In this regard, then, notice that the limit function consists of the set of points $(c, \lim_{n\to\infty} f_n(c))$. That is, for each $c \varepsilon [0,1]$ we can make the height denoted by $|f(c) - f_n(c)|$ smaller than any prescribed amount merely by choosing n to be greater than some number N_c, where by this notation we merely mean that the choice of N will usually depend on the choice of c (this will be the crucial point of our later discussion of uniform convergence).

In summary, then, from a pictorial point of view, each member of our sequence involves "freezing" n and looking at the curve $y = f_n(x)$ on [0,1], while the limit function involves our "freezing" x and looking at the sequence $f_1(x)$, $f_2(x)$, etc.

At any rate, with all the preceeding discussion as motivation, we are now ready to define what it means for a sequence of functions to converge to a limit. Namely,

Definition:

The sequence of functions $\{f_n\}$is said to converge (pointwise) to the function f on [a,b] if for each $x \varepsilon [a,b]$, $\lim\limits_{n\to\infty} f_n(x) = f(x)$.

Let us also notice at this point that our discussion, while emphasizing the concept of sequences, applies equally well to series. The reason for this is that every series may be viewed as a sequence of partial sums. For example, suppose we are now given the sequence of functions $f_1, f_2, \ldots, f_n, \ldots,$ etc. and we wish to find the sum:

$$f_1(x) + f_2(x) + \ldots + f_n(x) + \ldots = \sum_{n=1}^{\infty} f_n(x)$$

where $x \varepsilon \text{ dom } f_n$ for each n .

Well, just as we did in the case of numbers, we define a new sequence, say, $s_1(x)$, $s_2(x)$, ... where:

$$s_1(x) = f_1(x)$$

$$s_2(x) = f_1(x) + f_2(x)$$

$$s_3(x) = f_1(x) + f_2(x) + f_3(x)$$

$$s_n(x) = f_1(x) + f_2(x) + \ldots + f_n(x) \quad .$$

We then define $\sum\limits_{n=1}^{\infty} f_n(x)$ to be S(x) where S(x) $= \lim\limits_{n\to\infty} s_n(x)$, provided that the limit exists. In this way, we may talk about a series of functions convergeing to a limit in the sense that we look at the limit of the sequence of partial sums.

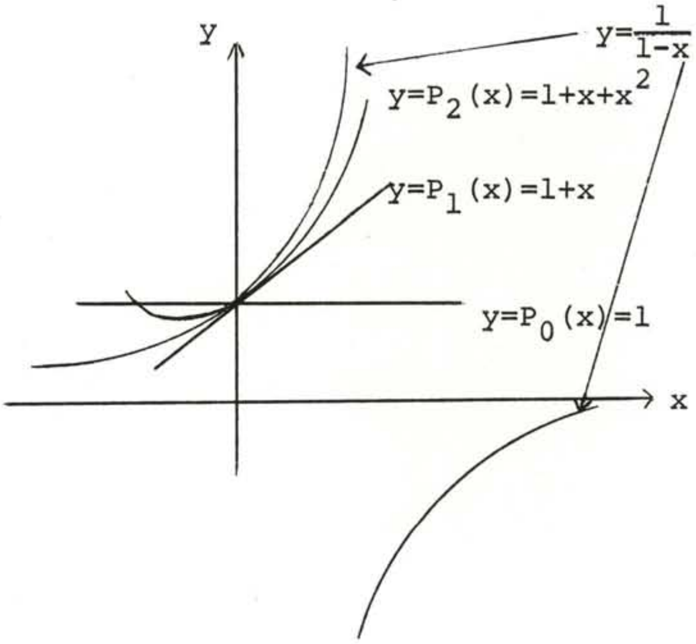
As an illustration that is well within our scope, let us return to the notion of a geometric series. Recall that we have already seen that $\sum_{n=0}^{\infty} x^n$ converges to $\frac{1}{1 - x}$ if $|x| < 1$. What this says in terms of our present discussion is that we may define, say, P_n by $P_n(x) = 1 + x + x^2 + \ldots + x^n$ for each $x \varepsilon (-1,1)$, and if we then define P by

$$P(x) = \lim_{n\to\infty} P_n(x) \quad ,$$

then $P(x) = \frac{1}{1 - x}$.

(Notice that if $x = 1$, $\lim_{n\to\infty} P_n\ (x) = \infty$.)

Again pictorially:



G. Uniform Convergence.

We have seen that e^x is represented by the power series $1 + x + x^2/2! + x^3/3! + \ldots + x^n/n! + \ldots$ for all real numbers, x. If we are now a bit clever, we can get an idea for a new attack on such integrals as $\int_0^1 e^{-x^2}\,dx$. Recall that this is a well-defined definite integral, but in terms of the first fundamental theorem of integral calculus we do not know (explicitly) a function g such that $g'(x) = e^{-x^2}$. In other words, we can write that $\int_0^1 e^{-x^2}dx = g(1) - g(0)$ where $g'(x) = e^{-x^2}$, but we do not have a "hold" on g.

Well, since

$$e^u = 1 + u + u^2/2! + \ldots + u^n/n! + \ldots. \qquad (1)$$

is valid for any real number u, and since u is just a symbol to denote a real number, (1) must remain valid if we replace the symbol u by, say, $-x^2$. If we do this, we obtain:

$$e^{-x^2} = 1 - x^2 + x^4/2! - x^6/3! + x^8/4! + \ldots + (-1)^n x^{2n}/n! + \ldots \qquad (2)$$

Consequently,

$$\int_0^1 e^{-x^2}\,dx = \int_0^1 \left(1 - x^2 + \ldots + \frac{(-1)^n x^{2n}}{n!} + \ldots\right)dx\,. \qquad (3)$$

Now, in (3), <u>if</u> our integral were the sum of a <u>finite</u> number of terms we already know that we could integrate term by term. Thus, we might expect that the same result holds for an infinite sum (although by now we hope you're suspicious of any example in which we try to predict what happens for the infinite based on our knowledge of the finite).

In any event, to compute

$$\int_0^1 \left(1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \frac{x^{10}}{5!} + \ldots\right) dx \,, \qquad (4)$$

we might be tempted to write:

$$\int_0^1 1\, dx - \int_0^1 x^2 dx + \int_0^1 \frac{x^4}{2!}\, dx - \int_0^1 \frac{x^6}{3!}\, dx + \ldots \quad . (5)$$

Notice, however, from a conceptual point of view (4) and (5) are <u>completely different</u>. In (4) we must <u>sum</u> the series <u>first</u> and then integrate while in (5) we must <u>integrate</u> each term <u>first</u> and then sum the resulting series.

From a different (notational) point of view let

$$P_n(x) = 1 - x^2 + \ldots + \frac{(-1)^n\, x^{2n}}{n!} \;.$$

Then (4) asks us to compute

$$\int_0^1 \lim_{n\to\infty} P_n(x)\, dx \quad . \qquad (4')$$

On the other hand, (5) is equivalent to computing

$$\lim_{n\to\infty} \int_0^1 P_n(x)\, dx \quad . \qquad (5')$$

To see this, notice that for any n (since n is <u>finite</u>),

$$\int_0^1 P_n(x)\, dx = \int_0^1 \left[1 - x^2 + \ldots + \frac{(-1)^n\, x^{2n}}{n!}\right] dx$$

$$= \int_0^1 1\, dx - \int_0^1 x^2 dx + \ldots + \int_0^1 \frac{(-1)^n\, x^{2n}}{n!}\, dx$$

$$\therefore \lim_{n\to\infty}\left[\int_0^1 P_n(x)\,dx \checkmark\right] = \int_0^1 1\,dx + \ldots + \int_0^1 \frac{(-1)^n\,x^{2n}}{n!}\,dx + \ldots$$

The crucial point is that it <u>need</u> <u>not</u> be <u>true</u> that

$$\int_a^b \lim_{n\to\infty} f_n(x)\,dx = \lim_{n\to\infty}\int_a^b f_n(x)\,dx \tag{6}$$

no matter how natural such a result may seem.

On the other hand, the ease of evaluating (5) or (5') compared with the problems in computing (4) or (4') makes it clear that we would like to know when (6) is true. The truth of (6) is connected with the topic of this section - <u>uniform convergence</u>. Before tackling this topic, however, let us show that (6) need not be true in every case.

By way of example, let us define the sequence of functions $\{f_n\}$ by:

$$f_n(x) = nx(1 - x^2)^n \text{ where dom } f_n = [0,1] \quad .$$

Let $f(x) = \lim\limits_{n\to\infty} f_n(x) = \lim\limits_{n\to\infty} nx(1 - x^2)^n = 0$*.

*If $|c| < 1$, then $\lim\limits_{n\to\infty} nc^n = 0$. That is $c^n \longrightarrow 0$ much faster than $n \longrightarrow \infty$. One "easy" way of proving this is to show (as we did in the last unit) that if $|c| < 1$ $\sum nc^n$ converges. We then use the fact that if $\sum a_n$ converges then $\lim\limits_{n\to\infty} a_n = 0$. In this case $\sum nc^n$ converges $\longrightarrow \lim\limits_{n\to\infty} nc^n = 0$. The point is that since dom $f_n = [0,1]$, $x \varepsilon$ dom $f_n \longrightarrow 0 \leqslant x \leqslant 1 \longrightarrow 0 \leqslant 1 - x^2 \leqslant 1$. Hence, if $0 < x \leqslant 1$ $nx(1 - x^2) \leqslant n(1 - x^2) \leqslant nc^2$ where $c = 1 - x^2 < 1$. Notice that this discussion does not hold for $x = 0$. Fortunately, however, when $x = 0$, $\lim\limits_{n\to\infty} nx(1 - x^2)^n = 0$ since $nx(1-x^2)^n = 0$.

At any rate we obtain:

$$\int_0^1 \lim_{n\to\infty} f_n\,(x)\,dx = \int_0^1 0\,dx = 0 \qquad (7)$$

On the other hand

$$\int_0^1 f_n(x)\,dx = \int_0^1 nx(1 - x^2)^n\,dx \qquad (\text{let } u = 1 - x^2)$$

$$= \int_1^0 n\,u^n \left(\frac{-du}{2}\right)^{*}$$

$$= \frac{n}{2}\int_0^1 u^n\,du$$

$$= \frac{n}{2}\left(\frac{1}{n+1}u^{n+1}\right|_0^1$$

$$= \frac{n}{2}\left(\frac{1}{n+1}\right)$$

$$= \frac{1}{2}\left(\frac{n}{n+1}\right)$$

$$\therefore \lim_{n\to\infty}\int_0^1 f_n(x)\,dx = \lim_{n\to\infty}\left[\frac{1}{2}\left(\frac{n}{n+1}\right) = \frac{1}{2}\lim_{n\to\infty}\left(\frac{n}{n+1}\right) = \frac{1}{2} \quad . \qquad (8)\right.$$

*Note the change in the limits of integration. Since we are now integrating with respect to u, not x .

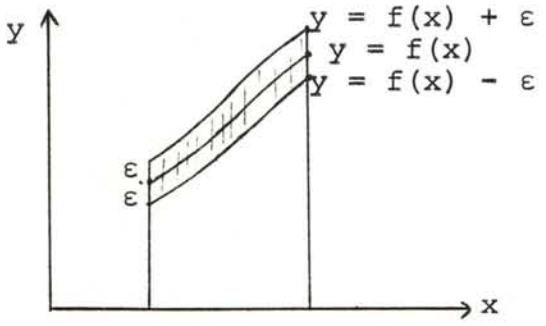
Since $0 \neq \frac{1}{2}$, a comparison of (7) and (8) shows that

$$\int_0^1 \lim_{n\to\infty} f_n (x)\ dx \neq \lim_{n\to\infty} \int_0^1 f_n (x)\ dx \quad . \qquad (9)$$

Thus, whether we like it or not, (9) establishes the fact that (6) need not be true.

To motivate uniform convergence, let us see if we can see pictorially what it might require for $\int_a^b \lim_{n\to\infty} f_n (x)\ dx$ to equal $\lim_{n\to\infty} \int_a^b f_n (x)\ dx$.

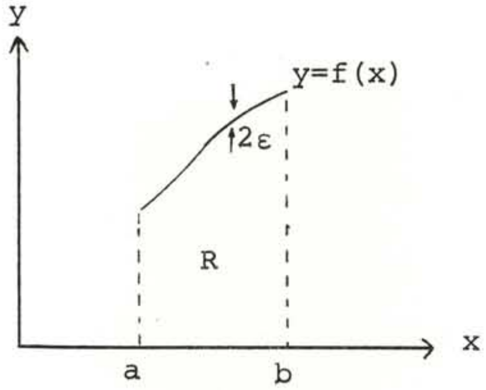
First of all, to view $f_n(x)$ as being within ε of $f(x)$ for all $x \varepsilon [a,b]$ means that we draw the three curves $y = f(x)$, $y = f(x) + \varepsilon$ and $y = f(x) - \varepsilon$. Thus,



(Figure 1)

Now, if we can find N such that $n > N$ implies $|f(x) - f_n(x)| < \varepsilon$ for every $x \varepsilon [a,b]$ it means, pictorially, that for $n > N$, $y = f_n(x)$ lies is the shaded region.

In particular, suppose we now pick ε so small that, as a length, it is less than the thickness of our pencil point. Then,



(Figure 1')

Our argument now is that since for n sufficiently large $y = f_n(x)$ lies "within" the curve $y = f(x)$, in computing the area of R it makes no difference whether we use

$$\int_a^b f(x)\,dx \quad \text{or} \quad \int_a^b f_n(x)\,dx$$ once n is sufficiently

large. That is, $\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx$.

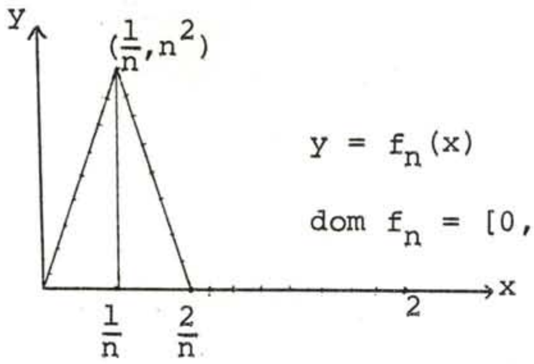
The crucial point is that the convergence <u>does not have to be as depicted in Figure 1</u>. That is, the fact that $\lim_{n\to\infty} f_n(x) = f(x)$ for each $x \varepsilon [a,b]$ does not guarantee that there exists <u>one</u> number N such that $n > N \longrightarrow |f_n(x) - f(x)| < \varepsilon$ for <u>every</u> $x \varepsilon [a,b]$. To be sure, it means that we can find one N for <u>each</u> x, but there are infinitely many x's in [a,b].*

*For a finite number of x's say $x_1, \ldots, x_k$ we could find $N_1, \ldots, N_k$ such that $n > N_i \longrightarrow |f_n(x_i) - f(x_i)| < \varepsilon$ $(i=1,\ldots k)$. We could then let $N = \max \{N_1, \ldots, N_k\}$ whereupon $n > N \longrightarrow |f_n(x_i) - f(x_i)| < \varepsilon$ for each x_i . With infinitely many N_i's, however, the selection process whereby we construct the maximum never terminates.

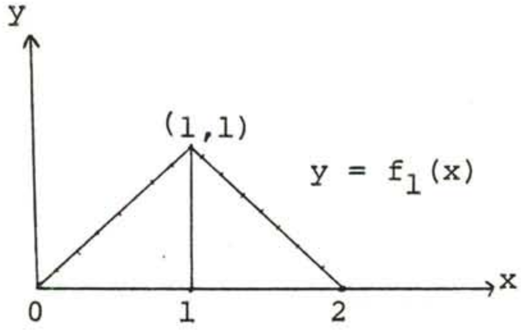
For example, define f_n by

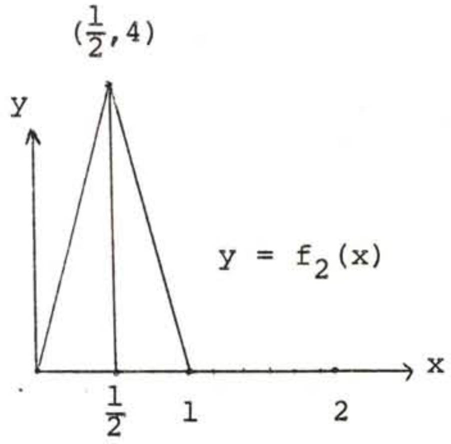
$$f_n(x) = \begin{cases} n^3x, \text{ if } 0 \leqslant x \leqslant \frac{1}{n} \\ -n^3x + 2n^2, \text{ if } \frac{1}{n} < x < \frac{2}{n} \\ 0, \text{ if } x \geqslant \frac{2}{n} \end{cases} .$$

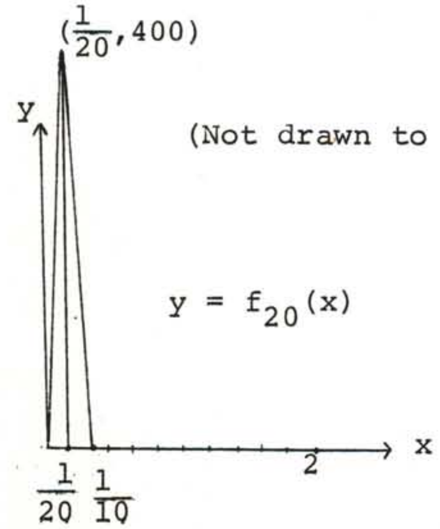
Pictorially,



In particular:







(Not drawn to scale.)

The idea here is to notice that as n increases the "peak" of $y = f_n(x)$ occurs closer to the y-axis $(x = 0)$. That is, while it might not be too obvious at first glance, notice that $\lim\limits_{n\to\infty} f_n(x) = 0$ for every $x \varepsilon [0,2]$. Yet $y = f_n(x)$ has such a great peak for large n that we can never find a large enough n so that all the curves $y = f_n(x)$ lie within ε (where ε is small) of $y = f(x)$ for <u>every</u> x.

To top off this example, observe that

$$\int_0^2 \lim_{n\to\infty} f_n(x)\ dx = \int_0^2 0dx = 0 \tag{10}$$

while:

$$\int_0^2 f_n(x)\ dx = \frac{1}{2}\ bh = \frac{1}{2}\left(\frac{2}{n}\right)n^2 = n$$

$$\therefore \lim_{n\to\infty} \int_0^2 f_n(x)\ dx = \lim_{n\to\infty} n = \infty \quad . \tag{11}$$

A comparison of (10) and (11) establishes most vividly the fact that in this example:

$$\int_0^2 \lim_{n\to\infty} f_n(x)\ dx \neq \lim_{n\to\infty} \int_0^2 f_n(x)\ dx \quad .$$

Thus, "ordinary" convergence may not take place "smoothly enough" to guarantee certain "nice" results.

With this in mind, we now define uniform convergence:

Definition:

The sequence of functions $\{f_n\}$ is said to converge uniformly to f on [a,b] if (1) it converges pointwise and (2) for each $\varepsilon > 0$ there exists a positive integer N (which depends on ε but not on x) such that $|f_n(x) - f(x)| < \varepsilon$ for every $x \in [a,b]$, provided only that $n > N$.

Thus, uniform convergence is "stronger" than ordinary (pointwise) convergence (that is, uniform convergence implies pointwise convergence, but the converse need not be true).

The importance of uniform convergence lies in the following three theorems (which need not be true if the convergence is not uniform).

Theorem 1

If the sequence of continuous functions $\{f_n\}$ converges uniformly to f on [a,b], then f is also continuous. (Note that in the example $f_n(x) = x^n$, dom $f_n = [0,1]$, f(x) was given by 0 if $x \neq 1$,1if $x = 1$. Thus, each f_n was continuous at $x = 1$ but f wasn't. This shows that uniform convergence is necessary if the theorem is to be true.)

Theorem 2

If the sequence of continuous functions $\{f_n\}$ converges uniformly to f on [a,b], then for each $x \in [a,b]$:

$$\lim_{n\to\infty} \int_a^x f_n(t)\,dt = \int_a^x \lim_{n\to\infty} f_n(t)\,dt$$

and this convergence is also uniform on [a,b] .

It would turn out to be nice if there were a theorem analogous to Theorem 2 but which pertained to differentiation instead of integration. Surprising as it may seem, there is no exact analog, although, as we shall soon state, there is a reasonably close substitute. Again, from a more intuitive point of view, differentiation requires more "smoothness" than does integration. As a result, f_n being a good approximation for f allows certain theorems to be true about continuity and integration (which is associated with continuous functions) but false about differentiation. In fact, even though both the proof and the example itself are beyond our needs, it turns out that there exist uniformly convergent sequences of differentiable functions whose limit function, while continuous, is <u>not</u> differentiable at <u>any</u> point. What is true, however, is

<u>Theorem 3</u>:

Suppose $\{f_n\}$ is a sequence of continuously differentiable functions (that is, not only is each function differentiable, but the derivative is also continuous) which is <u>pointwise</u> convergent to f on [a,b]. Then <u>IF</u> the sequence $\{f'_n\}$ converges uniformly on [a,b] it follows that:

(a) f' exists for each $x \in [a,b]$, and

(b) $f'(x) = \lim_{n\to\infty} f'_n(x)$

The proof of each of these theorems is rather straightforwardonce we observe the key idea. For example to prove Theorem 1 we have to show that $\lim_{x\to x_1} f(x) = f(x_1)$ for each x_1 in dom f . This, in turn, means that given an arbitrary $\varepsilon > 0$ we must be able to find $\delta > 0$ such that $|x - x_1| < \delta$ implies that $|f(x) - f(x_1)| < \varepsilon$.

By uniform convergence, what we do know is that for any given $\varepsilon > 0$ we can find N such that $n > N$ implies that both $|f(x) - f_n(x)|$ and $|f(x_1) - f_n(x_1)|$ are less than ε. With this in mind we rewrite $|f(x) - f(x_1)|$ in the following "clever" way:

$$|f(x) - f(x_1)| = |f(x) - f_n(x) + f_n(x) - f_n(x_1) + f_n(x_1) - f(x_1)|$$

$$\leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(x_1)| + |f_n(x_1) - f(x_1)| \quad . \tag{1}$$

Thus, as soon as, for example, each of the numbers $|f(x) - f_n(x)|$, $|f_n(x) - f_n(x_1)|$, and $|f_n(x_1) - f(x_1)|$ is less than $\frac{\varepsilon}{3}$ then $|f(x) - f(x_1)| < \varepsilon$. As we have just mentioned, however, given ε we can find N such that $n > N$ implies $|f(x) - f_n(x)|$ and $|f_n(x_1) - f(x_1)|$ are each less than $\frac{\varepsilon}{3}$. (In fact, by the way of review, by the uniform convergence we know that for this ε, $n > N$ implies $|f_n(x) - f(x)| < \varepsilon$ for each $x \varepsilon [a,b]$.)

At any rate, putting this information into (1) we see that if we choose a fixed $m > N$ then:

$$|f(x) - f(x_1)| < \frac{2\varepsilon}{3} + |f_m(x) - f_m(x_1)| \quad .(2)$$

Since each f_n is continuous, so, in particular is the function f_m in (2) . By definition of continuity, given $\frac{\varepsilon}{3}$ we can find $\delta > 0$ such that $|x - x_1| < \delta \longrightarrow |f_m(x) - f_m(x_1)| < \frac{\varepsilon}{3}$.

For this choice of δ, (2) yields:

Given $\varepsilon > 0$, we can find $\delta > 0$ such that $|x - x_1| < \delta$ implies $|f(x) - f(x_1)| < \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$, and the result is proved.

As for the proof of Theorem 2 we must show that for any $\varepsilon > 0$ we can find N such that

$$\left| \int_a^x f_n(t)\, dt - \int_a^x f(t)\, dt \right| < \varepsilon$$

whenever $n > N$ <u>for</u> <u>all</u> $x \varepsilon [a,b]$.

Our analysis then takes the following lines:

$$\left|\int_a^x f_n(t)\,dt - \int_a^x f(t)\,dt\right| = \left|\int_a^x [f_n(t) - f(t)]dt\right|$$

$$\leqslant \int_a^x |f_n(t) - f(t)|\,dt$$

and since a $\leqslant x \leqslant$ b, it follows that

$$\left|\int_a^x f_n(t)\,dt - \int_a^x f(t)\,dt\right| \leqslant \int_a^b |f_n(t) - f(t)|\,dt$$

$$\leqslant (b - a)\ \max|f_n(t) - f(t)| \quad . \tag{14}$$

With (14) as hindsight we may now say that if we are given $\varepsilon > 0$ we can find N such that $n > N$ implies $|f_n(t) - f(t)| < \frac{\varepsilon}{b - a}$ for all $t \varepsilon (a,x)$, by definition of uniform convergence. Putting this into (14) we find that:

Given $\varepsilon > 0$, then for N as above, we have: $n > N$ implies that

$$\left|\int_a^x f_n(t)\,dt - \int_a^x f(t)\,dt\right| < (b - a)\left[\frac{\varepsilon}{b - a}\right] = \varepsilon$$

and consequently, Theorem 2 is proved.

Theorem 3

Let g denote the limit functions to which $\{f'_n\}$ converges uniformly. Then we have

$$\int_a^x g(t)\ dt = \int_a^x \lim_{n\to\infty} f_n{}'(t)\ dt \quad . \tag{15}$$

Hence, by Theorem 2:

$$\int_a^x g(t)\ dt = \lim_{n\to\infty} \int_a^x f_n{}'(t)\ dt \quad . \tag{16}$$

But by the 2nd fundamental theorem of integral calculus

$$\int_a^x f_n{}'(t)\ dt = f_n(x) - f_n(a) \quad . \tag{17}$$

(This is where we invoke the fact that $f_n{}'$ is continuous.)

Now substituting (17) into (16), we obtain:

$$\int_a^x g(t)\ dt = \lim_{n\to\infty} [f_n(x) - f_n(a)]$$

$$= f(x) - f(a) \quad .$$

And since $f(x) - f(a) = \int_a^x f(t)\ dt$, we may differentiate both sides with respect to x, and using the first fundamental theorem of integral calculus, we obtain:

$$f'(x) = f(x) \qquad (= \lim_{n\to\infty} f_n(x))$$

and the theorem is proved.

Uniform convergence, as we shall see in the next section, plays a vital role in the application of power series. In closing this section, then, we should take a moment to make sure that it is clear that since a series may be viewed as a sequence of partial sums, that the entire theory of uniform convergence of sequences carries over in its entirety to series. More specifically, given the power series $\sum_{n=0}^{\infty} a_n x^n$, we may view this as being $\lim_{n\to\infty} \sum_{k=0}^{n} a_k x^k$, or, without the sigma-notation,

$$a_o + a_1 x + \ldots + a_n x^n + \ldots = \lim_{n\to\infty} (a_o + a_1 x + \ldots + a_n x^n \quad .$$

H. Uniform Convergence of Power Series.

In the previous section, we discussed the concept of uniform convergence. In this section we shall apply these concepts to power series. Before doing this, however, we would first like to obtain a more objective way of being able to decide whether a given convergence is uniform or not. To this end, we discuss what is known as the Weierstrass M-Test.

The test is the following:

Suppose $\sum_{n=0}^{\infty} M_n$ is a convergent series of positive numbers and that $\sum_{n=0}^{\infty} f_n(x)$ is a series of functions for which $|f_n(x)| \leqslant M_n$ for each n and for each $x \varepsilon [a,b]$. Then $\sum_{n=0}^{\infty} f_n(x)$ is uniformly (and absolutely) convergent on $[a,b]$.

The proof of the M-test begins with the comparison test for positive series. Namely, since $|f_n(x)| \leqslant M_n$ for each n and $\sum_{n=0}^{\infty} M_n$ is a convergent positive series, then $\sum_{n=0}^{\infty} |f_n(x)|$ also converges. This, in turn, says that $\sum_{n=0}^{\infty} f_n(x)$ converges absolutely for each $x \varepsilon [a,b]$. Let f, defined by $f(x) = \sum_{n=0}^{\infty} f_n(x)$, denote the function to which $\{\sum_{k=0}^{n} f_k(x)\}$ converges. So far, all we are sure of is that we have pointwise convergence. To prove that the convergence is uniform we must be able to show that $\lim_{n\to\infty} |f(x) - \sum_{k=0}^{n} f_k(x)| = 0$ independently of the choice of x. Thus, the next phase of our proof takes the form

$$|f(x) - \sum_{k=0}^{n} f_k(x)| = |\sum_{k=n+1}^{\infty} f_k(x)|$$

$$\leqslant \sum_{k=n+1}^{\infty} |f_k(x)|$$

$$\leqslant \sum_{k=n+1}^{\infty} M_k = \sum_{k=0}^{\infty} M_k - \sum_{k=0}^{n} M_k$$

$$\therefore \lim_{n\to\infty} |f(x) - \sum_{k=0}^{n} f_k(x)| \leqslant \lim_{n\to\infty} (\sum_{k=0}^{\infty} M_k - \sum_{k=0}^{n} M_k) = 0,$$

independent of x.

But $\lim_{n\to\infty} |f(x) - \sum_{k=0}^{n} f_k(x)| \geqslant 0$ since absolute values are non-negative. Hence $\lim_{n\to\infty} |f(x) - \sum_{k=0}^{n} f_k(x)| = 0$, independent of x.

We shall apply the Weierstrass M-test to power series, but as a "warm up" example, let us consider $f(x) = \sum_{n=1}^{\infty} \frac{\cos n^2x}{n^2}$.* (That is, $f(x) = \lim_{n\to\infty} f_n(x)$ where $f_n(x) = \sum_{k=1}^{n} \frac{\cos k^2x}{k^2}$.)

Clearly $|\frac{\cos n^2x}{n^2}| \leqslant \frac{1}{n^2}$ $(= M_n)$ for each n since $|\cos u| \leqslant 1$ for all real u.

Since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ $(= \sum_{m=1}^{\infty} M_n)$ converges, we conclude from the M-test that $\sum_{k=1}^{n} \frac{\cos k^2x}{k^2}$ converges uniformly to $f(x) = \sum_{n=1}^{\infty} \frac{\cos n^2x}{n^2}$. Suppose we now wish to compute $\int_0^x f(t)\,dt = \int_0^x \sum_{n=1}^{\infty} \frac{\cos n^2t}{n^2}\,dt$. By Theorem 2 of the preceeding section it follows that, since the convergence is uniform,

$$\int_0^x f(t)\,dt = \int_0^x \sum_{n=1}^{\infty} \frac{\cos n^2t}{n^2}\,dt = \sum_{n=1}^{\infty} \int_0^x \frac{\cos n^2t}{n^2}\,dt$$

$$= \sum_{n=1}^{\infty} \left\{ \frac{\sin n^2t}{n^4} \Big|_0^x \right\}$$

$$= \sum_{n=1}^{\infty} \frac{\sin n^2x}{n^4}$$

$$= \sin x + \frac{\sin 4x}{16} + \frac{\sin 9x}{81} + \ldots + \frac{\sin n^2x}{n^4} + \ldots$$

* $$\sum_{n=1}^{\infty} \frac{\cos n^2x}{n^2} = \sum_{n=0}^{\infty} \frac{\cos (n+1)^2x}{(n+1)^2}$$

This looks as though we had integrated $\int_0^x \sum_{n=1}^{\infty} \frac{\cos n^2 t}{n^2} dt$ term by term. In effect, we did, but notice that this is legitimate only because the infinite series of terms converges uniformly.

The major result that we shall prove in this section extends a result which we proved for power series concerning absolute convergence. The result is

For any power series $\sum_{n=0}^{\infty} a_n x^n$, there exists a non-negative number R ($R = 0$ and $R = \infty$ are included as special [extreme] cases) called the radius of convergence of the series such that

(a) the series converges (absolutely) if $|x| < R$, and

(b) the series diverges if $|x| > R$, and

(c) if R_1 is any number for which $0 < R_1 < R$ then $\sum_{n=0}^{\infty} a_n x^n$ converges uniformly for all $x \varepsilon [-R_1, R_1]$.

Actually, only (c) is new since (a) and (b) have been established in our discussion of absolute convergence.

The proof of (c) brings the M-test into play. Outlined, the proof goes like this:

Suppose $\sum_{n=0}^{\infty} a_n {x_0}^n$ converges. Then the nth term must approach 0. That is, $\lim_{n\to\infty} a_n {x_0}^n = 0$. Now, any convergent sequence is bounded. In particular, then, there exists an upper bound M such that $|a_n {x_0}^n| \leqslant M$ for all $n = 1,2,3,\ldots$

Thus, if $|x_1| < |x_0|$ we have:

$$|a_n x_1^n| = |a_n (\frac{x_1}{x_0})^n x_0^n| = |(a_n \; x_0^n)(\frac{x_1}{x_0})^n|$$

$$= |a_n \; x_0^n| \; |\frac{x_1}{x_0}|^n$$

$$\leq M \; |\frac{x_1}{x_0}|^n \quad .$$

(This last step establishes the previous result that $\sum a_n x_1^n$ converges absolutely since $\sum\limits_{n=0}^{\infty} M|\frac{x_1}{x_0}|^n$ is a geometric series which converges because $|\frac{x_1}{x_0}| < 1$.)

Finally, if we now pick any x for which $|x| < |x_1|$ then $|a_n x^n| \leq |a_n x_1^n|$ so that $\sum\limits_{n=0}^{\infty} a_n x^n$ converges uniformly by the M-test since $\sum\limits_{n=0}^{\infty} |a_n x_1^n|$ converges.

This proves our assertion that if R is the radius of convergence for $\sum\limits_{n=0}^{\infty} a_n x^n$ and $0 < R_1 < R$, then for all $x \varepsilon [-R_1, R_1]$, $\sum\limits_{n=0}^{\infty} a_n x^n$ "behaves like" a polynomial.

More specifically,

Theorem 2' (since this is Theorem 2 of the previous section restated for power series)

If the power series $F(x) = \sum\limits_{n=0}^{\infty} a_n x^n$ has radius of convergence R, then for any numbers a and b such that $-R < a < b < R$, $\int_a^b F(x)\,dx$ exists. Moreover,

$$\int_a^b F(x)\,dx \;(= \int_a^b \sum_{n=0}^{\infty} a_n x^n\,dx) = \sum_{n=0}^{\infty} \int_a^b a_n x^n\,dx = \sum_{n=0}^{\infty} \frac{a_n x^{n+1}}{n+1}\Big|_a^b$$

$$= \sum_{n=0}^{\infty} a_n (\frac{b^{n+1} - a^{n+1}}{n+1}) \quad .$$

(Our notation may not be the wisest even though it is quite standard. Do not confuse $\underline{a}$ with a_n. $\underline{a}$ is an endpoint of our interval of integration while a_n is the coefficient of x^n in the power series.)

Theorem 3'

If the power series $F(x) = \sum_{n=0}^{\infty} a_n x^n$ has radius of convergence R, then $F'(x)$ exists for each $x \epsilon (-R,R)$ and $F'(x) = \sum_{n=0}^{\infty} n\, a_n x^{n-1}$ $(= \sum_{n=1}^{\infty} n\, a_n x^{n-1}$ since when $n = 0$, $n\, a_n x^{n-1} = 0)$.

Again, by way of review, a power series may be integrated or differentiated term by term without affecting the radius of convergence.

We can now tie up all loose ends in terms of our sequence $\{P_n\}$ where $P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k$.

Since $\sum_{k=0}^{n} a_k x^k$ is continuous, we may conclude: if the radius of convergence for $\sum_{n=0}^{\infty} a_n x^n$ is R then $\sum_{n=0}^{\infty} a_n x^n$ $(= F(x))$ is a continuous function for $|x| < R$.

Moreover F' exists for $|x| < R$ and it is given by $F'(x) = \sum_{n=1}^{\infty} n\, a_n x^{n-1}$. Since the radius of convergence is still R we may repeat this process with $F'(x)$ to conclude:

$$F''(x) = \sum_{n=2}^{\infty} n(n-1)\, a_n\, x^{n-2}\,, \qquad |x| < R \quad .$$

We may continue this process indefinitely, and after k applications we have:

$$F^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\ldots(n-[k-1])\, a_n x^{n-k}$$

if $|x| < R$. (1)

In particular:

$$F^{(k)}(0) = \sum_{n=k}^{\infty} n(n-1)\ \ldots\ (n-k+1)\, a_n 0^{n-k} \quad .$$

Now $0^{n-k} = 0$ unless $n = k$. Putting this into (1) (that is, letting $n = k$) we find:

$$F^{(n)}(0) = n(n-1)\ \ldots\ (n-[n-1])\, a_n = n!\, a_n$$

$$\therefore a_n = \frac{F^{(n)}(0)}{n!} \quad . \qquad (2)$$

Equation (2) looks like a previously-obtained result from our discussion of polynomial approximations. Actually, it is the converse of this result. Namely, here we started with a power series having a known radius of convergence and showed that the coefficients of the limit function must be given by (2). If we now want to combine these two parts we obtain:

> <u>If</u> a given function f can be represented by* the power series $\sum_{n=0}^{\infty} a_n x^n$ for all $x \in (-R,R)$ then f must possess derivatives of all orders for $x \in (-R,R)$. In particular, $f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\ \ldots\ (n-k+1)\, a_n x^{n-k}$.
>
> Moreover, the coefficients a_n are uniquely determined by:
>
> $$a_n = \frac{f^{(n)}(0)}{n!} \quad .$$

*We say that $\sum_{n=0}^{\infty} a_n x^n$ represents f(x) if $f(x) = \sum_{n=0}^{\infty} a_n x^n$; that is, if $f(x) = \lim_{n\to\infty} \sum_{k=0}^{n} a_k x^k$.

Notice that this result does not tell us whether there is a series $\sum_{n=0}^{\infty} a_n x^n$ which represents a given $f(x)$. That must be determined by, for example, Taylor's Remainder Theorem. What we have proved is that if f if representable by a power series then (1) f must have derivatives of all orders for $x \in (-R,R)$, and (2) there is only one such series representation of f - the one in which $a_n = \frac{f^{(n)}(0)}{n!}$.

Thus , for example, $f(x) = \sqrt{x}$ cannot be represented by a series $\sum a_n x^n$ since $f'(0) = \frac{1}{2\sqrt{0}}$ implies $f'(0)$ does not exist. That is, if $f(x) = \sqrt{x}$, f does not have derivatives of all orders at $x = 0$. In still other words if f is to be represented by $\sum_{n=0}^{\infty} a_n x^n$ it is necessary (but not sufficient *) that f possess derivatives of all orders in which case $a_n = \frac{f^{(n)}(0)}{n!}$.

With all this as background, we may now return to our problem which was used to motivate uniform convergence, namely

$$\int_0^1 e^{-x^2}\, dx \quad .$$

We now know that $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{n!}$ converges uniformly to e^{-x^2} . (Before our discussion of uniform convergence, we could only be sure that we had absolute convergence.)

*Even if f has derivatives of all orders there is no guarantee that f can be represented by a power series. A classic example is $f(x) = \begin{cases} e^{-1/x^2} & x \neq 0 \\ 0 & x=0 \end{cases}$. It can be shown that $f^{(k)}(x)$ exists for each k. However $f^{(k)}(0) = 0$, hence $\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n \equiv 0$ not $f(x)$.

Therefore,

$$\int_0^1 \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{n!}\,dx = \sum_{n=0}^{\infty} \int_0^1 \frac{(-1)^n x^{2n}}{n!}\,dx$$

and we are, thus, allowed to use the procedure outlined in Section G. In other words:

$$\int_0^1 e^{-x^2}\,dx = \sum_{n=0}^{\infty} \left[\int_0^1 \frac{(-1)^n x^{2n}}{n!}\,dx\right] = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)n!}\Big|_0^1$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)n!}$$

$$= \frac{1}{1(0!)} - \frac{1}{3(1!)} + \frac{1}{5(2!)} - \frac{1}{7(3!)} + \frac{1}{9(4!)} - \frac{1}{11(5!)} + R$$

(where $0 \leqslant R \leqslant \frac{1}{13(6!)}$ since we have an alternating series)

$$= 1 - \frac{1}{3} + \frac{1}{10} - \frac{1}{42} + \frac{1}{216} - \frac{1}{1320} + R\,,$$

$$0 \leqslant R \leqslant \frac{1}{9360} \quad .$$

Since $\frac{1}{9360} = 0.0001+$, $1 - \frac{1}{3} + \frac{1}{10} - \frac{1}{42} + \frac{1}{216} - \frac{1}{1320}$ is correct to at least three decimal places as a value of $\int_0^1 e^{-x^2}\,dx$.

Hence, to three decimal place accuracy

$$\int_0^1 e^{-x^2}\,dx = 0.747 \quad .$$

In any event, this completes our discussion of uniform convergence and also of power series. Additional applications are left for the exercises.

Resource: Calculus Revisited
Herbert Gross